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

PROCEEDING

CONTRIBUTED PAPER SESSION

VOLUME 8

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 8)

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 8, 2019. 425 pages

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

Preface

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

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

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

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

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

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

Thank you.

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

Scientific Programme Committee of the 62nd ISI WSC 2019

Chair

Yves-Laurent Grize, Switzerland

Vice-Chairs

Kwok Tsui, Hong Kong, China
Jessica Utts, USA

Local Programme Committee Chair and Co-Chair

Rozita Talha, Malaysia
Ibrahim Mohamed, Malaysia

Representatives of Associations

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

At-Large Members

Alexandra Schmidt, Canada/Brazil
Tamanna Howlader, Bangladesh

Institutional/Ex Officio

Helen MacGillivray, ISI President, Australia
Fabrizio Ruggeri, Past SPC Chair, Italy

Liaison at ISI President's Office

Ada van Krimpen, ISI Director
Shabani Mehta, ISI Associate Director

Local Programme Committee of the 62nd ISI WSC 2019

Chairperson

Rozita Talha

Co-Chairperson

Prof. Dr. Ibrahim Mohamed

Vice Chairperson

Siti Haslinda Mohd Din
Daniel Chin Shen Li

Members

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

TABLE OF CONTENTS

Contributed Paper Session (CPS): Volume 8

Accuracy of tomographic scan-derived implant placement

Laura Antonucci[1], Corrado Crocetta[2], Filiberto Mastrangelo[1], Anna Eleonora Carrozzo[3], Luigi Salmaso[3]

[1] Department of Clinical Science & Dentistry, University of Foggia, Italy
[2] Department of Economics, University of Foggia, Italy
[3] Department of Management and Engineering, University of Padova, Italy

Abstract

This paper analyses the accuracy of a computer-guided template-based implant dentistry by comparing implant position in the virtual project and the actual reached in the bones. We have studied 23 healthy patients subjected to an implant surgical session with the insertion of implants with an external connection. Our Data do not follow a (multivariate) normal distribution and they are not obtained by a well-defined sampling procedure, for this reason, we used a nonparametric procedure based on permutation tests and nonparametric combination.

Keywords

Computer-guided implant dentistry; permutation tests; nonparametric combination; multivariate analysis.

1. Introduction

In dentistry there is a growing interest in new treatments with a reduced number of implants, designed to support fixed prostheses with high aesthetic and functional results. The standard protocol for implant surgery is often based on a diagnostic phase entrusted to radiographic examination and by a clinical examination of the edentulous maxillary bone sites. Two dimensional radiographic examination does not provide an accurate analysis of the bone structures and, consequently, they do not allow to acquire sufficient data for the functional and esthetic design of the implant-prosthetic complex.

An accurate before intervention analysis of the patient's stomatognathic apparatus can be very useful. In this paper we consider a new method for the flapless positioning of endosseous implants in the edentulous patient with the aid of surgical mucosal support and computer guided techniques, based on diagnostic 3D imaging, able to measure the bone volume available for implant surgery, its quality and possible anatomical variations. Thanks to the matching of images of prostheses acquired through optical scanners and three-dimensional radiological images it is possible to study the virtual prosthetic rehabilitations of the patients. Proper virtual planning allows to accurately perform mini-invasive surgery and to use pre-established prostheses made with CAD/CAM methods for an immediate and faithful functionalization of virtual planning. The scientific literature shows that the flapless surgical approach for implant placement has a long-term survival rate similar to the open-flap conventional surgical techniques.

The aim of the study is to evaluate the accuracy of clinical results of dental implant, positioned in total edentulous patients, with CAD/CAM surgical guides, produced after 3D software planning. Through a specific software for the evaluation of three-dimensional deviations it is possible to detect, in the three spatial coordinates, the discrepancies between the project position and the clinical position actually reached. Data available for this study do not allow to consider traditional parametric technics because collecting data did not follow a well-designed sampling procedure and distributional assumptions are difficult to justify. In order to study the discrepancies between the projected position and the actual position, we opted for nonparametric technics in the field of permutation tests.

It is known that in many circumstances permutation tests perform better than parametric tests by providing a valid statistical test with much weaker assumptions (Arboretti et al., 2018, 2017; Pesarin et al., 2016).

2. Methodology

For the study 23 patients were - not randomly – selected after clinical examination. All patients showed total maxillary edentulism and the need to receive a full-arch immediate implant-prosthetic rehabilitation. Patients underwent an implant surgical session, with the insertion of implants with an external connection. Immediately after, there is the load of the prosthetic device. Six months after loading, a control 3D cone beam computed tomography (CBCT) radiographic examination was detected to evaluate the deviations between the virtual project and the clinical position of the fixtures, guided by the surgical template.

Differences for three spatial coordinates (X, Y, Z) between the virtual planning implant position and the clinical actual position in the bone were observed both at the apex and at the entry point of each implant. From a statistical point of view, coordinates measured on the same implant are dependent, whereas different implants can be assumed independent.

Data $\mathbf{D}_j$ = (X_j, Y_j, Z_j), $j \in$ {apex, entry point} are differences given by underlying paired observations *pre* and *post*-surgery. In fact, we may consider observable variables as

$$\mathbf{D}ij = (X_{ij}(post) - X_{ij}(pre), Y_{ij}(post) - Y_{ij}(pre), Z_{ij}(post) - Z_{ij}(pre)),$$

$i = 1,...,n$ and $j \in$ {apex, entry point}. Formalizing we are interested in testing the following system of hypotheses

$$H_0: Pj,post = Pj,pre \qquad \forall j$$

$$H_1: P_{j,post} \neq P_{j,pre} \qquad \text{for at least one } j$$

where $P_{j,post}$ and $P_{j,pre}$ are the multivariate distributions of responses *post* and *pre*surgery respectively, and $j \in$ {*apex, entry point*}. If we assume in both groups the multivariate errors of positioning to be normally distributed, an unconditional solution is represented by the parametric paired Hotelling T^2 test. However, this distributional assumption may not be true, and departures from this assumption can potentially lead to incorrect conclusions. Furthermore, the T^2 test fails to provide an easily implemented one-sided (directional) hypothesis test (Blair et al., 1994) and it does not allow to investigate on partial aspects involved (marginal coordinates), giving only a global result. It is also worth to underline that patients enrolled in this study were not randomly selected, that is one of the assumptions regarding the validity of Hotelling T^2 test. For this reason, we used the permutation test and in particular the nonparametric combination methodology described in Pesarin and Salmaso 2010. Permutation tests are conditional inferential procedures in which conditioning is with respect to the sub-space associated with the set of sufficient statistics in the null hypothesis for all nuisance entities, including the underlying known or unknown distribution. A sufficient condition for properly applying permutation tests is that the null hypothesis implies that observed data are exchangeable with respect to groups. When exchangeability may be assumed in H_0, the similarity and unbiasedness property allow for a kind of weak extension of conditional to unconditional inferences, irrespective of the underlying population distribution and the way sampling data are collected. Therefore, this weak extension may be made for any sampling data, even if they are not collected by welldesigned sampling procedure (Pesarin, 2002). Permutation tests do not require assumptions and/or approximations that may be difficult to meet in real data. In order to solve the global hypothesis testing we have to face separately but simultaneously - the two (multivariate) testing problems (one for each $j \in$ { apex, entry point}) and then combining them through the nonparametric combination (NPC) methodology (Pesarin and Salmaso, 2010). A detailed description of the algorithm used is described in (Antonucci et all 2019).

3. Result

We performed 10,000 permutation tests using the Fisher omnibus combining function, considering different situations: the whole set of implants, the upper and lower arch and front and back mouth.

The whole set of implants

In the first instance we considered the whole set of implants regardless for the position of implants in the mouth. From Table 1 we can see that globally there is a significant difference between the virtual planning implant position and the clinical actual position in the bone (p = 0.00).

Table 1: Results on the whole set of implants

	Apex	Entry point	Global
Combined	0.00	0.00	0.00
X	0.00	0.02	
Y	0.29	0.05	
Z	0.03	0.00	

We can see that this significant result is referred to the spatial X-coordinates, (pX_{apex} = 0.00, $pX_{entry\ point}$ = 0.02) and to the spatial Z-coordinates, (pX_{apex} = 0.03, $pX_{entry\ point}$ = 0.00). Assuming the null hypothesis were true, the P-value approach considers the probability of observing a more extreme test statistic in the direction of the alternative hypothesis than the one observed. If the P-value is small, say less than (or equal to) α, the null hypothesis, that the virtual planned implant position and the clinical actual position are the same, is rejected in favour of the alternative hypothesis. If the P-value is large, say more than α, then the null hypothesis is not rejected. That means that the distances between planned and actual positions of implants are so small that can be consider equal to zero. Considering the implants of the 23 patients considered we can say that the computer aided technique tested has still to be improved

Upper and Lower arch

In this section we split out the entire set of implants into two groups depending on the fact that they are positioned in the upper or in the lower arch of the mouth. Results are shown in Tables 2.

	UPPER		
	Apex	Entry point	Global
Combined	0.00	0.00	0.00
X	0.00	0.31	
Y	0.47	0.14	
Z	0.96	0.35	
	LOWER		
	Apex	Entry point	Global
Combined	0.00	0.00	0.00
X	0.00	0.00	
Y	0.03	0.15	
Z	0.00	0.00	

Table 2: Results on the implants in upper and lower part.

We can see that globally there is a significant difference for implants in the lower and in the upper part of the mouth (p = 0.00). Investigating on partial aspects we can see that differences refer both to apex and entry point (both combined p-values are significant: p_{apex} = 0.00 and $p_{entry\ point}$ = 0.00. In particular, we can see that the technique tested performs better in the upper part of the mouth, where the differences between planned and actual positions for Y and Z coordinates for apex and entry point are not significant. In the lower part, we can only say that the p value for Y coordinates at the entry point (P=0.15) is not significant.

Front and Back

In this section we split out the entire set of implants into two groups depending on the fact that they are positioned in the front (positions starting with 1,2,3) or in the posterior (positions starting with 4,5,6) part of the mouth. Results are shown in Tables 3. We can see that globally there seem to be not significant differences both for the front and back part. Further investigating on partial aspects, we can see that in the front part there are significant differences for the Y-coordinate both at the apex and at the entry point (p_{Xapex} = 0.*11* and $p_{Xentry\ point}$ = 0.*60*). In the back part of the mouth there is significant difference for the Y-coordinate but only at the apex (p_{Xapex} = 0.*9*0) and for the Z-coordinate but only at the apex (p_{Xapex} = 0.*39*).

Table 3: Results on the implants in front and back part.

	FRONT		
	Apex	Entry point	Global
Combined	0.00	0.00	0.00
X	0.00	0.00	
Y	0.11	0.60	
Z	0.03	0.13	
	BACK		
	Apex	Entry point	Global
Combined	0.00	0.00	0.00
X	0.00	0.55	
Y	0.90	0.04	
Z	0.39	0.00	

4. Discussion and Conclusion

In this paper we evaluate the accuracy of the clinical results of dental implant positioned in total edentulous patients with CAD/CAM surgical guides produced after 3D software planning, by analysing the discrepancies between the project position and the clinical position actually reached. Since data available from this study do not allow to consider traditional parametric technics we opted for nonparametric technics in the field of permutation tests.

The results of the analysis highlight that there exist discrepancies between the planned and actual implant position. In general, from the analysis it emerges that the *X*-coordinate (that refers to movements in the internal-external direction of the mouth) is that mainly subjected to errors, both for the apex and the entry point. In particular, it appears that errors are more evident for the implants in the lower part with respect to those in the upper part of the mouth. Furthermore, errors in the implants on the front part refer both to apex and entry point to *X*-coordinate, whereas for those implants on the back of the mouth refers to the apex *X*-coordinate and to the nec Y and Z coordinates. We can conclude that the result analysed lead us to think that further development it is necessary to let this method more reliable.

References

1. Arboretti, R., Carrozzo, E., Pesarin, F., and Salmaso, L. (2017). A multivariate extension of union–intersection permutation solution for two-sample testing. *Journal of Statistical Theory and Practmjnice*, 11(3):436–448.
2. Arboretti, R., Carrozzo, E., Pesarin, F., and Salmaso, L. (2018). Testing for equivalence: an intersection-union permutation solution. *Statistics in Biopharmaceutical Research*, 10(2):130–138.
3. Blair, R. C., Higgins, J. J., Karniski, W., and Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace hotelling's t2 test in prescribed circumstances. *Multivariate Behavioral Research*, 29(2):141–163.
4. Pesarin, F. (2002). Extending permutation conditional inference to unconditional ones. *Statistical Methods and Applications*, 11(2):161–173.
5. Pesarin, F. and Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. Wiley.
6. Pesarin, F., Salmaso, L., Carrozzo, E., and Arboretti, R. (2016). Union–intersection permutation solution for two-sample equivalence testing. *Statistics and Computing*, 26(3):693–701.

Estimating a panel data model with structural change and panel heterogeneity

Sarah B. Balagbis
Supervising Statistical Specialist, Philippine Statistics Authority, Quezon City, Philippines

Abstract

The forward search algorithm and nonparametric bootstrap are used in the context of the back fitting algorithm to estimate a panel data model with structural change and panel heterogeneity. Simulated data with two covariates are used to illustrate the procedure. The method is comparable to time series cross section regression (estimated using generalized least squares) with respect to predictive ability in scenarios where there is actually no perturbation or when there is structural change in the data. The method, however, is superior when there is panel heterogeneity and both panel heterogeneity and structural change in the data. The proposed procedure yields robust covariate parameter estimates. Further, it yields efficient and reliable covariate parameter estimates which are comparable to the time series cross section regression estimated using generalized least squares when there are no real perturbations or when there is structural change in the data.

Keywords

Forward search algorithm; nonparametric bootstrap; back fitting algorithm

1. Introduction

Panel data enable the study of the dynamics of a phenomenon better than either a cross-section or time series alone. Gujarati (2003) as cited by Yaffee (2003) noted that the combination of time series with cross-sections can enhance the quality and quantity of data. Panel data control the heterogeneity of the units and gives more informative data, more variability, and less collinearity among variables. Panel data analysis can also provide a rich and powerful study of a set of units, if one is to consider both the space and time dimension of the data (Yaffee, 2003).

This paper proposes an estimation procedure for panel data modelling in the presence of structural change or panel heterogeneity. The method takes advantage of the benefits from the forward search algorithm and the bootstrap to hopefully come up with robust estimates. It adapts the spatial-temporal model proposed by Landagan and Barrios (2007) but omits the spatial component in the system. A nonparametric bootstrap and the forward search algorithm from Campano (2008) are incorporated into the back fitting procedure proposed by Dumanjug (2007).

Specifically, this paper aims to:

1. Use the back fitting procedure, forward search algorithm, and nonparametric bootstrap to the estimation of panel data model with structural change and panel heterogeneity.
2. assess the robustness, efficiency, reliability and predictive ability of the estimators through simulation.
3. compare the results of the proposed procedure with the time series cross section regression estimated using generalized least squares.

2. Methodology

This section presents the procedure to estimate a panel data model with structural change and/or panel heterogeneity. Estimation will use backfitting procedure, forward search algorithm, and nonparametric bootstrap

The proposed model without perturbation is given by

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \ldots + \varepsilon_{it} \text{ where } \varepsilon_{it} = \rho\varepsilon_{i(t-1)} + a_t \,. \quad (1)$$

In the presence of panel heterogeneity, some cross-sections will follow the following model:

$$Y_{it} = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) X_{1it} + (\beta_2 + \delta_2) X_{2it} + \ldots + \varepsilon_{it} \text{ where } \varepsilon_{it} = \rho\varepsilon_{i(t-1)} + a_t \,. \quad (2)$$

On the other hand, in the presence of structural change, for some time points the model becomes:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \ldots + \varepsilon_{it} \text{ where } \varepsilon_{it} = (\rho + \delta)\varepsilon_{i(t-1)} + a_t \,. \quad (3)$$

Furthermore, if both panel heterogeneity and structural change are present, the model is expressed as

$$Y_{it} = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) X_{1it} + (\beta_2 + \delta_2) X_{2it} + \ldots + \varepsilon_{it} \text{ where } \varepsilon_{it} = (\rho + \delta)\varepsilon_{i(t-1)} + a_t \,. \quad (4)$$

The models assume constant covariate effect (β) across locations and time, and constant temporal effect (ρ) across locations. Only two covariates (X_1, X_2) are considered in this paper. The proposed procedure also assumes the presence of panel heterogeneity and/or structural change.

The models (1) to (4) are a modification of the spatial-temporal model proposed by Landagan and Barrios (2007). It only omits the spatial component. This paper considers only the covariate effect (β) and the temporal effect (ρ).

To estimate the parameters of the model

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \ldots + \varepsilon_{it} \text{ with } \varepsilon_{it} = \rho\varepsilon_{i(t-1)} + a_t$$

given data layout above, the following procedure is proposed:

1. For each t=1,2,..., T in a give panel data realizations $\{Y_{it}\}$, i=1,...,n

a. Choose *m* observations as "close" to each other as possible. This is accomplished by fitting the model $y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} +..$ for all observations and choosing m observations with the smallest residuals.

b. Using the m observations from step (a), fit the model $y_{it} = \beta_0^1 + \beta_1^1 X_{1it} + \beta_2^1 X_{2it} + \ldots$ and store the parameter estimates. Compute the residuals of the remaining *(n-m)* observations and choose one observation with the smallest residual to be added to the m observations in step (a). This yield (*m*+1) observations.

c. Using the (*m+1*) observations from step (b), fit the model again and store the parameter estimates. Compute the residual of the remaining *[n – (m+1)]* observations and choose one observation with the smallest residual.

d. Continue the process of estimating the parameters while Cook's distance for the newly entered observation is less than ε, otherwise stop the process and get the series of parameter estimates you have stored.

2. For each of the series of parameter estimates, compute a bootstrap estimate:

a. Draw a simple random sample of size n with replacement.

b. Compute the mean. $\hat{\theta}^{(b)} = \frac{1}{n}\sum_{k=1}^{n} \hat{\theta}_k$

c. Repeat (a) and (b) B times, where B is large.

d. Compute the bootstrap estimat $\hat{\theta}_{BS} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{(b)}$, and the standard error $\hat{\sigma}_{BS} = \left[\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{BS})^2\right]^{1/2}$

3. Given the bootstrap estimates of the parameters do the following:

a. Compute $e_{it} = Y_{it} - \hat{\beta}_0 - \hat{\beta}_1 X_{1it} - \hat{\beta}_2 X_{2it} - \cdots$

b. For each i = 1, 2, ...N, fit the model $\varepsilon_{it} = \rho\varepsilon_{i(t-1)} + a_t$ and store the parameter estimates as ρ^i, i = 1,..., N.

c. Compute the bootstrap estimate $\hat{\rho}_{BS}$ by following step 2 above.

4. Generate new series $Y_{ij}^1 = Y_{ij} - \hat{\rho} e_{i(t-1)}$ and iterate from step 1. Continue the iteration until there is no substantial change in the values of the parameter estimates

Simulation Studies

A simulation study is performed to illustrate and evaluate the performance of the proposed estimation procedure. Seventy (70) panel points and eighty (80) time points are considered resulting to a panel with 5,600 data points. The first independent variable *(X_1)* is generated from a uniform distribution from the interval (20, 30). The second independent variable (X_2) is generated from N(10,1).

The dependent variable *(Y)* is generated by first simulating the independent variables and the error terms *(a_t)*. Eighty (80) a_t's are simulated from a standard normal distribution. Given a_t's and setting ρ=0.5, the error component $\varepsilon_{it} = \rho\varepsilon_{i(t-1)} + \alpha_t$ is computed. Note that we need the initial value ε_{i0} . We set $\varepsilon_{i0} = 0$ which is tantamount to setting the first value $\varepsilon_{11} = a_1$. All the other value of ε_{it} for t=1 will be initialized by the previous value in the data layout. Using the simulated independent variables X_1 and X_2, and the computed error component ε_{it}, and setting β_0= 1.2, β_1=0.4, and β_2=0.6, the dependent variable $Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \varepsilon_{it}$ is computed. This constitutes the dependent variable without perturbation.

Panel heterogeneity is introduced into the system by replacing the Y values of the randomly chosen panels for all T =1, 2, …, 80 by setting β_0= 0.8, β_1=1.4, β_2=2.1. Structural change is introduced into the system by choosing time points at the start, middle, end, and time points located in all 3 locations (start, middle, and end) and changing the Y values for these time points for all N = 1, 2, …, 70 by setting ρ=0.8. Panel heterogeneity and structural change are incorporated by combining the two strategies above and by setting β_0= 0.8, β_1=1.4, and β_2=2.1, and ρ=0.8.

Twenty data sets without perturbation are simulated in the study and for each of these, 5% (≈280 observations) and 10% (≈560 observations) panel heterogeneity or structural change are incorporated. For the structural change and a mixture of structural change and panel heterogeneity, four scenarios are considered: (i) perturbations at the start; (ii) perturbations at the middle; (iii) perturbations at the end; and (iv) perturbations at the start, middle, and end, i.e., spread all throughout the 3 locations (Table 1).

For the 5% panel heterogeneity, 3 panels are randomly selected while for 10% panel heterogeneity, 7 panels are randomly selected. Four time points are selected at each different location (start, middle, end, throughout the 3 locations) for the 5% structural change, and 8 time points are selected for the 10% structural change. For a mixture of panel heterogeneity and structural change, 14 panels are randomly chosen and 20 time points at each location are selected for the 5%, while for the 10% panel heterogeneity and structural change, 28 panels are randomly chosen and 20 time points at each location are selected. A total of 360 data sets with perturbation will be used in the study:

Table 1. Simulated data sets with perturbation

PERTURBATION									
	Panel Heterogeneity	Structural Change				Panel Heterogeneity & Structural Change			
%		Start	Middle	End	All	Start	Middle	End	All
5%	20	20	20	20	20	20	20	20	20
10%	20	20	20	20	20	20	20	20	20

Comparison with Time Series Cross Section Regression (Generalized Least Squares)

The estimates are compared with time series cross section regression estimated using generalized least squares based on the following measures:

1. Robustness $\%bias = \left|\frac{trueparametervalue - estimate}{trueparametervalue}\right| x100\%$
2. Efficiency $MSE = Var(\hat{\theta})$ (if bias is not a serious problem

   $MSE = Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2$ (if bias is a serious problem)
3. Reliability $CV(\%) = \frac{se}{mean}\ x\ 100$

   Where se = standard error of the bootstrap estimate and mean = bootstrap estimate
4. Predictive Ability MAPE $= \frac{\sum \frac{|y-\hat{y}|}{y}}{NT}$X100 where NT is the total number of observations.

3. Result

The mean among the estimates from twenty (20) data sets are used in the evaluation. Note that the biases of β estimates from the proposed method are generally tolerable while the biases of the generalized least squares parameter estimates are generally not tolerable. The tolerable bias of the parameter estimates from the proposed method can mean that the additivity assumption of the backfitting algorithm is satisfied, and that the robustness from the forward search algorithm is inherited by the proposed method.

In general, the behavior of the parameter estimates using the proposed method is comparable to the generalized least squares estimates when there is no perturbation in the data and when structural change is present in the data (Table 2). Likewise, the behavior of the parameter estimates using the proposed method is comparable to the generalized least squares estimates in the presence of panel heterogeneity and a mixture of panel heterogeneity and structural change.

Without perturbation or with structural change in the data, the proposed method and the time series cross section regression (estimated using generalized least squares) yields comparable β estimates which are both near to the true value. In the presence of panel heterogeneity and a combination of panel heterogeneity and structural change, the proposed method yields β estimates which are generally nearer the true value than the generalized least squares (Table 2).

Without perturbation or with structural change, the β1 and β2 estimates from the proposed method and the generalized least squares β1 and β2 estimates are comparable with respect to efficiency, reliability, robustness, and predictive ability (Table 2). In the presence of panel heterogeneity and a mixture of panel heterogeneity and structural change, however, the β1 and β2 estimates from the proposed method are generally more robust and have better predictive ability than their generalized least squares counterpart though they are less efficient and less reliable (Table 2).

The ρ estimate from the proposed method is generally inferior to the generalized least squares estimate (efficiency, reliability, and robustness) (Table 2).

Table 2. Summary of comparison: Proposed method vs. time series cross section regression estimated using generalized least squares

		Without perturbation	Panel heterogeneity		Structural change		Panel heterogeneity and structural change	
			5%	10%	5%	10%	5%	10%
Estimates	β0	C	*	*	C	C	* except All	*
	β1	C	*	*	C	C	*	*
	β2	C	*	*	C	C	*except All	*
	ρ	x			x	x		
Efficiency	β0	x	x	x	x	x	x	x
	β1	C	x	x	C	C	x except ALL	x
	β2	C	x	x	C	C	x	x
	ρ	x	x	x	x	x	x	x
Reliability	β0	x	x	x	x	x	x	x
	β1	C	x	x	C	C	x	x
	β2	C	x	x	C	C	x	x
	ρ	x	x	x	x	x	x	x
Robustness	β0	*	*	*	*except: ALL	*except: END	* except ALL	*
	β1	C	*	*	C	C	* except ALL	*
	β2	C	C	*	C	C	* except ALL	*
	ρ	x	x	x	x	x	x	x

Predictive Ability		C	*	*	C	C except: END	*	*

*Legend C – two methods comparable; x – proposed method is inferior to the time series cross section regression (GLS); * - proposed method better than GLS; blank – both values are far from the true value.*

4. Discussion and Conclusion

The following conclusions are arrived at based on the simulation results:

1. Covariate parameter estimates change minimally only regardless of the extent of structural perturbation and/or panel heterogeneity (5% or 10%) and regardless of the location of the perturbation except when 5% panel heterogeneity and structural change is spread all throughout the three (3) locations.
2. Robustness: (a) With no perturbation or with structural change, β1 and β2 estimates are robust and are comparable for the two methods; (b) With panel heterogeneity, β1 and β2 estimates are better for the proposed procedure except β2 with 5% panel heterogeneity which proves to be just comparable; (c) With panel heterogeneity and structural change, β1 and β2 estimates are better for the proposed procedure except when 5% perturbation are spread all throughout the three (3) locations.
3. Efficiency: (a) With no perturbation or in the presence of structural change, the covariable parameter estimates for the two methods are comparable; (b) With panel heterogeneity or a mixture of panel heterogeneity and structural change, GLS is more efficient though the β1 estimates of the proposed method can also be considered efficient since the values are generally small (except when 5% panel heterogeneity and structural change is spread all throughout the three (3) locations.
4. Reliability: With no perturbation or with structural perturbation, the β1 and β2 estimates of the proposed method are reliable and comparably reliable to its GLS counterpart.

Assessing the effect on the ρ estimate when the temporal parameter is estimated first before the covariate parameters are estimated or when the temporal parameter is estimated simultaneously with the covariate parameters would be informative.

References

1. Baltagi, Badi H. (1995). Econometric Analysis of Panel Data. Chichester: Wiley.
2. Baltagi, Badi H, et. al. (2003). Testing for Serial Correlation, Spatial Autrocorrelation and Random Effects Using Panel Data. Econometric Society 2004 Australasian Meetings, No. 338.
3. Campano, Wendell Q (2008). An Estimation Procedure of an ARIMA Model in the Presence of Structural Change. Unpublished MS thesis. UP School of Statistics.
4. Dumanjug, Charlotte F. (2007). Bootstrap Procedure in an Iteratively Estimated Spatial-Temporal Model. Unpublished MS thesis. UP School of Statistics.
5. Guarte, J.M. (2004). Estimation Under Purposive Sampling. Unpublished MS Thesis. UP School of Statistics.
6. Guarte, J.M. and E. B. Barrios (2006). Estimation Under Purposive Sampling. Communication in Statistics – Simulation and Computation, Volume 35, Issue 2.
7. Landagan, O. Z. and E. B. Barrios (2007). An Estimation Procedure for Spatial-Temporal Model. Statistics and Probability Letters, 77:401-406.
8. Yafee, Robert A. (2003). A Primer for Panel Data Analysis. http://www.nyu.edu/its/statistics/Docs/pda.pdf

How trees can help to learn statistics

Rima Kregzdyte
Lithuanian University of Health Sciences, Kaunas, Lithuania

Abstract

Teachers know and use different methods of teaching statistics. Many years' experience in teaching statistics to biomedical students allowed to appear one more way to make statistics more close to nontechnical people. Suggested way for teaching statistics is based on principle of growing tree. Practical application of statistical methods in undergraduate works and master's theses of biomedical students at Lithuanian University of Health Sciences was reviewed Students, who were taught statistics using statistical tree, correctly chose and applied more various methods than students, who were taught statistics as mathematically-oriented explanation of a particular method. Principle of growing statistical tree makes understanding of choosing and application of statistical methods easier and more attractive.

Keywords

Teaching Statistics; Statistics Education; Learning Statistics

1. Introduction

Curious and creative people observe nature, try to understand processes happening around, and try to evaluate existing complex systems. Statistical knowledge and ability to apply them is necessary in order to achieve the tasks. There are many methods and tricks of teaching statistics (Gelman & Nolan, 2002; Čekanavičius & Murauskas,2000; Čekanavičius & Murauskas, 2002). But no one of them is universal, good for all target groups.

Modern biomedical students also need to understand the choice of the statistical method and the results of performed statistical analysis. As many biomedical students dislike statistics, therefore the most appropriate teaching way should be found. Suggested way for teaching statistics is based on principle of growing tree.

2. Methodology

Two different methods of teaching statistics were used at Lithuanian University of Health Sciences. One was mathematically-oriented explanation, based on teaching about a particular method, without generalized idea of statistical analysis. Another was problem-oriented explanation, based on systematic approach and visualized by statistical tree. Statistical tree consists of stem (data), branches (tasks) and leafs (tests). Tree can grow into different directions, adding additional branches. Different sorts of trees are appropriate for different levels of students.

Practical applications of statistical methods in undergraduate works and master's theses of biomedical students learnt differently were reviewed.

3. Results

Classical statistical teaching begins from getting acquaintance with probabilities, following by detailed explanation of a particular statistical test. Such teaching methodology is used many years at the Department of Physics, Biophysics and Mathematics of Lithuanian University of Health Sciences. Students learn name of statistical test and how to calculate its statistics.

Other statistical teaching practice is used at the Department of Public Health. This way is problem-oriented approach to statistics. First of all, a scientific problem is formulated. Next step is looking for the statistical solution of this problem.

The base of the successful analysis should be strong data. Data can be imagined as the stem of a tree (Figure 1).

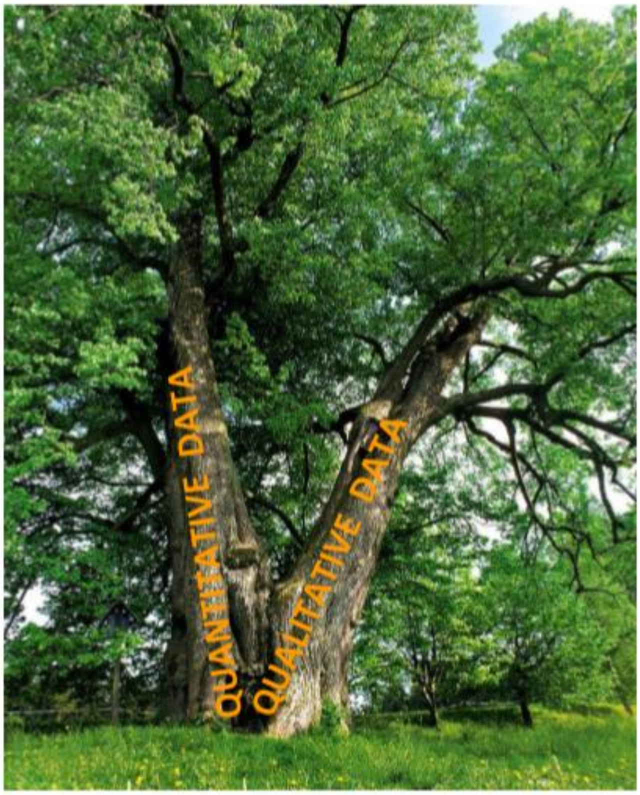


Figure 1. Presentation of data types.
(Lime tree at Lipka in Czech Republic – finalist of the European Tree of the Year 2011)

When data type is defined it is possible to look at the task. All main statistical tasks may be aggregated into three generalized tasks: to describe group, to compare groups, to find relationship. The tasks can be imagined as branches (Figure 2). When students decide what generalized task is related to the formulated problem, they are able to find more detailed task. Detailed statistical task could be: to describe quantitative variable in two groups, to compare two independent groups, to compare 2 dependent groups, to compare more than 2 independent groups, to compare more than 2 dependent groups, to evaluate relationship.

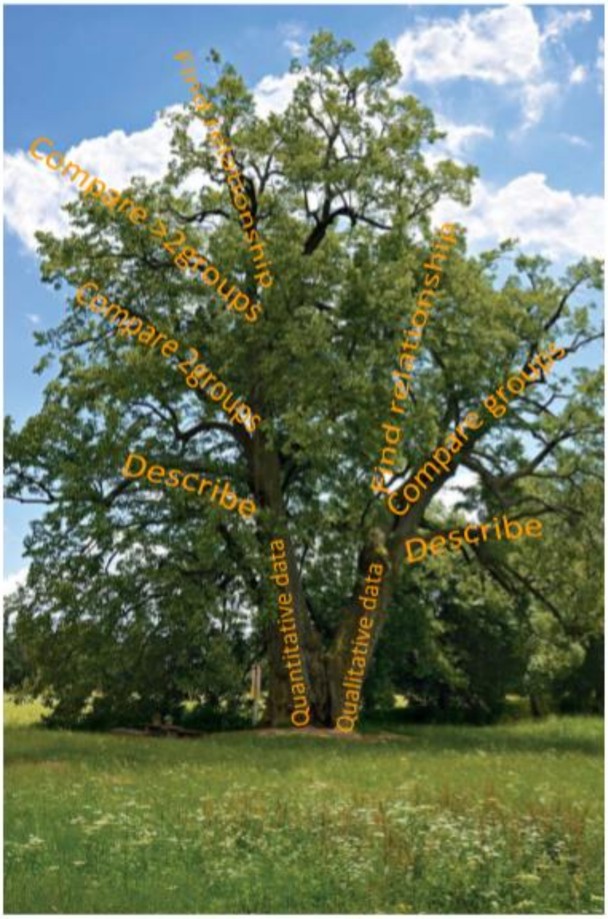


Figure 2. Presentation of statistical tasks.
(Lime tree at Lipka in Czech Republic – finalist of the European Tree of the Year 2011)

For each task the appropriate statistical test must be chosen. Separate tests can be imagined as leaves of a tree. Basic tests can be presented as in general as in the simplified trees (Figure 3 and Figure 4). Similar trees eliminate fear of mathematical formulas. Students become more flexible in discussions about data analysis. They see wider picture of possible tests and are able to find the most appropriate statistical test.

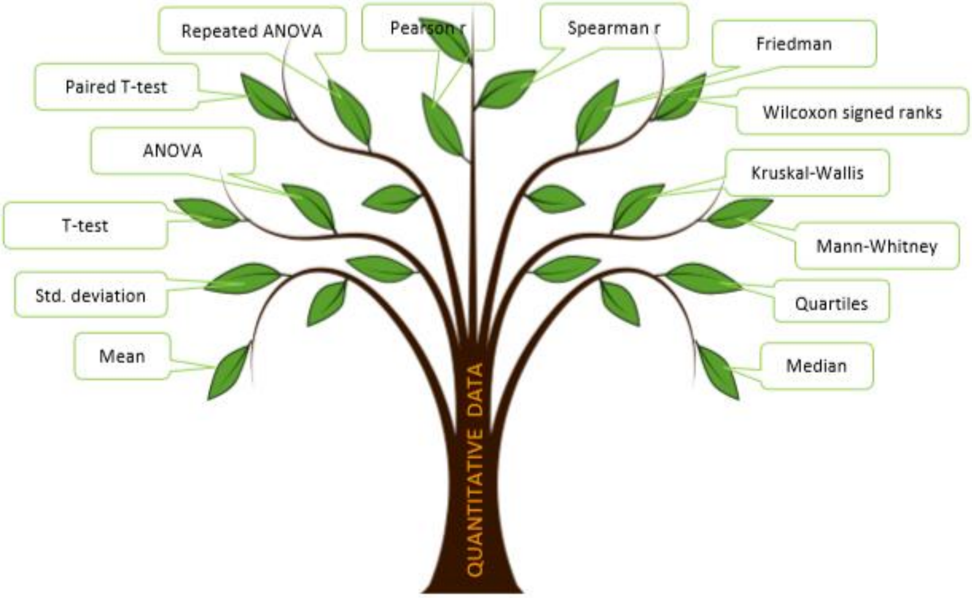


Figure 3.
Simplified tree of statistical tests for quantitative data analysis.

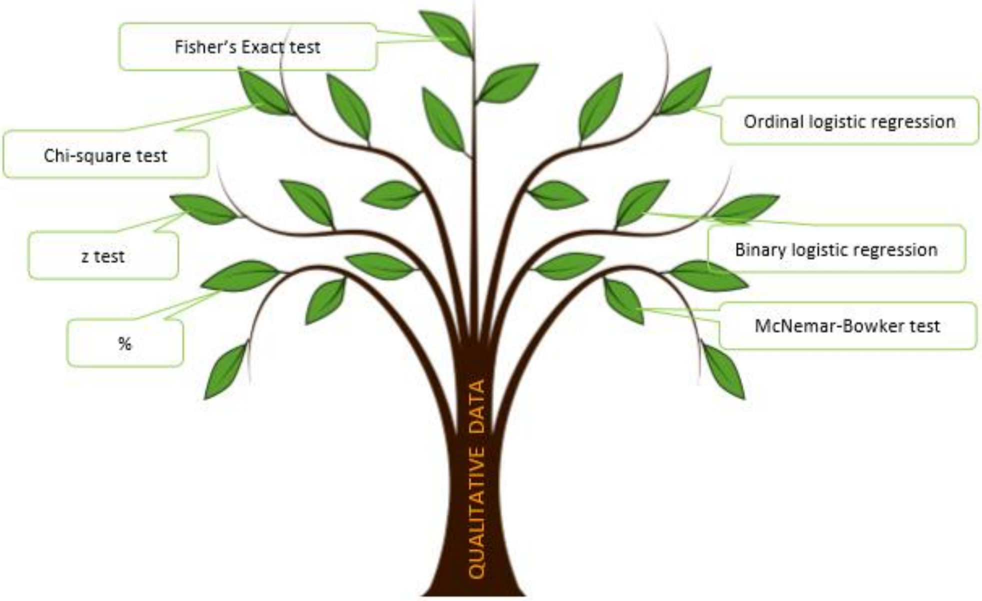
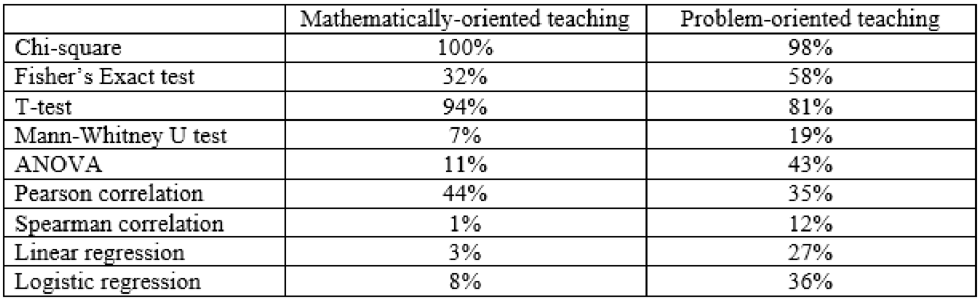


Figure 4.
Simplified tree of statistical tests for qualitative data analysis.

Abilities to choose a statistical test can be seen in final works of students. Students, who were taught statistics as mathematically-oriented explanation of a particular method, used mostly Chi-square and T-test.

Students, who were taught statistics using statistical tree, correctly chose and applied more various methods, such as Chi-square, Fisher’s Exact test, T-test, Mann-Whitney U test, ANOVA, linear or logistic regression (Table 1).

Table 1. Practical application of statistical tests in undergraduate works and master’s theses of biomedical students learnt differently at Lithuanian University of Health Sciences.



	Mathematically-oriented teaching	Problem-oriented teaching
Chi-square	100%	98%
Fisher’s Exact test	32%	58%
T-test	94%	81%
Mann-Whitney U test	7%	19%
ANOVA	11%	43%
Pearson correlation	44%	35%
Spearman correlation	1%	12%
Linear regression	3%	27%
Logistic regression	8%	36%

Some of the latter students considered strength and weakness of their works in relation to the collected data and statistical analysis. Most of them were able to formulate and solve real-world data problems.

4. Discussion and Conclusion

Mathematically-oriented explanation of statistical methods is well written in many books (Kallen, 2011; Yandell, 1997; Zar, 1999). Such methodology is appropriate for mathematically-thinking students. But it is heavily acceptable for non-mathematically-thinking students.

Creative teachers create various means that makes understanding of statistical secrets easier, more tangible. There are ideas to use songs (Lesser, 2001), fun elements (Lesser & Pearl, 2008; Kranzler, 2017), gaming with camels (Lyford et al., 2019). But these tricks are understandable and acceptable not in all cultures.

Students and teachers are part of nature. Surrounding environment and culture makes influence on thinking. People in different countries protect trees. Since 2011 the contest of the European Tree of the Year is held by the Environmental Partnership Association. The winner is announced at the awards ceremony in the European Union Parliament. Lithuanian people love and take care of nature and especially trees, celebrate traditional holydays near the trees (Figure 5).

Teaching of different subjects based on human nature is more acceptable and successful. Principle of growing statistical tree makes understanding of choosing and application of statistical methods easier and more attractive. Statistical tree is helpful educational mean as for bachelor as for master students.

Figure 5. Folk traditional celebration near the oldest oak (more than 1000 years) in Lithuania

References

1. Čekanavičius, V. & Murauskas, G. (2000). Statistika ir jos taikymai, I. (Statistics and its applications, I). Vilnius. TEV.
2. Čekanavičius, V. & Murauskas, G. (2002). Statistika ir jos taikymai, II. (Statistics and its applications, II). Vilnius. TEV.
3. Gelman, A. & Nolan, D. (2002). Teaching statistics. A bag of tricks. Oxford. Oxford University Press.
4. Kallen, A. (2011). Understanding biostatistics. Chichester. John Wiley & Sons.
5. Kranzler, J.H. (2017). Statistics for the terrified, 6th ed. Lanham, MD: Rowman and Littlefield.
6. Lesser, L.M. (2001). Musical means: Using songs in teaching statistics. Teaching Statistics, 23 (3), 81-85.
7. Lesser, L.M. & Pearl, D.K. (2008). Functional fun in statistics teaching: Resources, research, and recommendations. Journal of Statistics Education, 16 (3), 1-11.
8. Lyford, A. Rahr, T., Chen, T., & Kovach, B. (2019). Using camels to teach probability and expected value. Teaching Statistics, 41 (1), 18-24.
9. Yandell, B.S. (1997). Practical data analysis for designed experiments. London. Chapman&Hall. 10. Zar, J.H. (1999). Biostatistical analysis. New Jersey. Prentice Hall International.

A study on correlation between literacy and sex ratio in a Northeast State of India

Rajkumari Latasana Devi[1], Rajkumari Sanatombi Devi[2], Sumati Rajkumari[1]
[1] Ghana Priya Women's College
[2] Sikkim Manipal Institute of Medical Sciences, Sikkim Manipal University

Abstract

One of the basic demographic characteristics of a population is the sex composition. The sex composition of a population has a considerable impact on health, social, economic condition of a country. According to census of India "a person aged more than 6 years and who can both read and write with understanding in any language has taken as a literate" and the sex ratio is defined as the number of females born per 1000 males. The sex ratio is considered as a tool to determine the gender equity of the population. The objective of the present study is to determine the correlation between the literacy and sex ratio in Sikkim, India. The study was based on census report 2011 published by the Registrar General, Government of India. The literacy was considered as an independent variable, while sex ratio is considered as a dependent variable. Using Spearman's rank correlation method, it was observed that literacy and sex ratio were negatively correlated which was not statistically significant (r = - 0.40, P=0.60). The study reveals that there was a low negative correlation between literacy and sex ratio of the population in Sikkim. In other word, an indirect relationship was found between the literacy and sex ratio in the state.

Keywords

Population; Census; Gender equity; Northeast India

1. Introduction

One of the important indicators of social development is the level of literacy and a high level of which is considered to be an important factor in the process of modernization. The level of social development in different countries, regions or even in the various political divisions of the same country may be compared on the basis of literacy. Literacy is one of the most important components uses to measure the quality of life or wellbeing of a country. The United Nations has defined literacy as the ability of a person to read and write with understanding a short simple statement on his everyday life. According to census of India a person aged more than 6 years and who can both read and write with understanding in any language has taken as a literate. Literacy rate is defined as the percent of literate persons in the age group 7 and above, to population in ages 7 and above. It is calculated as

Literacy rate=Number of Literates/(Population of age 7+)*100. In the 1991 census, by definition all children below the age of 7 were considered as illiterate. According to 2011 census literacy rate of India was found to be a total of 74.04% with 65.46% literate females and 82.14% males. This was a 9.81% increase since the last census. India has 48.53% female population compare to 51.47% male population as per census 2011. Literacy plays a very important role development of the society at large. The level of literacy indicates the level of economic development, living standards, status of women in a society. The sex ratio, in India, is defined as the number of females born per 1000 males (femininity ratio) in the population whereas, internationally, the sex ratio is defined as number of males per 100 females (masculinity ratio). Because of its significant demographic, economic and social implication, it is essential to identify the factors that affect the sex composition of a population. The most fundamental implication of the sex ratio is to define the limits of the society's' reproductive potentials. The sex ratio is also considered as a tool to determine gender equity of a population. Historically, the sex ratio in India remained favourable to males than females. Literacy plays a vital role in the improvement of the sex ratio of a population. The sex ratio generally shows a balance between the number of male and female of a population. In a patriarchal society, sex ratio is always unfavourable for females and in matrilineal society the sex ratio is favourable to females. Since India is a patriarchal society, the sex ratio is favourable to males as compared with females.

Study area

The state Sikkim is spread below Mount Kanchanjunga (8,534 m), the third highest peak in the world and it became the 22nd state of India on 26 April 1975.Sikkim is the second smallest state in term of areas with 7096 sq. km after Goa. It is located in the northeastern part of India. Sikkim lies between 27.04 degree to 28.07-degree North latitude and 80.00 degree to 88.55-degree East longitude. It is bound on the North by the Tibet plateau, on the East by Chumbi valley of Tibet and Bhutan, on the West by Nepal and on the South by Darjeeling district of West Bengal. It shares a 97.8-kilometre-long international boundary with Nepal, 30.90 kilometre long border with Bhutan and a 220.35 kilometre long border with Tibet. The state constitutes 0.22% of India's geographical area and 2.7% of the Northeast states of India. It the least populated state in India with a population of 607,688 in which 52.9% are males and 47.1% are females according to 2011 provisional population census. The density of Sikkim is 86 per sq km. There are four districts in the state. The four districts are: East, West, North, and South district. East district is having the highest population of 281,293 followed by South district (146,742). North district is having the least population of 43,354 persons in the state. The population of West district is 136,299 persons.

Trends of literacy and sex ratio in India and Sikkim between 1991 and 2011

Fig.1 showed the progress of literacy in India and Sikkim from 1991 to 2011. As per census 2011, the literacy rate of India has 74.04% and it was also observed that from 1991 to 2011 the literacy rate in Sikkim was higher than the national literacy rate. The rate of increased in literacy in the state was 44.88% while the rate of increased in literacy in India was 41.18%.

Fig.1 Comparison literacy rate between India and Sikkim between 1991 and 2011

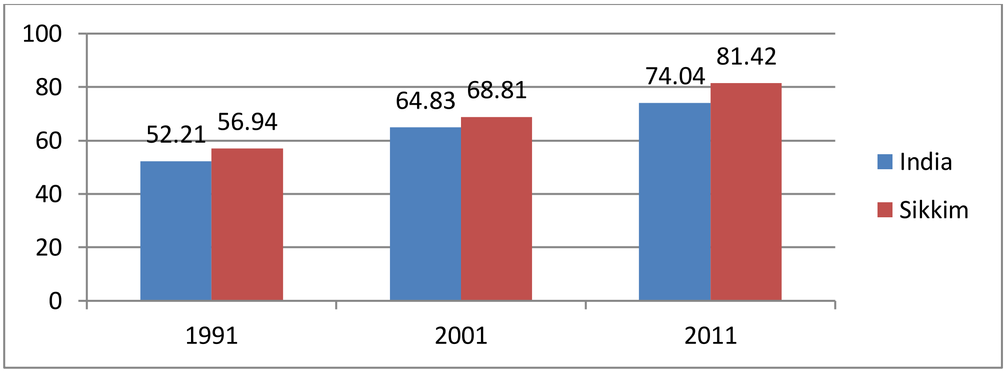


Fig.2 showed the changing pattern of sex ratio in India and Sikkim. There was an improvement in overall sex ratio in India between 1991 and 2011. The sex ratio of India in 2011 was 940 females per 1000 males registering an improvement of 7 points on the 2001 census of 933 females per 1000 males. In Sikkim, it was 889 females per 1000 males which were lower than the national sex ratio of 940 females per 1000 males.

Fig.2 Comparison sex ratio between India and Sikkim (1991 – 2011)

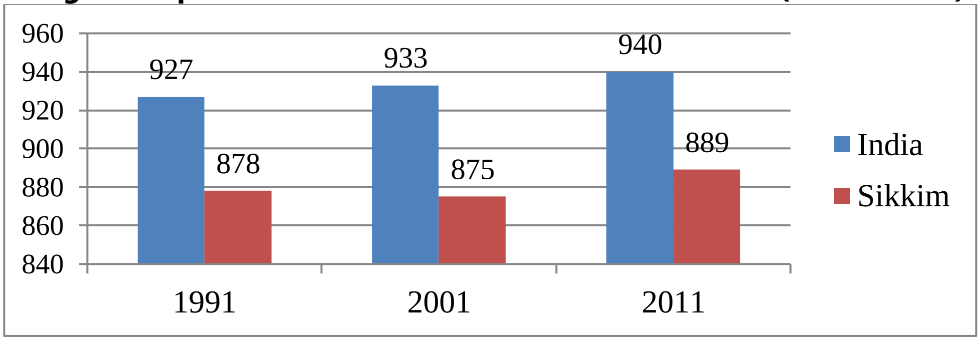
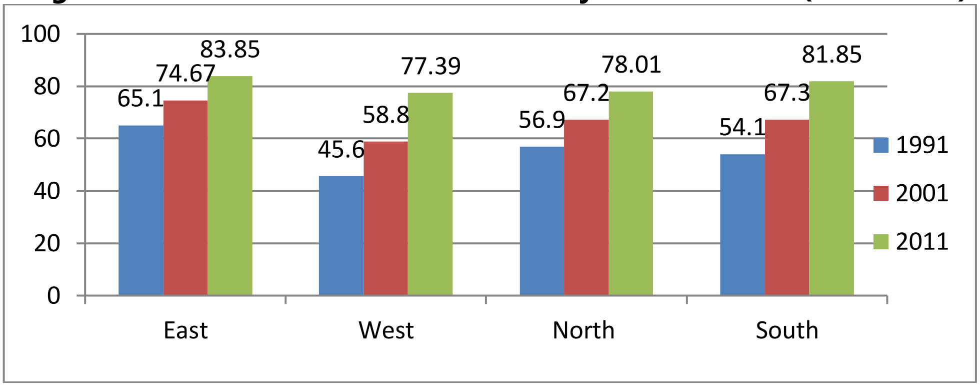


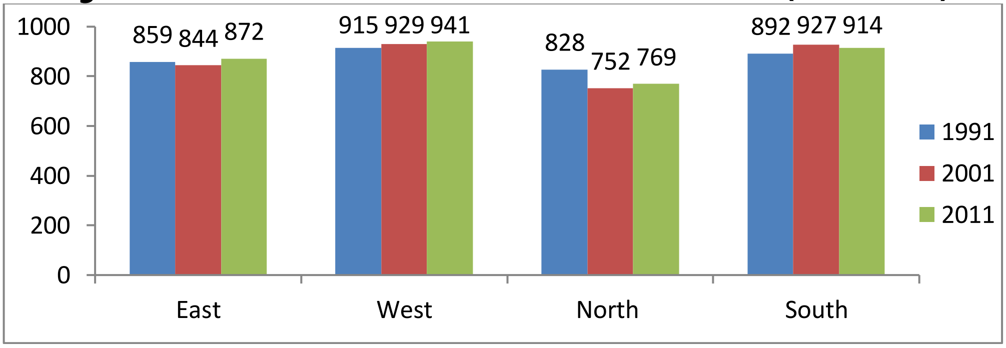
Fig.3 and Fig 4 showed the changing pattern of literacy and sex ratio in the four districts of Sikkim over the last three decades. East district had the highest literacy rate in all the three decades followed by South district. West district had the lowest literacy rate as compared with the three districts during periods.

Fig. 3 District wise distribution of literacy rate in Sikkim (1991-2011)



West district had the highest average sex ratio and followed by South district. It was lowest in North district during 1991 to 2011.

Fig. 4 District wise distribution of sex ratio in Sikkim (1991-2011)



Objective: The objective of the present study is to determine the correlation between the literacy and sex ratio in Sikkim, India.

2. Methodology

The present study was based on secondary data collected from the report of Census of India, 2011 published by the Registrar General, Government of India.

Definition of Correlation: Whenever two variables are so related that an increase in the one is accompanied by an increase or decrease in the other, then the two variables are said to be correlated. The number showing the degree of extent to which the two variables are related to each other is called the correlation coefficient. It is denoted by the symbol r or greek letter ρ. The value of correlation coefficient has a range of -1.0 to +1. Spearman's Rank

Correlation is a measure of the relationship between two variables using the ranked data. Spearman's rank correlation is the number r, given by the formula

$$r = 1 - \frac{6 \Sigma d^2}{n (n^2 - 1)}$$

Where d is the difference of the corresponding ranks and n is the number of pair's observation. In the present study, literacy is considered as an independent variable, while sex ratio is considered as a dependent variable.

3. Results

Table 1 showed the literacy and sex ratio of India, Sikkim and its four districts as per Census 2011. According to 2011 census, the literacy rate in the state was higher than the national figure. All the 4 districts had higher literacy higher than the national literacy rate. Average literacy rate higher than state were East and South district. It was 83.8% and 81.8% respectively. The lowest literacy rate was in West district (77.4%) followed by North district (78.0%). Even though the state' sex ratio was lower than the national' sex ratio, West district had highest sex ratio (942 females per 1000 male) among the four districts of the state. The second heighted sex ratio was in South district (915 females per 1000 males) followed by East district (873 females per 1000 males). North District had the lowest sex ratio (767 females per 1000 males) in the state in 2011.

Table 1: Literacy rate and Sex Ratio in India, Sikkim and its 4 districts in 2011

	Literacy rate	Sex Ratio
India	74.04	940
State	81.42	889
District of State		
East	83.8	873
West	77.4	942
North	78.0	767
South	81.4	915

# Source: Census of India, Sikkim, 2011

Table 2: Correlation between Literacy rate and Sex Ratio in 4 districts of Sikkim as per 2011 Census

Name of district	Literacy	Sex Ratio	R_1 (literacy)	R_2 (Sex Ratio)	d= $(R_1 - R_2)$	d^2
East	83.8	873	1	3	-2	4
West	77.4	942	4	1	3	9
North	78.0	767	3	4	-1	1
South	81.4	915	2	2	0	0
						14

Source: Census of India, Sikkim, 2011

Using Spearman's rank correlation method, we get

$$r = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

$$= 1 – \frac{6*14}{4\ (4^2\text{-}1)}$$

$$= 1 – \frac{84}{4\ (16\text{-}1)}$$

$$= 1 – \frac{84}{4 * 15}$$

$$= 1 - \frac{84}{60}$$

$$= 1\text{-}\ 1.4$$

$$= \text{-}\ 0.40$$

From the above calculation, the value of Spearman’s Rank Correlation coefficient was found to be -0.40. The minus sign indicate the direction of the relationship between literacy and sex ratio and the value 0.40 showed that the strength of the relationship was low. It means that there were low negative correlations between literacy and sex ratio in the state. We can interpret in another way that with the increase in literacy, the sex ratio decreases, or with the decrease in literacy, the sex ratio increases. Both the two variables are of opposite of each other. We can also say that both literacy and sex ratio has an indirect correlation. Hence, from the above result we concluded that literacy and sex ratio had a low negative correlation between them but it was insignificant (r = - 0.40, P=0.60).

4. Discussion

We know that Correlation measure the linear relationship between two or more quantitative variables. Here, Spearman's rank correlation method was applied by assigning ranks to the two variables: literacy rate and sex ratio. Clearly, the coefficient of Correlations (r) was found to be negative and it was 0.40. This meant that literacy and sex ratio are opposite of each other. So, it is a case of low degree of negative correlation. This shows that with the increase in literacy, the sex ratio decreases or decrease in literacy, the sex ratio increases. The finding of the present study revealed that literacy and sex ratio were negatively correlated. But it was not statistically significant (P=0.60). A study conducted in Gujarat concluded that there was negative correlation (-0.52) between literacy and sex ratio in their study. In a study conducted in Maharashtra to find out the relationship between literacy and sex ratio using Spearmen’s Rank Correlation method revealed that the literacy and sex ratio had positive but insignificant correlation in Maharashtra (r = 0.07). A study conducted in Rajasthan also revealed that there was a negative correlation between literacy and sex ratio in Rajasthan. The correlation between literacy and sex ratio was found in low degree and negative correlation (r = - 0.37) which was similar with the result of the present study. However, a study conducted in Kerala, having the highest literacy rate in India found that there was a strong and positive correlation (r=.66) between literacy rate and sex ratio in the state. However, a study conducted in Meghalaya, which is one of the 8 states of northeast India found that literacy rate and sex ratio was independent of each other (r = 0.0). This meant that the change in one variable that is literacy was not followed by changes in another variable that is sex ratio. This showed that literacy and sex ratio was not correlated in their study.

5. Conclusion

As per census 2011, Sikkim had literacy rate of 81.42% which is higher than the national literacy rate 74.04%. Literacy rate of East Sikkim and South district were higher than the state literacy rate. It was 83.85% in East and 81.85% in South district. West district had the lowest literacy rate (77.39%) in Sikkim. On the other hand, West district having the lowest literacy rate was having the highest average sex ratio (941 females per 1000 males) in the state. North district was having the lowest sex ratio, but it stood third in term of the ranking of literacy rate in 2011. East district stood in third ranked having the lowest sex ratio in the state even though the literacy rate was highest ac compared with other three districts of the state. In census 2011, South district stood in second ranked in both the literacy rate and sex ratio in the state. The study concluded that when the sex ratio was highest, the literacy was lowest, when the literacy was highest, the sex ratio was declined. This meant that increasing the literacy rate would not make an increase in sex ratio in the population. Applying the Spearman's rank correlation method, it was observed that literacy and sex ratio were having a low degree negative correlation of each other. In other words, an indirect relationship between literacy and sex ratio was observed in the present study.

References

1. Directorate of Census Operations, Sikkim. Provisional Population Totals, Part 1 of 2011, Sikkim, Series 12, Gangtok, Sikkim.
2. Birth and Deaths Cell, Department of Health Care, Human Services and Family Welfare, Government of Sikkim. Annual report on the working of the registration of births and deaths act, 1969, for the year 2012.
3. Jasim, H.R. (2017). Correlation between literacy rate and sex ratio in Thiruvananthapuram district: a geographical study*. International Journal of Applied and Pure Science and Agriculture*, 3 (7), 12 – 16.
4. Desai, H., & Oza, V. (2016). A study on correlation between literacy and sex ratio in Gujarat. *Abhinav International Monthly Refereed Journal of Research in Management & Technology,* 5 (8), 37-42.
5. Chaudhari, S.R., & Ahire, R.C. (2015). Correlation between literacy rate and sex ratio in Maharashtra: a spatial analysis. *Indian Streams Research Journal*, 5 (8), 1-7.
6. Bhat, F.A., & Manzoor, S. (2015). Falling Sex ratio in Jammu and Kashmir: Trends, Determinants and Consequences. *International journal of Sociology and Anthropology,* 7(3), 64 – 67.
7. Yadav, N. (2015). Correlation between literacy and sex ratio in Rajasthan: A geographical analysis. *International Journal of Engineering Development and Research*, 3 (4), 1108-1111.

Wavelet methods in statistics: Application in forecasting exchange rate volatility

Chiraz KARAMTI [1], Aida KAMMOUN [1], Ahmed TRABELSI [2]

[1] Higher Institute of Business Administration (ISAAS), Sfax, Tunisia

[2] PhD, Tunis University.

Abstract

Real effective exchange rate (REER) is a useful summary indicator of essential economic information. However, the predictability of exchange rate movements is still a major puzzle in international finance and even in official statistics because of the conventional models' inability to produce accurate forecasts of the exchange rate volatility. This article suggests a novel technique for modelling and forecasting EURO/USD exchange rate in time and frequency, based on Wavelet transforms and GARCH models with high frequency return data. The purpose is to check whether the performance of these models is uniform along different frequencies or whether it is driven by certain frequencies. The main finding of this work is that the predictability of exchange rates varies along the different frequencies.

Keywords

Wavelet analysis; forecasts, GARCH models, exchange rate.

1. Introduction

In the context of globalization and economic integration, each country is obliged to maintain commercial relations with several other countries. Exchange rates of local money with the currencies of many partner and competitor countries and their variations affect the volume of trade with international markets. An aggregate indicator measuring the evolution of the exchange rate against a set of other currencies can be constructed; it combines various bilateral rates into a single indicator which is Real Effective Exchange Rate (REER). The REER is the nominal effective exchange rate divided by a price deflator or index of costs (Schmitz et al. 2012). It aims to assess a country's competitiveness (in terms of price or cost) relative to its principal partners and competitors in international markets. Referring to the Bank for International Settlements (2019), "Real effective exchange rates are calculated as weighted averages of bilateral exchange rates adjusted by relative consumer prices". Changes in competitiveness depend on exchange rate variations as well as on inflation trend. An increase of the index means competitiveness deterioration. To maintain its competitiveness, each country must maintain constant or decrease its REER. Or, this rate depends on exchange rates and both domestic and world market inflation. Faced with a free floating exchange rate system, and in order to maintain stable REERs, countries must control inflation rates through the adoption of appropriate monetary policies. In this context, and as part of a preventive approach, the prediction of exchange rates helps authorities to plan monetary policies to be adopted in the future in order to maintain and strengthen country competitiveness. However, exchange rate prediction has long been recognized as a puzzle. Many models have been developed by theorists and analysis in order to determine the exchange rate. This paper tries to develop a new approach for exchange rate prevision using the wavelet method.

Wavelet theory has provided statisticians with powerful new techniques for nonparametric inference by combining conventional forecasting methods with insights gained from applied signal analysis. Only recently few research (Tan et al. (2010); Ismail et al., 2016) used this novel forecasting method based on wavelet transform approaches combined with ARIMA and GARCH models, and it was compared with some of the most recently published price forecasting techniques. The comparative results clearly showed that the proposed forecasting method was far more accurate than the other forecasting method.

This article suggests using this novel technique for forecasting the exchange rate EURO/USD, based on Wavelet transforms and ARIMA/GARCH model. High frequency return data (5 minutes) from 01/05/2017 until 12/12/2016 with a total sample of 44508 observations is used for this study. Results show that the predictability of exchange rates varies along the different frequencies. Besides, there is significant improvement in predictability as we move from a short forecast horizon to a long forecast horizon. The omission of these features in forecasting the REER indicators for the euro may have serious consequences since the REER is used as indicator for monetary and exchange rate policies, and as policy makers may use it to forecast current account and trade balance in the country.

2. Methodology

Based on prior evidence of unparallel performance of wavelet technique and following Tan et al. (2016) and Ismail et al. (2016) methodology, we implement the wavelet transformation in conjunction with Autoregressive Moving Average (ARMA) and the Exponential Generalized Autoregressive Conditional Heteroscedasticity (EGARCH) model to accurately predict exchange rates.

2.1 Wavelet transform

Wavelet theory is a powerful mathematical tool for time series analysis. It provides a time-frequency representation of a time series $X(t)$ (in our study, $X(t)$ is the exchange rate return), and it can be used to analyze non-stationary time series, which are very common in finance, given the continuous presence of abrupt changes and volatility. Recently, this methodology has received great interest in the financial literature (Rua and Nunes, 2009; He et al., 2009; Masih et al., 2010; Jammazi, 2012). Researchers, have consistently endorsed that the wavelet tool is superior to the conventional statistical ones that have been used to decompose, filter and denoise signals. Wavelet transformations has the capacity to breakdown macroeconomic variables into their scale parcels. According to the Mallat's (2001) theory, the original discrete time series $X(t)$ can be decomposed into a series of linearity independent approximation and detail signals by using wavelet transform. For that, the wavelet technique uses multiresolution analysis by which different frequencies are analyzed with different resolutions. There is a scaling function $\phi(t)$ (also called father wavelet) such that:

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j - k) \tag{1}$$

where the level *j* controls the degree of stretching of the function (the larger the *j*, the more stretched is the basis function); the smaller the scale, the higher the frequency of the decomposed series, and *k* is the parameter that controls the translation of the basis function. Assuming that the detail space, $\{W_j\}$, are orthogonal to each other, we can define a sequence $\{\psi_{j,k}(t)\}_k$(called mother wavelet) of orthonormal basis function that spans $L^2(\mathbb{R})$:

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \tag{2}$$

where wavelets are generated by shifts and stretches of the mother wavelet $\psi_{j,k}(t)$. Let $X(t)$ the original time series, we represent the multiresolution representation of $X(t)$ by:

$$\begin{aligned} X(t) &= \sum\nolimits_k s_{j,k}\phi_{j,k}(t) + \sum\nolimits_j \sum\nolimits_k d_{j,k}\psi_{j,k}(t) \\ &= S_J(t) + d_J(t) + d_{J-1}(t) + \cdots + d_1(t) \end{aligned} \tag{3}$$

The series $S_J(t)$ provides a smooth of original time series $X(t)$ (called also approximation) at levels *J* and captures the long term properties (i.e. the low-frequency dynamics), and the series $d_J(t)$ for $j = 1, \dots, J$ denote wavelet details and capture small variations (i.e. the higher-frequency characteristics) in the data over the entire period at each scale. The last expression (3) denotes the decomposition of $X(t)$ into orthogonal components at different resolutions and represents the so-called wavelet multiresolution analysis (MRA).

2.2 Wavelet-based EGARCH model

To model the time-varying dynamics of exchange rates, the wavelet details (d1-d7) are used as input in the ARMA-EGARCH model of Nelson (1991). Recall that the EGARCH model was developed to allow for asymmetric effects between positive and negative shocks on the conditional variance of future observations. For each detail j, the mean equation of the model is given by:

$$d_{jt} = \mu_t + \varepsilon_t = \mu_t + \sqrt{h_t}\eta_t \quad \eta_t \sim N(0,1) \tag{4}$$

The conditional variance equation (in logarithm) is formulated as follow:

$$\ln(h_t) = \omega + \alpha \left|\frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}}\right| + \gamma \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} + \beta \, ln \, h_{t-1} \tag{5}$$

where d_{tj} is the modified wavelet transform (MODWT) return series, h_{t-1} is the conditional variance or volatility at $t-1$, while β measures the extent to which a present volatility shock goes into the future volatility and $(\alpha+\beta)$ measures the rate at which this effect dies in the future. The parameter γ is the chief causative agent of the asymmetry in volatility. When $\gamma > 0$, a positive return shock increases volatility, and when $\gamma < 0$, a positive return shock reduces volatility. η_t is a standardized error. This implies that the leverage effect is exponential, rather than quadratic, and the forecasts of the conditional variance are guaranteed to be non-negative. To evaluate the accuracy of the models, two performance criteria such as RMAE and MSE. These criteria are given below:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(x_t - \hat{x}_t)^2}{T}} \quad and \quad MAE = \sum_{t=1}^{T}\left|\frac{x_t - \hat{x}_t}{T}\right|$$

Where x_t is the actual and $\hat{x}_t$ is the forecasted value of period t, and T is the number of total observations.

3. Empirical analysis

The data consists of average returns of the exchange rates between the European Euro vis-à-vis the US dollar. Five minutes observations cover the period from May 1, 2016 to December 12, 2017 and that makes a sample of 44508 observations. We split the full sample in two sub-periods, pre-Brexit and post-Brexit referendum to investigate the volatility dynamics in the presence of high uncertainty.

4. Result

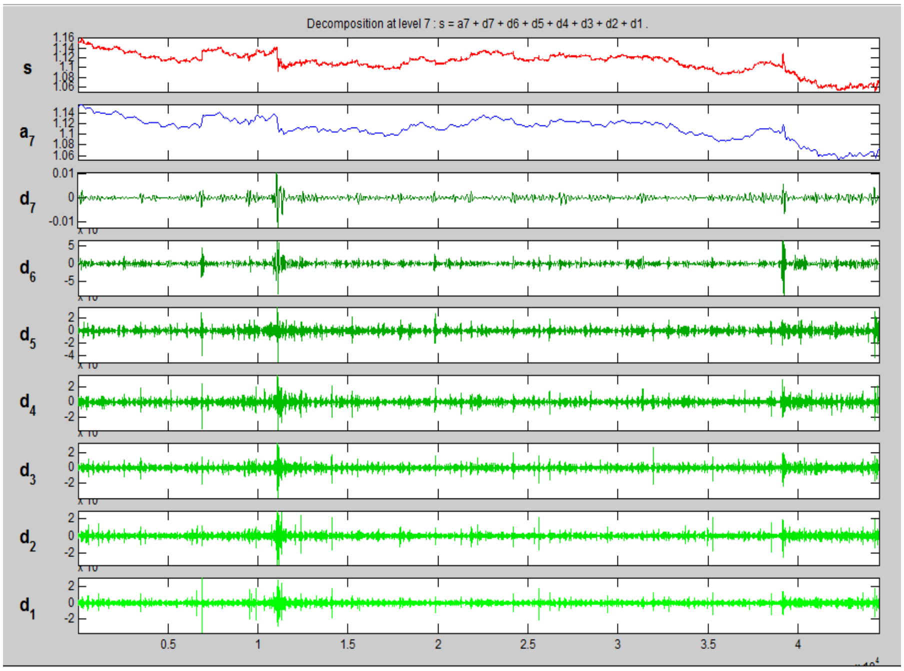
We used 5mn data and our decomposition goes to scale 7. The first level detail d1 represents the variations within [10;20] mn, while the next level details d2-d7 represent the variations within $\left[2^j, 2^{j+1}\right]$ (5 minutes) horizon (Table 1).

Tableau 1. Interpretation of time scales

Horizon $[2^j, 2^{j+1}]$ $\; j = 1, 2, \ldots, 7$		
Scale	Component	Frequency resolution
Scale 1	d1	10-20 mn
Scale 2	d2	20-40 mn
Scale 3	d3	40-80 mn
Scale 4	d4	80-160 mn
Scale 5	d5	160-320 mn
Scale 6	d6	320-640 mn
Scale 7	d7	640-1280 mn$\approx 1\, day$
Approximation	A7	>1280

The wavelet coefficients and scale coefficients of the exchange rate level time series derive a Mallat decomposition algorithm using Daubechies least asymmetric (LA) wavelet filter of length L=8 are shown in Fig. 1.

Fig 1. Decomposed wavelet sub-time series components (d) and Approximation (a) of crude oil



The outcomes of estimating the parameters of the model given in Eq. (5) using the MODWT transformed exchange rate (EUR/USD) return series are reported in Table 2[1].

[1] The MODWT-ARMA$_{(2,0)}$ model is the best fitting model for the mean equation according to the conventional information criteria (AIC and BIC). Given the amount of results, the mean equation

Tableau 1. Parameter estimates of wavelet based EGARCH (1,1) models

	Full sample			Before Brexit			After Brexit		
	α	γ	β	α	γ	β	α	γ	β
d_1	0.0017^{***}	0.0212^{***}	0.3253^{***}	-0.0048^{***}	0.0408^{***}	0.0203^{***}	-0.0422^{***}	0.0110^{***}	0.0165
d_2	0.0136^{***}	0.0235^{***}	0.2020^{***}	0.0818^{***}	-0.0047^{***}	0.0645	-0.0115^{***}	0.0360^{***}	0.1916^{***}
d_3	0.1719^{***}	0.0587^{***}	0.9264^{***}	0.0401^{***}	0.0140^{***}	0.0924	-0.0007	0.0121^{***}	0.0613
d_4	0.2768^{***}	-0.0001	0.9834^{***}	0.2798^{***}	-0.0004	0.9807^{***}	0.2758^{***}	0.0001	0.9741^{***}
d_5	0.2743^{***}	0.0011	0.9783^{***}	0.2936^{***}	-0.0012	0.9737^{***}	0.2686^{***}	0.0018	0.9696^{***}
d_6	0.2580^{***}	0.0004	0.9596^{***}	0.2058^{***}	0.0005	0.9656^{***}	0.2721^{***}	8.15E-05	0.9582^{***}
d_7	0.1156^{***}	-0.0061^{***}	0.3188^{***}	0.1937^{***}	-0.0004	0.9721^{***}	0.1866^{***}	-0.00003	0.9625^{***}

Although every uncertainty may not cause volatility, uncertainty about major events can result in volatile market. In our case, the results indicate that the effect of previous volatilities is high and stable before and after Brexit over the medium and long term. However, Brexit appears to have a significant negative impact on the currency market volatility and shows that the volatility of returns is higher before the Brexit and decreases at the post-Brexit period. Lots of uncertainty before the Brexit brought this increase in the volatility and realizing the outcome of the referendum and cutting the interest rate by the central bank resulted in investors asking less risk premium which led to the decrease in volatility. Before major events like Brexit, investors lose their trust to the central bank to be able to have policies that positively interfere with the market. As the event passes, the realization of the outcome appears to reduce the need for the risk premiums to be introduced to the market thus the implied volatility is reduced. Furthermore, the asymmetric effect in most wavelets after Brexit is positive and small indicating an unfavorable asymmetric reaction to good news increasing volatility more than bad news. However, this effect disappears in the medium to long run (sacles 4-7). The most striking evidence from the above results is that no general pattern can be found since volatility at each scale have its own dynamics. However, these results in general indicate the dominance of low frequency elements (high scales) in the exchange rates market.

The Akaike Information Criterion (AIC) is used to decide the best fitting model for each MOWT exchange series and the values are presented in Table 3. The forecasting ability of EGARCH models for different scales was judged

results based on applying the iterative Box-Jenkins procedure, are not reported, only the conditional variance part is presented.

on the basis of root mean square error (RMSE) and absolute mean error (AME) are also reported in Table 3. When looking at the two sub periods depicted in Table 3, we can see that on average the forecast error of the three EGARCH models decreased significantly in higher scales (long-run) compared to lower scales (short-run). Specially, after Brexit, it is obvious that the uncertainty is higher, so a longer-term forecast seems more reliable (d6 and d7). It follows that a forecast at a period of high volatility is better in the long run and that the accuracy of the euro/dollar exchange rate forecast depends on the frequency of the data.

Tableau 1. The AIC, RMSE and MAE comparisons for different EGARCH models

	Full sample			Before Brexit			After Brexit		
	α	γ	β	α	γ	β	α	γ	β
d_1	0.0017***	0.0212***	0.3253***	-0.0048***	0.0408***	0.0203***	-0.0422***	0.0110***	0.0165
d_2	0.0136***	0.0235***	0.2020***	0.0818***	-0.0047***	0.0645	-0.0115***	0.0360***	0.1916***
d_3	0.1719***	0.0587***	0.9264***	0.0401***	0.0140***	0.0924	-0.0007	0.0121***	0.0613
d_4	0.2768***	-0.0001	0.9834***	0.2798***	-0.0004	0.9807***	0.2758***	0.0001	0.9741***
d_5	0.2743***	0.0011	0.9783***	0.2936***	-0.0012	0.9737***	0.2686***	0.0018	0.9696***
d_6	0.2580***	0.0004	0.9596***	0.2058***	0.0005	0.9656***	0.2721***	8.15E-05	0.9582***
d_7	0.1156***	-0.0061***	0.3188***	0.1937***	-0.0004	0.9721***	0.1866***	-0.00003	0.9625***

5. Discussion and Conclusion

This paper proposes the application of the Wavelet-EGARCH technique for the modelling of euro/dollar exchange rate series. The MODWT-EGARCH models are obtained by combining two methods, an EGARCH model and discrete wavelet transforms. The main objective is to verify if the frequency of the data would have an impact on the reliability of the forecasts. Accordingly, the series was decomposed at 7 decomposition levels (10mn until one day). The sum of the effective details and the approximation component were used as inputs to the EGARCH model. The performance of the proposed MODWT-EGARCH models was compared to forecasting using regular criteria. Comparison of the results indicated that the MODWT-EGARCH model was substantially more accurate in higher scales, i.e. the medium and long-terms. This study shows that wavelet transform technique, joined with GARCH models, are particularly useful in forecasting foreign exchange volatility in periods of either low or high volatility. Indeed, forecasts seem less reliable in periods characterized by greater uncertainty, in our case due to the Brexit announcement. Thus, over a period of high volatility, a long-term scale is found to be the most effective in yielding an accurate forecast, whereas before Brexit, the best forecast is given by medium-term wavelets. In all cases a short-term forecast has proved unreliable.

Finally, the frequency component affects the predictive performance of various models at both short horizons and long horizons. The volatility dynamics are not uniform across scales. Accordingly, as might have been expected, wavelets provide a degree of refinement and flexibility not available using conventional forecasting methods. With wavelets, one can choose the scale at which the forecast is to be made. As evidenced by our results, each scale level has to be treated as a separate series for forecasting purposes. These findings suggest that *forecasting* is more delicate than has been recognized so far and that forecasts need to be expressed conditional on the relevant scales (Gallegati and Semmler, 2014; Yousefi et al., 2005).

References

1. Bank for International Settlements, Real Broad Effective Exchange Rate for Euro Area [RBXMBIS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/RBXMBIS, January 30, 2019.
2. Gallegati, M. and Semmler, W. (eds.), *Wavelet Applications in Economics and Finance*, Dynamic Modeling and Econometrics in Economics and Finance 20, (2014) DOI 10.1007/978-3-319-07061-2__1. Springer International Publishing Switzerland.
3. Gallegati, M., Ramsey, JB., Semmler, W. (2013). Time scale analysis of interest rate spreads and output using wavelets. Axioms 2:182–207.
4. He, K., Xie, C., Chen, S., Lai, K.K., (2009). Estimating var in crude oil market: A novel multi-scale non-linear ensemble approach incorporating wavelet analysis and neural network. Neurocomputing, 72, 3428- 3438.
5. Ismail, M.T., Audu, B., Tumala, M.M. (2016). Volatility forecasting with the wavelet transformation algorithm GARCH model: Evidence from African stock markets. *The Journal of Finance and Data Science*, 2(2), 125-135.
6. Jammazi, R., 2012. Oil shock transmission to stock market returns: Wavelet-multivariate markov switching garch approach. *Energy*, 37, 430-454.
7. Mallat. S., (2001). A Wavelet Tour of Signal. *Processing.* Academic Press, San Diego*.*
8. Marchand-Blanchet, F. (1998). Une approche de la compétitivité de la zone euro : le taux de change effectif de l’euro, Bulletin De La Banque De France, n°60, 103.
9. Masih, M., Alzahrani, M., Al Titi, O., (2010). Systematic risk and time scales: New evidence from an application of wavelet approach to the emerging Gulf stock markets. *International Review of Financial Analysis*, 19, 10-18.
10. Rua, A., Nunes, L.C., (2009). International comovement of stock market returns: A wavelet analysis. *Journal of Empirical Finance*, 16, 632-639.
11. Schmitz, M., De Clercq, M., Fidora, M., Lauro B., and Pinheiro, C. (2012). Revisiting The Effective Exchange Rates Of The Euro. European Central Bank Occasional Paper series, No 134, June 2012.

12. Tan Z, Zhang J, Wang J, Xu J. (2010). Day-ahead electricity price forecasting using wavelet transform combined with ARIMA and GARCH models. *Appl Energy*, 87(11),3606-3610.
13. Yousefi S, Weinreich I, Reinarz D (2005) Wavelet-based prediction of oil prices. Chaos Solitons Fractals 25, 265–275
14. Yousr ZRIBI, (2017). Appréciation de la compétitivité prix de l'économie tunisienne. Notes et analyses de l'ITCEQ N° 52.

A local dynamic multipollutant air quality indicator

Giuliana Passamani, Paola Masotti, Matteo Tomaselli
Department of Economics and Management, University of Trento, Trento

Abstract

Given the worldwide concern on urban air pollution and its impact on public health and damage to the environment, we aim to contribute to a better reporting of information about air quality by suggesting a methodological procedure leading to the estimation of a local pollution indicator. The suggested procedure combines daily measurements of air pollutants with the observed meteorological conditions, taking into account also the effects of lagged pollution. The advantage of the dynamic factor model used for the empirical analysis is that we can consider the dynamics of local air pollution as a main determinant, together with the weather conditions, of what we actually observe, and we can use the same estimated model for forecasting future air pollution, given the meteorological predictions. The application regards pollution data collected at some monitoring sites in the alpine province of Trento.

Keywords

Air pollution; Air quality indicator; Dynamic-factor model

1. Introduction

It's well known that air, land and water pollution harms human health and damages the environment. For these reasons, since the end of the past century, environmental agencies and organizations, across the world, have been working with the aim of reducing pollution and improving the quality of the environment, through increasing information on the consequences of pollution and working towards the introduction of environmental laws and directives, in order to establish new and appropriate regulations. If we focus our attention just on-air pollution, properly measuring it is the first and main objective when we want to evaluate air quality in a particular region. To this aim, the EU legislation defines evaluation and management methods for air quality and set the standards for the monitoring networks. In the specialized literature an important effort has been done towards quantifying air pollution and observing its evolution, in particular, a substantial number of Air Quality Indices (AQIs) have been proposed with the aim of combining observations on a variety of pollutants at multiple monitoring sites, and giving rise to a simple indicator summarizing air quality, as in Bruno and Cocchi (2002), where they obtain a synthetic value by means of hierarchical median-maximum aggregation processes, or as in Plaia, Di Salvo, Ruggeri and Agrò (2013), where they suggest an index based first on a spatial aggregation and then on a pollutants synthesis.

Our proposal is to provide a synthetic measure of air pollution in a given town/place by means of a stochastic dynamic-factor model where we combine the daily measurements of air pollutants with the meteorological conditions, taking into account also the pollution measure observed the day before. That is, we suggest a multipollutant synthesis based on human-caused and natural sources emissions levels and their relationship with meteorological factors and with the lagged pollution level. Giving that meteorological conditions can change within a distance of some miles, the daily measures for each pollutant cannot simply be aggregated over different monitoring sites, but, first of all, we must analyse and aggregate different pollutants at the same monitoring site for which we have also the corresponding observations on meteorological variables. Once the synthetic air pollution measure has been calculated in a given place, an aggregated AQI could be suggested for an entire area of interest. After these premises we state that the focus in the present paper is the proposal of a statistical methodology aiming at evaluating the different ambient air qualities in the Province of Trento that covers an almost entirely mountain area located in the northern part of Italy. The data set is made up of the time series observations relative to three pollutants, particulate matter 10 (PM10), nitrogen dioxide (NO_2) and ozone (O_3): these are now generally recognized as the three main pollutants that most significantly affect human health[1]. The pollutants are observed at different monitoring sites located in traffic and non-traffic areas. Time series observations of some meteorological variables are also available for the same monitoring sites. The empirical analysis is based on a two years' period, 2014-2015.

The plan of the paper is the following. In Section 2 we discuss the methodological approach and the dynamic-factor model adopted for analysing the available daily time series dataset. In Section 3 we describe the results in terms of the estimated pollution indicator for each site and in Section 4 we conclude the paper and outline possible lines for further research.

2. Methodology

The principle which is at the basis of a dynamic-factor model is that few unobservable dynamic factors drive the co-movements observed in a higher dimensional vector of endogenous time series variables which can also be affected by exogenous variables, as well as by a vector of mean-zero idiosyncratic disturbances. For the purpose of our approach in which we assume the existence of a single dynamic factor underlying the observed

[1] European Environment Agency: https://www.eea.europa.eu/themes/air/intro

pollutants, the stochastic dynamic-factor model that represents the framework for the empirical analysis has the following form[2]:

(1) $$\mathbf{y}_t = \gamma f_t + \boldsymbol{\beta}' \mathbf{x}\ \mathbf{u}_t + {}_t, \quad f_t = \alpha f_{t-1} + \nu_t, \quad \mathbf{u}_t = \boldsymbol{\Phi} \mathbf{u}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{y}_t$ denotes the $(k\times 1)$ vector of observed endogenous variables, k being the number of pollutants analysed, and $\mathbf{x}_t$ the $(m\times 1)$ vector of observed exogenous variables, m being the number of meteorological variables taken into account; f_t denotes the unobserved common factor that represents a synthetic measure of air pollution; $\mathbf{u}_t$ and ε_t are $(k\times 1)$ disturbance vectors and ν_t is a scalar disturbance. The unknown parameters are $\boldsymbol{\gamma}$, a $(k\times 1)$ vector containing the unknown dynamic factor loadings, α, the autoregressive factor parameter and $\boldsymbol{\Phi}$, a $(k\ k\times\ )$ matrix of autoregressive parameters for the disturbances. These idiosyncratic disturbances are assumed to be uncorrelated with the factor disturbance at all leads and lags. Model (1) can thus be considered as a stochastic dynamic model with vector autoregressive errors, where the conditional mean of the unobserved factor is assumed to vary over time according to an AR(1) model.

The first issue at hand is to estimate the factor. With the further assumptions that the disturbances in the vector ε_t are i.i.d. N(0, σ_i^2), $i=1,\ldots,\ ,k$, the estimation is performed using a maximum likelihood approach, implemented by writing the model in state space form and by using a stationary Kalman filter and diffuse De Jong Kalman filter for calculating the log likelihood. This approach unites the statistical efficiency of the state space approach with the robustness and convenience of the principal components approach: state space/Kalman filter estimates can produce substantial improvements in estimates of the factors and common components if the time behaviour of the common component is persistent, as in the present case. An advantage of this parametric state space formulation is that it can handle data irregularities[3]. Once the model has been estimated, we use it in order to predict the unobserved common variable $\hat{f}_t$. The prediction method estimates the states at each time by a Kalman smoother and using all the sample information.

3. Results

The data set consists of daily time series observations on three pollutants: PM10, NO_2 and O_3. The observations on PM10 correspond to the 24-hour running means, while the observations on NO_2 and O_3 are the maximum

[2] A similar model was used by Fontanella et al. (2007) for the analysis of environmental pollution in the Milan district, following the work of Forni et al. (2000).

[3] For a detailed presentation of the methodology, we advise to refer to Stock and Watson (2011).

hourly concentrations within the day. They are obtained from continuous measurements of each pollutant and their unit of measurement is µg/m^3. The three pollutants are characterized by a clear seasonal variation, with PM10 and NO_2 positively correlated and both negatively correlated with O_3. In addition, we have the availability of hourly observations on some meteorological variables of which we retain that wind speed and precipitations can have a significant impact on air pollution. Therefore, we take into account wind speed (m/s) and precipitations (mm), measured as daily means.

The seven monitoring sites are:

(i) Trento Parco S. Chiara, an urban non-traffic area in the main town;

(ii) Rovereto Largo Posta, another non-traffic area in the second main town of the province; (iii) Borgo Valsugana and (iv) Riva del Garda as sub-urban sites, but with a quite different climate as the latter is mitigated by the winds coming from lake Garda, the largest Italian lake; (v) Piana Rotaliana, as a rural non-traffic area; (vi) Trento VBZ, which is an urban traffic area; (vii) Monte Gaza, a nontraffic mountain area. Due to the fact that for Trento VBZ we do not have any measurement on ozone and Monte Gaza is a purely mountain area far from any town, we restrict our attention to the other five monitoring sites.

Given the methodological approach adopted for the empirical analysis, the air pollutant data need not to be pre-processed in order to standardize pollutants with different orders of magnitude, as could be done using linear interpolation, particularly recommended by US Environmental Protection Agency (2006). The standardization procedure must be applied when the purpose of the analysis is to calculate air quality indices by means of some aggregation functions, as in a large part of the literature has been done (see, inter alia, Murena, 2004, Plaia et al., 2013, Li et al., 2014).

For each monitoring site we have estimated model (1). The estimates show that the unobserved common factor is quite persistent and it's a significant predictor of the observed variables, that is the factor loadings are highly significant and they show a positive sign, indicating that the three pollutants contribute together to determining the pollution indicator.

The two meteorological variables that we assume to determine the level of pollutants, show different effects: wind affects negatively and significantly PM10 and NO_2, but positively and significantly O_3 in any monitoring station; precipitations affect significantly and negatively O_3 in Trento and Rovereto, significantly and negatively O_3 and significantly and positively NO_2 in Piana Rotaliana, while they have no significant effect in Borgo Valsugana, and significant and negative effects on PM10 and NO_2 and significant and positive effect on O_3 in Riva del Garda. Briefly, wind has a clear impact everywhere, while precipitations have not.

In order to have a better understanding of the results obtained with the suggested methodological procedure, we make use of the following graphs for two monitoring sites, Trento PSC and Piana Rotaliana, but similar graphs are available also for the other sites.

In Fig. 1 and Fig. 2 we represent with a thick line the estimated common dynamic factor. In the same graph we can observe the standardized measurements of the three pollutants whose linear combination, together with the weather conditions, determine the pollution indicator. Given the differences in the order of magnitude of the three pollutants, we have to use standardized data obtained using the linear interpolation method if we want to compare the estimated pollution indicator with the observed air pollutants. It's rather surprising to note how the procedure has been able to summarize pretty well, in both cases, the daily observations with a single smooth indicator.

4. Discussion and Conclusion

In Fig. 3 we represent the estimated air pollution indicators for the five monitoring stations in one graph. As expected, the two urban monitoring sites, Trento and Rovereto, show higher air pollution levels with respect to the other sites. Borgo Valsugana seems to suffer from pollution less than the two urban areas, though, according to official data recorder by APPA, it is affected by pollution even in a worse manner than the others. APPA is the provincial environment protection agency that collect the data on pollutants and calculate the air quality indices in accordance with the European and national directives. Those AQIs are then transmitted to the European Environmental Agency that Figure 1: The estimated pollution indicator and the standardized pollutants for Trento PSC

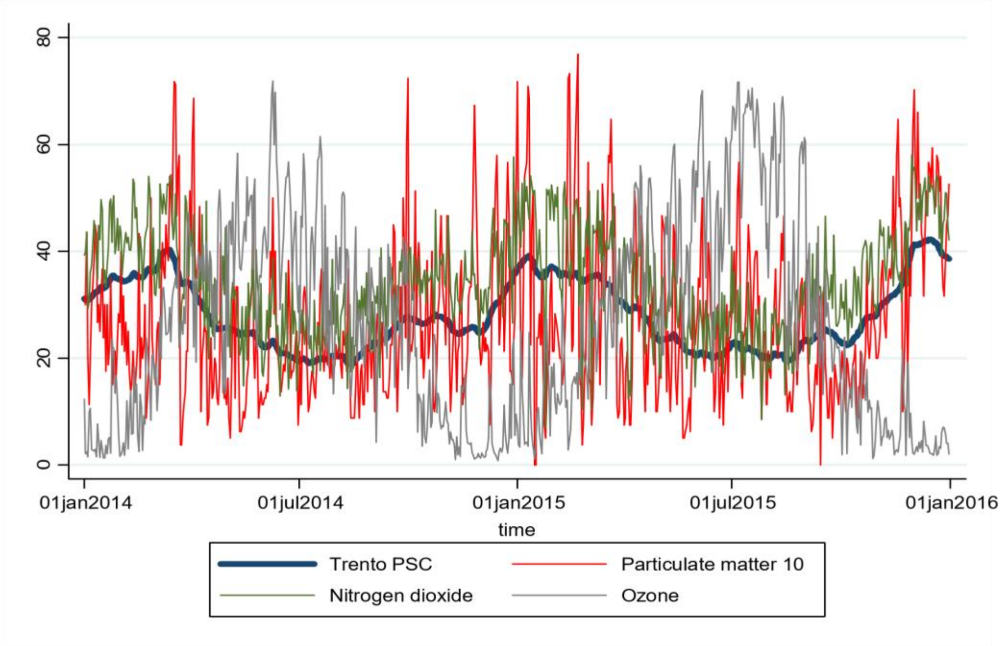
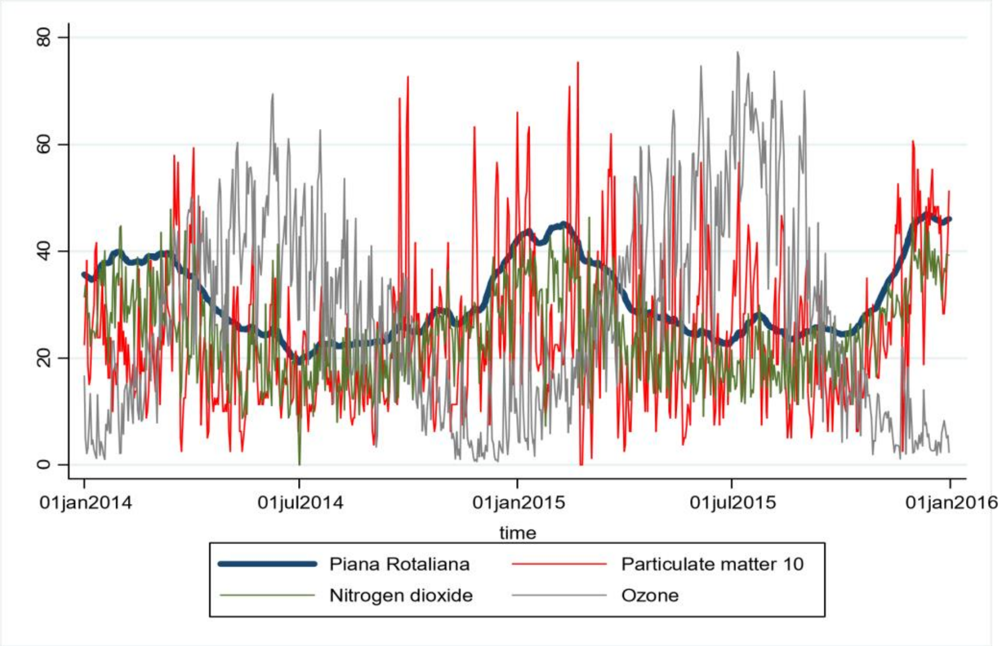


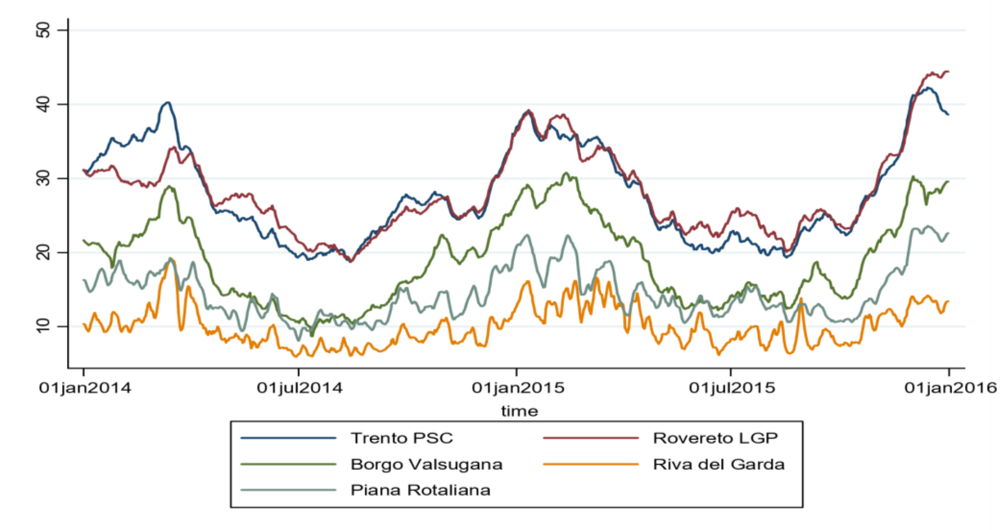
Figure 2: The estimated pollution indicator and the standardized pollutants for Piana Rotaliana



make them publish on its website[4]. If we look at those indices, we can see that Borgo Valsugana is worse affected mainly by particulate matter of a smaller dimension, PM2.5, primarily due to road transport.

[4] EEA: https://www.eea.europa.eu/themes/air/air-quality-index

Fig. 3: The estimated pollution indicators



Data on PM2.5 are not available for all the sites and for this reason this pollutant has not been taken into consideration in the empirical analysis. For sure, better and interesting results could be obtained if we had data even on sulphur dioxide, SO_2, and carbon oxide, CO, as well as on PM2.5. In any case, the purpose of this paper is principally the proposal of a statistical procedure to be applied for analysing pollution data within a dynamic model, and not just to calculate air quality indices. The advantage of the dynamic-factor model used for the empirical analysis has been shown and further research could be done, particularly in the direction of being able to better forecasting future air pollution, given the predicted weather conditions. Another appealing further issue would be the suggestion of a procedure for combining the estimated air pollution indicators in just a single one. This could be of particular interest especially in the case we want to synthetize in a single measure the pollution data collected by means of several monitoring sites covering a large area with similar characteristics, like a metropolitan area. This last issue would not be meaningful for the dataset analysed in this paper, given the spatial dispersion across a mountain province of the monitoring stations from which our data are collected.

References

1. Bruno, F. and Cocchi, D. (2002). A unified strategy for building simple air quality indices. Environmetrics, 13, 243-261.
2. Fontanella, L., Ippoliti, L. and Valentini, P. (2007). Environmental Pollution Analysis by Dynamic Structural Equation Models. Environmetrics, 18, 265-83.
3. Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). The Generalized Dynamic-Factor Model: Identification and Estimation. The Review of Economics and Statistics, 82, 540-54.
4. Li, L., Qian, J., Ou, CQ., Zhou, YX., Guo, C. and Guo, Y. (2014). Spatial and temporal analysis of Air Pollution Index and its timescale-dependent relationship with meteorological factors in Guangzhou, China, 2001–2011, Environmental Pollution, 10, 75-81.
5. Murena, F. (2004). Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples. Atmospheric Environment, 38, 6195-6202.
6. Plaia, A., Di Salvo, F., Ruggieri, M. and Agrò, G. (2013). A Multisite-Multipollutant Air Quality Index. Atmospheric Environment, 70, 387-391.
7. Stock, J.H. and Watson, M.W. (2011). Dynamic factor models. Chapter 2 in Clements, M. J. and
8. Hendry, D. F. (Eds.), Oxford Handbook of Economic Forecasting, OUP, 35-59
9. US Environmental Protection Agency (2006). Guidelines for the Reporting of Daily Air Quality – the Air Quality Index (AQI). U.S. EPA Office of Air Quality Planning and Standards Research, Triangle Park, North Carolina.

Transmission network analysis of infectious disease in Portugal

M Lutfor Rahman[1], Carla Nunes[2]

[1] Institute of Statistical Research & Training, University of Dhaka, Bangladesh

[2] National School of Public Health, NOVA University of Lisbon, Portugal

Abstract

Proper identification of the key factors responsible for an infectious disease transmission would contribute to control rapid spread of the disease. This study aims to model transmission network of an infectious disease focused on a case-study of tuberculosis network in Portugal. Characteristics of index-patients (confirmed TB diseased), contacts and exposition were considered. There were 495 participants in the study, considering 69 were index patients (confirmed TB diseased) and 426 identified contacts. Identified contacts were screened at the TB centre and other relevant information was collected at the time of screening. Some different methods were applied in a sequential order: simple measures of association, binary logistic regression, hierarchical logistic regression and network logistic regression. These different models have different assumptions, namely independency between observations and index-contact relationship of the network, and the results were compared. Based on binary logistic regression only the factors symptomatic period of index patients, sputum, age of contacts, and exposure duration were responsible for contacts being infected. The hierarchical logistic regression selected the factors age, diabetes mellitus, sleeping together as important variables for TB infection. The network regression reveals that the older age of the contacts and household interaction are the risk factors of TB infection. The network logistic regression was better to model the TB transmission using a lower number of variables, being a promising tool in the analysis of infectious disease with lower methodological limitations.

Keywords

Tuberculosis; transmission; network logistic regression; risk factors

1. Introduction

Infectious disease, also known as communicable or transmissible disease, is caused by infectious agents including viruses, bacteria, fungi, parasites, and other macro parasites. Emerging infectious diseases continue to expose national and global unpreparedness for prevention and control of disease outbreaks. Tuberculosis is an infectious disease usually caused by the Mycobacterium, thus the disease known as Mycobacterium tuberculosis or M. Tuberculosis. According to World Health Organization (WHO) tuberculosis which is one of the ten leading infectious diseases causes deaths of 1.3 million people around the globe in 2016. It is imperative to identify the leading factors responsible for spreading the TB infection as well as to quantify the likelihood of normal individuals being infected in a natural environment, particularly taking into account the TB networks.

Binary logistic regression assumes independency between observations. When there is nonindependence in the data, namely due to the presence of a hierarchical structure, logistic mixed models should be used to model binary outcomes (Mendes 2013). Is relatively frequent, but not theoretically adequate, the use of binary logistic regression even if in the presence of a hierarchical structure. Mendes (2013) used logistic model for interpreting TB data and to relate TB infection with the risk factors, but their model does not consider the dependency structure of the data.

Additionally, a more challenging structure can be presented, considering not only a non-independency between observations but also the index-contact network relationship. In this case, a more complex statistical approach but with lower methodological limitations could be based on network analysis, popularly known as social network analysis (SNA), rooted in the mathematical graph theory. The SNA has established itself as a powerful tool for studying structure and complex dynamics of systems (Lusseau et.al. 2008). It has been employed across the disciplines prominently in studies of transportation systems (Sen et.al. 2013*)*, infectious disease spread Rothenberg et al. 1995; PastorSatorras et.al. 2001; Bell et.al. 1999; Cook et al. 2007; computer viruses through the internet (Newman, 2002), animal behaviour (Wey, 2019); leadership evaluation (Hoppe et.al. 2010). The methodological and performance of networks tools evaluation have been reported by Freeman (1978); Brandes (2001); and Butts (2008). Further, social network analysis has extensively been reviewed by Wasserman and Faust (1994); Martínez-López et al. (2009). This study aims to model transmission network of an infectious disease (event under analyses - infection), particularly considering the casestudy of Tuberculosis (TB) network in Portugal. Three levels of risk factors were considered: index patients, contacts and characteristics of exposure. In methodological terms, three different models were used and the results were compared.

2. Methodology

The main interest in this study was to model the binary response (infected or not infected). Three levels of risk factors were considered: index-patients, contacts and characteristics of exposure: 1) index-patients: biologic products (Bronchoalveolar lavage, Sputum), chest radiography (Without cavitation, Cavitation), symptomatic period (in days); 2) contacts: age (years), sex (Male/Female), diabetes (Yes/No), Chronic renal failure (Yes/No); 3) exposure:

time spent together/relation (hours in 12 weeks), type of contacts (Co-habitant/ family/ friend/ work colleague/ known/ teacher/ patient/ others; grouped as casual/household), sleep together (Yes/No), eat together (Yes/No), size of exposure site ($<12\text{m}^2$ - "small", $\geq 12\ \text{m}^2$- "large"), location of contact (Car/ Bedroom/ Living room/Dining room/ Restaurant/ Open space/ coffee/ Other), ventilation facility of the exposure site (Yes/No). As each of the index patients are connected to several contacts, a hierarchical logit model can be employed to model the infection status considering risk factors. The hierarchical models are designed to handle mutual dependence among data points [23]. The last one is network logistic regression which could be more appropriate but also more complicated to implement in this study and as it is less known and used, some basic theoretical details are presented here and contextualized in our case-study.

Some network concepts: Metrics and Logistic Regression

In the present study, as it is one step network, i.e. the network of index-contact exists only, but no index-index or no contact-contact relations. For this reason, many of the social networks analyzing tools, though exist, do not carry meaningful interpretations and therefore, they are not explored for the current data. However, implementation of network logistic regression might be a realistic tool for studying any infectious disease including the current TB index-contact network. Because, network logistic regression considers the structural position of the TB patients interacting contacts as well as cofactors associated with index and contacts.

The logistic network regression framework is a simple basis for the modelling of joint edge/vertex dynamics with various orders of temporal dependence (Almquist and Butts, 2014). The models can be associated with dynamic network logistic regression or conventional cross-sectional network logistic regression. Given a random graph *G* on support Ψ, Almquist and Butts (2014) defined the general form of dynamic network logistic regression as follows:

$$P(G=g \mid w,\ \beta)=\frac{\exp\left(\beta^{T} w(g)\right)}{\sum_{g' \in \Psi} \exp\left(\beta^{T} w(g)\right)} \mathrm{I}_{\psi}(g) \qquad (1)$$

where *P(.)* is the probability mass function of its arguments, Ψ is the support of *G*, g is the realized graph, *w* is the function of sufficient statistics, β is the vector of parameters, and I_{Ψ} (g) is the indicator function. The indicator function I_{Ψ} (g) takes value 1 when the argument is in the support of Ψ and 0 otherwise. The general framework of the model described in (1) has been implemented with specific examples computationally for dynamic networks as well as cross-sectional version of the dynamic network logistic regression described in Almquist and Butts (2014).

Adjacency Matrices of Index and Contacts and Their Cofactor Matrices

We define the adjacency matrices of response variable originated from the interaction of index and contacts along with the risk factors connected to index and contacts in the network, of where 1 denotes infected and 0 denotes not infected and in the covariance related adjacency matrices 1 denotes presence of the characteristics and 0 denotes absence of that characteristics or in the case of continuous variable, just value of the covariate has been placed in the point of intersection of an index and contacts. The following algebraic notations have been presented to illustrate the ongoing discussion:

$$Y = \begin{array}{c} 1 \\ 2 \\ \cdots \\ \vdots \\ N \end{array}\begin{bmatrix} 0 & Y_{12} & \cdots & \cdots & Y_{1N} \\ Y_{21} & 0 & \cdots & \cdots & Y_{2N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_{N1} & Y_{N2} & \cdots & \cdots & 0 \end{bmatrix} \qquad X_i = \begin{array}{c} 1 \\ 2 \\ \cdots \\ \vdots \\ N \end{array}\begin{bmatrix} 0 & X_{i12} & \cdots & \cdots & X_{i1N} \\ X_{i21} & 0 & \cdots & \cdots & X_{i2N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{iN1} & X_{iN2} & \cdots & \cdots & 0 \end{bmatrix}$$

where Y is an adjacency response matrix, Xi is an *i*th covariate associated with index or contacts; i =

$$\text{logit } [Y] = \beta_0 + \beta_1 \mathrm{X}_1 + \ldots\ldots + \beta_{\mathrm{i}} \mathrm{X}_{\mathrm{i}} + \ldots\ldots\ldots + \beta_{\mathrm{p}} \mathrm{X}_{\mathrm{p}} + \varepsilon \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

1,2, … … … … … p . Now the network logit model can be defined as follows where *β*$_1$,*β*$_2$,.....,*β*$_i$,..........*β*$_p$ are model parameters and is an error term. In the current study, the order of adjacency matrix is 495 x 495 as there are 69 index patients and 426 contacts.

In the analysis, two statistical software were used: SPSS 20 to produce all the initial results (univariate and bivariate) as well as to model binary multiple and hierarchical logistic regressions; R, particularly *igraph* and *SNA* packages, for the network matrices and for the network logistic regression implementation. The *igraph* and *SNA* packages comprise a range of tools for social network analysis, including node and graph-level indices, structural distance and covariance methods, structural equivalence detection, network regression, random graph generation, and 2D/3D network visualization [17]. Statistical significance was set to 0.05 and we followed stepwise forward selection procedure for all three logistic models to identify significant variables.

3. Results

Comparison of Multiple, Hierarchical and Network Logistic Regressions

This section presents crude (simple logistic), adjusted (multiple logistic), hierarchical (mixed logistic), and network logistic regression odds ratio and their corresponding confidence intervals. Focused on infected status (dependent variable: infected or non-infected), the results of fitting the simple and multiple logistic regression models to the TB data are shown in Table 4. Table 5 presents the results of hierarchical and network logistic regressions.

In the Table 4, the crude estimates show that the covariates age, diabetes mellitus, exposure duration, sleep together, and eat together are significant at 1% or 5% level of significance. Whereas, the variables biological product, symptomatic period, age, contact type, and exposure duration are found to be significant in multiple logistic regression. The household contacts are 6.819 times more likely to be infected than casual contacts. The variables in multiple logistic regression have been selected by stepwise backward method based on likelihood ratio.

Table 4: Estimates of parameters in simple (crude) and multiple logistic regression (adjusted) analysis of TB networks in Portugal.

Characteristics	**Category**	**Binary Logistic Regression**					
		Simple Logistic			**Multiple Logistic (Model 1)**		
		Coefficient	**p-value**	**Crude OR (95% CI)**	**Coefficient**	**p-value**	**Adjusted OR (95% CI)**
Intercept					-3.070	0.000	
Index patient characteristics							
Biological product studied	Bronchoalveolar lavage			1			1
	Sputum	0.026	0.917	1.027 (0.623,1.692)NS	0.673	0.045	1.960 (1.016-3.782) *
Symptomatic period days		0.004	0.064	1.004 (1.000,1.009)NS	0.007	0.006	1.007 (1.002-1.013) **
Chest radiography	Without cavitation			1			
	Cavitation	-0.031	0.899	0.970 (0.602,1.561) NS			
Contacts' characteristics							
Sex	Female			1			
	Male	0.421	0.081	1.523 (0.949,2.444)NS			
Age in years		0.020	0.001	1.020 (1.009,1.032)**	0.024	0.003	1.024 (1.008-1.041) **
	No			1			

Diabetes mellitus	Yes	0.840	0.043	2.316 (1.028,5.217)*			
Chronic renal failure	No			1			
	Yes	-0.203	0.601	0.816 (0.381,1.748)			
Exposure characteristics							
Contact type	Casual			1			
	Household	1.585	0.000	4.877 (2.550,9.327)**	1.920	0.000	6.819 (2.906,16.000)**
Exposure duration (hours)		0.001	0.016	1.001 (1.000,1.002)*	-0.084	0.037	0.919 (0.8490.995)*
Sleeping together	No			1			
	Yes	1.681	0.001	5.373 (2.053,14.064)**			
Eat together	No			1			
	Yes	0.540	0.026	1.716 (1.067,2.760)*			
Size of the exposure site	Large			1			
	Small	0.105	0.728	1.110 (0.615,2.004)NS			
Ventilation of the exposure site	Yes			1			
	No	-0.157	0.709	0.855 (0.374.1.953)NS			

*Note: * means significant at 5% level, ** means significant at 1% level*

The results of fitting the hierarchical and network logistic regression models to the TB data are shown in Table 5. In Table 5, the results from hierarchical model show that the covariates age, diabetes mellitus, and contact type are significant at 1% or 5% level of significance. Whereas, the variables age, and contact type are found to be significant in network logistic regression analysis at 5% level. The network logistic regression considers the structural position of the covariates in a network. Interestingly, household contacts are 38.016 times more likely to have TB infection than the casual contacts. The results in network logistic regression provide more precise selection of variables that can be emphasized by the public health community in the real-life practice.

Table 5: Estimates of parameters in hierarchical and network logistic regression analysis of TB networks in Portugal

Characteristics	**Category**	**Hierarchical Logistic Regression (Model**			**Network Logistic Regression**		
		Coefficient	**p-value**	**Adjusted OR (95% CI)**	**Coefficient**	**p-value**	**Adjusted OR**
Intercept		-2.539	0.000		-8.801	0.000	
Contacts' characteristics							
Age in years		0.033	0.000	1.034 (1.015,1.053)*	0.0925	0.045 *	1.0969
Diabetes mellitus	No			1			1
	Yes	0.788	0.035	2.198 (1.059,4.566)*	1.128	0.483 NS	3.0895
Exposure characteristics							
Contact type	*Casual*			1			1
	Household	1.346	0.001	3.842 (1.766,8.359) **	3.638 *	0.042 *	38.016
Sleeping together	*No*						1
	Yes				1.438 NS	0.493	4.210

In multiple logistic regressions, Nagelkereke R^2 is 0.167, in hierarchical regression R^2 is not available and in network logistic regression (NLR) pseudo-R2 is 0.580 which indicates that the variation in having TB infection is better explained by network logistic regression as the NLR takes into account information on index-contact relation in addition to the exposure characteristics. However, there could be lack of information fitting all the models as the value of R^2 is not strong enough in all models. One of the reasons of poor performance of this network logistic regression approximation is that there could be some interaction terms those were not been considered in the model. This can be considered as a limitation of the network logistic regression process.

4. Discussion and Conclusion

This study has taken into account the network analysis of Tuberculosis patients particularly considering a TB network from Portugal. In this endeavor, we have compared the estimates from crude, multiple, hierarchical, and network logistic regressions to explore the factors influencing TB infection possibilities.

The crude estimate shows that age, diabetes mellitus, contact type, sleeping together, and eating together are important covariates. The multiple logistic regression reveals that index characteristics- biologic product and symptomatic period (days) and contact characteristics age, contact type, and exposure duration are important factors. However, simple and multiple logistic regressions do not consider the dependency of responses. The hierarchical logistic regression considers the dependency of responses while modelling TB network data. The hierarchical logistic regression finds that the variables age, diabetes mellitus, and contact type are significant for TB infection in contacts. The crude, multiple, and hierarchical models do not consider the structure of TB network. At this stage, the network logistic regression appeared to be a good tool for analyzing TB network data. The network logistic regression shows that only age and contact type are much valuable to interpret TB network data.

The current study was supplemented by two models viz. hierarchical and network logistic models. The benefit of these models was visible as they provide more precise list of predictors for TB infection. Hierarchical model indicates age, diabetes mellitus, and contact type being important and further network logistic analysis limits the predictors to age and contact type to be central for TB infection.

To sum up the discussion, we emphasize on the factors age and contact type (household or casual) to be the most important factors for TB infection when people interact TB patients in the real life. In older age, people are vulnerable to any disease including TB as their immune system become weaker, thus older people are more likely to have TB infection than the younger people. Contact type- particularly household contacts appeared to be more exposed to TB infection than others as found in all methods. The other risk factors e.g. eating together, sleeping together are broadly represented by household contacts. None of the exposure characteristics i.e. exposure site (big or small) and ventilation facilities (yes or no) are found to be important for TB infection among contacts.

The current work can be extended to dynamic network system where for each of the time point in a follow up study, the number of contacts (edges) and number of vertices (nodes), number of infected individuals can also be predicted. If dynamic network logistic regression is in use, it would be possible to forecast mean degree (number of nodes on average) and network size for a particular future time point. The similar approach, particularly network logistic regression, can be replicated in the investigation of other infectious diseases.

Acknowledgement

The author MLR greatly acknowledges to NOVA University of Lisbon for awarding a postdoctoral fellowship during this research.

References

1. Almquist ZW and Butts CT (2014). Logistic Network Regression for Scalable Analysis of Networks with Joint Edge/Vertex Dynamics, *Sociological Methodology*, 44: 273-321. Bell DC, Atkinson JS, & Carlson JW. (1999) Centrality measures for disease transmission networks. *Social Networks*, 21: 1-21.
2. Brandes U., (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2), 163-177.
3. Butts CT., (2008). Social network analysis with sna. *Journal of Statistical Software*; 24: 1-51. Cook VJ, et al. (2007) Transmission network analysis in tuberculosis contact investigations. The *Journal of infectious diseases*; 196(10), 1517-1527.
4. Freeman LC (1978). Centrality in social networks: conceptual clarification. *Social Networks*; 1: 215–239. Hoppe, B., & Reinelt, C. (2010). Social network analysis and the evaluation of leadership networks. *The Leadership Quarterly*, 21: 600-619.
5. Lusseau D, Whitehead H and Gero S., (2008). Incorporating uncertainty into the study of animal social networks, Animal Behaviour 75: 1809-1815
6. Makagon MM, McCowan B, Mench JA(2012). How can social network analysis contribute to social behavior research in applied ethology? Applied *Animal Behaviour* Science 2012; 138:152161
7. Martínez-López B, Perez AM, and Sánchez-Vizcaíno JM., (2009). Social network analysis. Review of general concepts and use in preventive veterinary medicine *Transboundary and Emerging Diseases*; 56: 109-120.
8. Mendes MA, et al. (2013) Contact screening in tuberculosis: can we identify those with higher risk? *European Respiratory Journal*; 41: 758-760; DOI: 10.1183/09031936.00164612 Newman ME, Forrest S, & Balthrop J. (2002) Email networks and the spread of computer viruses. Physical Review E; 66(3), 035101.
9. Pastor-Satorras R, and Vespignani A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev*. Lett 2001; 86: 3200–3203.
10. Rothenberg RB et al. (1995). Choosing a centrality measure: epidemiologic correlates in the Colorado Springs study of social networks. *Social Networks* 17: 273-297. Sen P, et al. (2013) Small-world properties of the Indian railway network. Physical Review E; 67: 036106.
11. Wasserman S and Faust K., (1994). *Social Network Analysis*: Methods and Applications; Vol. 8. Cambridge University Press, Cambridge.

12. Weber N, *et al*. Badger social networks correlate with tuberculosis infection. Current Biology 2013 Oct 21; 23:R915-6.
13. Wey T, Blumstein DT, Shen W, & Jordán F. (2008) Social network analysis of animal behaviour: a promising tool for the study of sociality. *Animal Behavior*, 75(2), 333-344

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

Anchored SEM – Merging structural equation models

Lukasz Widla-Domaradzki
Polish Agency for Enterprise Development

Abstract

When performing Classical Structural Equation Model a researcher quickly discovers one main limitation: preparing SEM is very data-consuming process[1]. In most cases that means either find another statistical tool which allows produce desired estimates or try to simplify proposed model, for example cutting number of variables. Anchored SEM is the analytical tool that allows to analytically create a connection between independent SEM models drawn from the same sample. This allows to build smaller SEMs independently and merge them on the latter stage. Anchored SEM may be a solution when complex model is not computable. At this approach one SEM is embedded in another not fully, but by the anchor: as part of other SEM model. An anchor in this case is this part that allows to estimate to what extent both SEMs are connected and – therefore – what estimates of the hypothetical larger SEM (built from SEM1 and SEM2) should look like.

Keywords
Structural Equation Model; SEM

1. Introduction

Structural Equation Models (SEM) are still one of the most recognized tools for preparing complex and composite indicators[2]. They are often used when the relation between latent variables is known and a researcher can assign specific observed variables to the latent constructs. However, SEMs have several limitations: 1) they require large sample to obtain satisfactory results[3]; 2) the variables cannot be multicollinear; 3) Structure of the model should be known and well described in the theory. As a result very complex indicators may not be computable because of that constraints. In this paper I propose how to overcome the first problem. It usually occurs when sample is too small but may also be useful to simplify very complex Structural Equation Models.

[1] *"The sample size, as a rule of thumb, is recommended to be more than 25 times the number of parameters to be estimated, the minimum being a subject-parameter-ratio of 10:1"*, Nachtigall C., Kroehne U., Funke F., Steyer S. "(Why) Should We Use SEM? Pros and Cons of Structural Equation Modeling" in Methods of Psychological Research Online, 2003

[2] E.g. Composite Learning Index: http://www.niagaraknowledgeexchange.com/wp-content/uploads/sites/2/2014/10/2010CLI-Booklet_EN.pdf

[3] Barbara M. Byrne, *Structural Equation Modeling with AMOS*, Routledge, 2009

The solution proposed here is based on anchoring dependent SEM models into the basic one to obtain better analytical result. Through this method we are able to build smaller models independently and merge them analytically at the later stage.

2. Methodology

Anchored SEM is the analytical tool that allows to analytically create a connection between at least two independent SEM models drawn from the same sample. Those models should be – at the later stage – part of one complex solution. It is possible to include more than two models in this procedure, however, in this paper I focus on the simple solution when only two models are merged. Another – mentioned above – constraint is models should be dependent. In other words, one of the models should be able to explain a part variability of the second one. If this condition is not satisfied (for example with models which are related, but there is no dependency between them) Anchored SEM can't be used.

In the analysis presented below I worked with two independent SEM models drawn from the same sample. Sample for this study was obtained from the population of over 100 thousand of Polish enterprises. One of the aims of the study was to analytically contribute to the large and complex indicator of Polish Enterprises Innovation Index. Theory for the proposed index[4] is well described, so the SEM might be used.

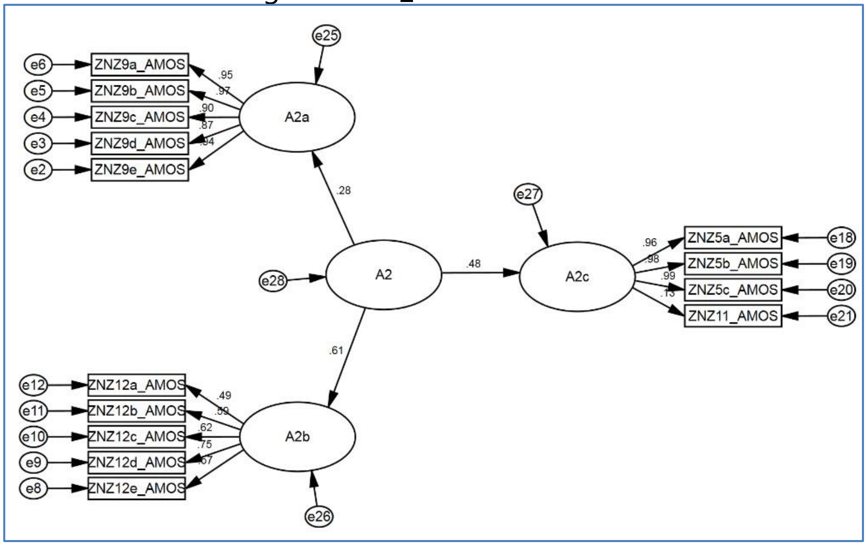
The database used for presented analysis was coming from a pilot study and contained 1327 cases. Difficult and multi-layered model that was needed to construct a nationwide index demands bigger sample, but I attempted to build a preliminary model on limited data and revalidate it after several waves of the study become available. In order to do this, I had to develop some kind of shrinking or folding models to connect some of the SEM latent variables to each other. As a first step of this exercise, I prepared two connected (at the theory level) models A2 and A3 (see Fig 1). Model A2 contained variables connected with innovative infrastructure (such as: "level of process automation"; "software for supporting ERP (enterprise resource planning)" or "the number of employees who are responsible for developing innovations" etc.), while model A3 included variables connected with innovative management (such as level of agreement with the statements: "innovations are supported by company managers", "there is an innovation management system", "there are processes that allow us to effectively manage the development of a new product" etc.). In theory innovative infrastructure is a

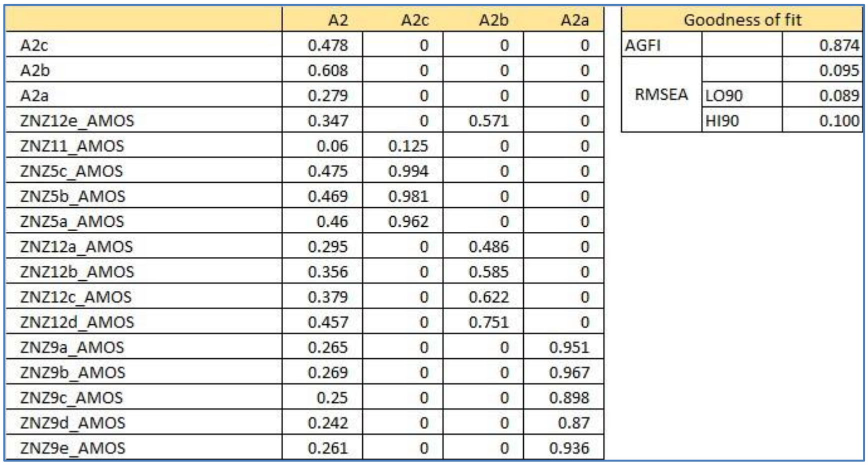
[4] Theory of innovation is mainly based on expectancy-value theory first proposed by John William Atkinson and expanded for the field of innovation studies by Martin Fishbein (in Stephen W. Littlejohn's "Theories of Human Communication")

part of management, that’s why model A2 (infrastructure) is anchored with a part of model A3.

All of the input variables were standardized before being put into the model. Models visualisation and main standardized estimates are presented below on graphs. Tables include all total standardized estimates between latent and observable variables.

Fig.1: Model_1. A2 – main model

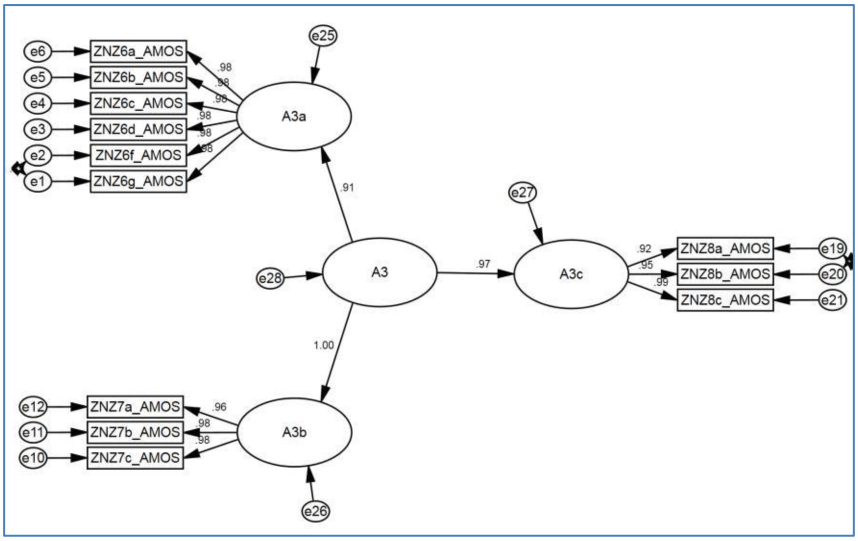


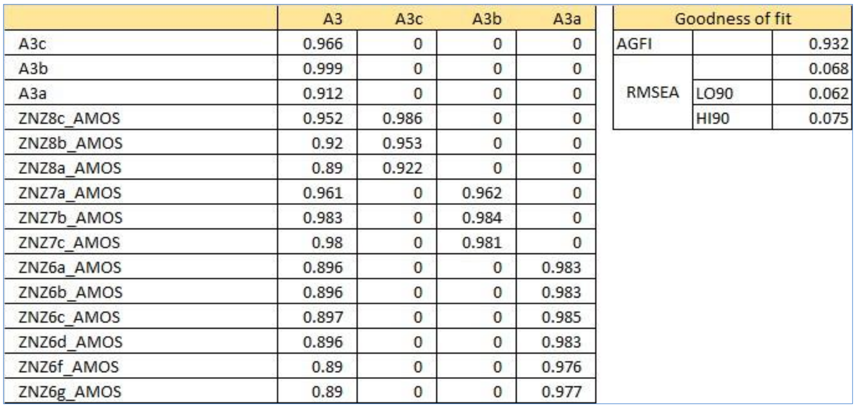


	A2	A2c	A2b	A2a
A2c	0.478	0	0	0
A2b	0.608	0	0	0
A2a	0.279	0	0	0
ZNZ12e_AMOS	0.347	0	0.571	0
ZNZ11_AMOS	0.06	0.125	0	0
ZNZ5c_AMOS	0.475	0.994	0	0
ZNZ5b_AMOS	0.469	0.981	0	0
ZNZ5a_AMOS	0.46	0.962	0	0
ZNZ12a_AMOS	0.295	0	0.486	0
ZNZ12b_AMOS	0.356	0	0.585	0
ZNZ12c_AMOS	0.379	0	0.622	0
ZNZ12d_AMOS	0.457	0	0.751	0
ZNZ9a_AMOS	0.265	0	0	0.951
ZNZ9b_AMOS	0.269	0	0	0.967
ZNZ9c_AMOS	0.25	0	0	0.898
ZNZ9d_AMOS	0.242	0	0	0.87
ZNZ9e_AMOS	0.261	0	0	0.936

Goodness of fit		
AGFI		0.874
RMSEA		0.095
	LO90	0.089
	HI90	0.100

Fig.2: Model_2. A3 – main model



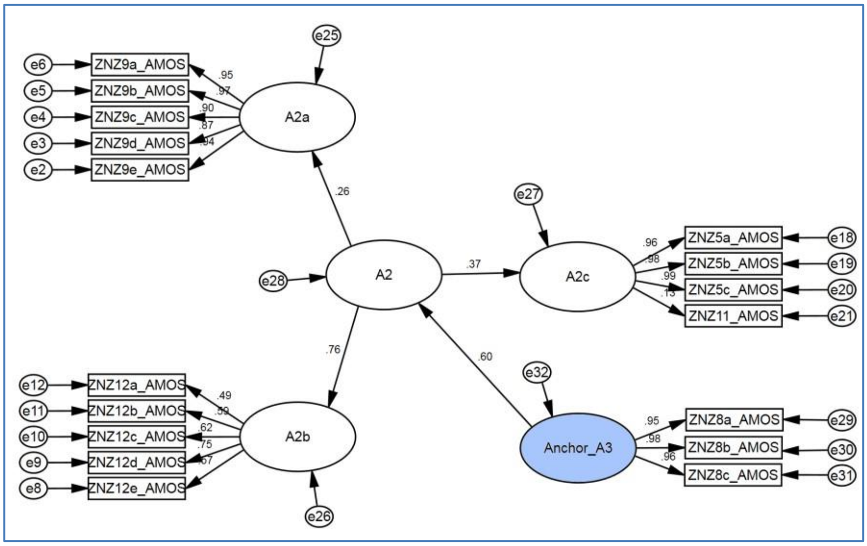


	A3	A3c	A3b	A3a
A3c	0.966	0	0	0
A3b	0.999	0	0	0
A3a	0.912	0	0	0
ZNZ8c_AMOS	0.952	0.986	0	0
ZNZ8b_AMOS	0.92	0.953	0	0
ZNZ8a_AMOS	0.89	0.922	0	0
ZNZ7a_AMOS	0.961	0	0.962	0
ZNZ7b_AMOS	0.983	0	0.984	0
ZNZ7c_AMOS	0.98	0	0.981	0
ZNZ6a_AMOS	0.896	0	0	0.983
ZNZ6b_AMOS	0.896	0	0	0.983
ZNZ6c_AMOS	0.897	0	0	0.985
ZNZ6d_AMOS	0.896	0	0	0.983
ZNZ6f_AMOS	0.89	0	0	0.976
ZNZ6g_AMOS	0.89	0	0	0.977

Goodness of fit		
AGFI		0.932
RMSEA		0.068
	LO90	0.062
	HI90	0.075

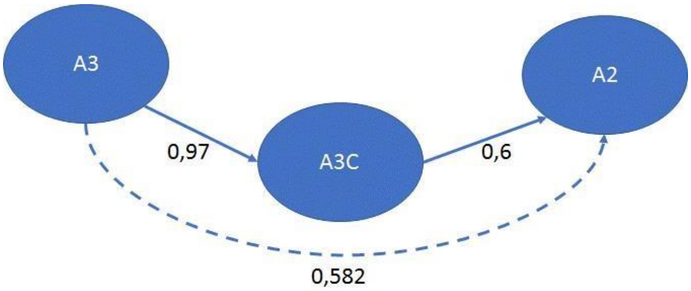
To merge those two models I anchored part of A3 model (Anchor_A3) in the A2 model. Anchor is therefore a part of one model (in this case model_A3) that becomes embedded in the other. Anchor allows us to establish the connection between two models without actually connecting them. There is a little flexibility from the researcher side to decide which model can be anchored in which. As is written at the introduction of this paper this should be embedded in the underlying theory. When the connection is established, anchor can be made – always at the side of the dependent model. In this case, anchor represents a part the higher model (A3) embedded in the model A2. Graphical representation is presented below:

Fig.3: Anchored A2 model



Knowing the estimates from the A3 model (presented at figure 2), we may now estimate connection between A2 and A3 models. Note, that I used reflective notation for main (A2 & A3) model construct and anchor is added as a formative component. In this case I assume that my final model will have model A2 as a dependent one for model A3:

Fig.4: Dependency level between model A2 (anchored) and model A3



To summarize steps undertaken so far:

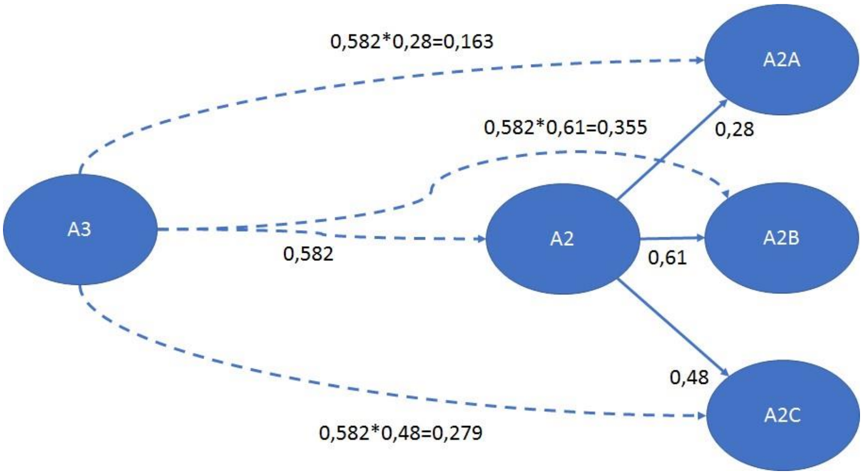
1) I assumed my desired indicator should be built by two SEM models, one of them dependent from another
2) I built two independent SEM models: A2 and A3 to get the total standardized estimates of the parameters
3) I amended model A2 (dependent model) by adding an anchor, which was the part of model A3. I got total standardized estimates of the parameters for anchored model

4) Because estimates for the model A3 are known, I can now merge analytically both models. In this case my anchor will show strength of the connection between model A2 and A3.

3. Results

Finally, I can merge both models. Below I present the simplified version with only latent variables, to illustrate how the connection is made and what can we assume on its strength. Of course, there is a possibility of computing the rest of the estimates for observed variables from A2A, A2B and A2C parts of the model as well as to compute connection between components of the A3 model and components of A2 model. But what can be achieved by using an anchor is visible even in the simplified version is presented below:

Fig.5: Simplified dependency between latent variables from model A2 and model A3



Of course, using anchor changes specification of A2 model as well as goodness of fit. That's why it's important to have the estimates of the non-anchored model before adding an anchor. In the example shown above I used original A2_model estimates (0,28; 0,61; 0,48). It's important, because anchor in fact is interfering with the model! In this case I knew my model A2 will be built from three latent and 14 observed variables: this is how "innovative infrastructure" was described by the theory. Anchor, however useful in case of merging models is not the part of the real A2 model – it's just a tool added to it. Here is matrix of total standardized effect for both (anchored and original) models:

Fig. 6: Differences in estimates between anchored and original model A2

	A2 (anchored)	A2 (original)
A2c	0.374	0.478
A2b	0.763	0.608
A2a	0.255	0.279
ZNZ12e_AMOS	0.436	0.347
ZNZ11_AMOS	0.047	0.06
ZNZ5c_AMOS	0.372	0.475
ZNZ5b_AMOS	0.367	0.469
ZNZ5a_AMOS	0.36	0.46
ZNZ12a_AMOS	0.376	0.295
ZNZ12b_AMOS	0.449	0.356
ZNZ12c_AMOS	0.473	0.379
ZNZ12d_AMOS	0.57	0.457
ZNZ9a_AMOS	0.243	0.265
ZNZ9b_AMOS	0.247	0.269
ZNZ9c_AMOS	0.229	0.25
ZNZ9d_AMOS	0.222	0.242
ZNZ9e_AMOS	0.239	0.261

Summarizing this part, anchoring one model in another researcher is able to estimate connection between them without simplifying the models itself or getting bigger sample (which is always a good – but in most cases, impossible – solution).

4. Discussion and Conclusion

Merging SEM models using an anchor may be a solution for datasets with insufficient number of cases. Another solution is to fold one of the models: in this scenario one model (for e.g. Model A2) is estimated, imputed to the dataset as an observed variable and used as another variable in larger model. In this scenario a lot of variability of the first model is lost, since there is a single variable representing the whole partial model. Anchoring allows to uphold whole variability from one model and connect it with another. One known problem which is needing further development – is that there is no information about goodness of fit of the merged model, as the merging is made not as a SEM procedure

References

1. Handbook on Constructing Composite Indicators, OECD, 2008
2. Barbara M. Byrne, Structural Equation Modelling with AMOS, Routledge, 2009
3. Rick H. Hoyle, Handbook of Structural Equation Modelling, Guilford Press, 2014
4. John W. Atkinson, George H. Litwin, Achievement Motive and Text Anxiety Conceived as Motive to Approach Success and Motive to Avoid Failure, Bobbs-Merrill Company, 1960.
5. Stephen W. Littlejohn, Theories of Human Communication, Wadsworth Publishing Company, 1996

The national data infrastructure in statistics Portugal and the data access for scientific research purposes – Evolution and challenges

José A. Pinto Martins, Maria João Zilhão
Statistics Portugal, Lisbon, Portugal

Abstract

The access to anonymized official microdata by researchers in Portugal is guaranteed by Statistics Portugal. This paper describes how researchers access the data free of charge, highlighting the most important issues of the process, namely the legal framework and the cooperation protocol with the Ministry of Education and Science, describing the process accreditation of researchers. Some statistics about the use and demand by accredited researchers to databases are presented. The paper also presents the Statistics Portugal already in movement project the National Data Infrastructure (NDI), that will allow data access maximization, namely for scientific and research purposes, and for the production of Official statistics for better decision making.

Keywords

Microdata databases; Researchers' accreditation; anonymized individual statistical data; free access

1. Access to anonymized official microdata

Statistics Portugal (SP) being aware of the fact that the academic community has special needs regarding statistical information, namely for the development of research projects and for the preparation of Master's and PhD theses, has established a Protocol with the Ministry of Education and Science, specifically the Foundation of Science and Technology (FCT - entity responsible for funding R&D in Portugal) and the General Directorate of Statistics for Education and Science (DGEEC), in order to facilitate the access of (accredited) researchers to the official statistical information needed to carry out their activities. The protocol concerns researchers from universities and other legally recognized higher education and research institutions.

Researchers with proven affiliation in international organisations such as: specialised agencies of the United Nations (International Labour Organisation (ILO), Food and Agriculture Organisation (FAO), United Nation Educational, Scientific and Cultural Organisation (UNESCO), The World Bank Group and the International Monetary Fund (IMF), and OECD, are also eligible for accreditation concerning their expertise and reputation in quality scientific research.

DGEEC is responsible for accrediting users and providing them with the necessary information and each researcher must sign a form and a Statement of Confidentiality Commitment.

The accreditation granted by DGEEC is valid during the declared length of the research project and only for the data identified in the request. It requires signature of a Code of Conduct by the applicant and the research institution of affiliation.

Under the protocol three access modes are possible to be authorized, including provision of fully anonymised data files and ready-made tables that allow no form of re-identification of statistical units; and exceptionally, on-site access in a safe environment, allowing the use of indirectly identifiable microdata under strictly controlled conditions (subject to a previous additional assessment by SP and an external group of experts in the statistical domain of the request).

PhD students have access under the same conditions as other researchers. Master's students need to fulfil an additional condition: the request and the statement of commitment must be also signed by the supervisor.

Non-resident researchers can access statistical data under the same conditions as the Portuguese ones in case they participate in a Foundation for Science and Technology Portuguese training scholarship or in cooperation programmes in R&D with Portugal.

2. How to Apply & Modes of Access

The researcher must submit a request for accreditation for a specific research project, naming all the researchers involved and specifying the duration, goals and methods of the project. The request is examined by DGEEC for accreditation purposes and then forwarded to SP if approved. When access is granted, then a contract must be signed between the researchers and SP, specifying obligations of the parties and sanctions in case of breach. This contract specifies in particular the duty of confidentiality, non-replication of data, data security conditions and the obligation to destroy data after its use, as well as the implicit legal responsibility issues.

Exceptionally, access to a secure environment (safe centre) can be provided and the researcher can access the microdata in raw format but without direct identification. Outputs are always subject to confidentiality checks before release, to assure that the final information does not contain any direct or indirect identification.

3. Legal Framework

According to the Portuguese Statistical Law (Law nº 22/2008, 13 May), which defines the principles, rules and structure of the National Statistical System, concerning statistical confidentiality it states that "*All individual statistical data* *collected by statistical authorities are confidential*". However, it is also stated that access to anonymised data "…*shall only be supplied for SCIENTIFIC PURPOSES, if ANONYMISED, upon an AGREEMENT signed between the statistical authority supplying the data and the entity requesting them…*".

The applying process is based on three phases, as represented in Figure 1:

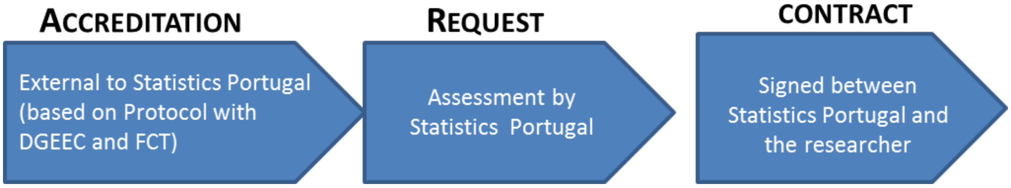

ACCREDITATION
External to Statistics Portugal (based on Protocol with DGEEC and FCT)
REQUEST
Assessment by Statistics Portugal
CONTRACT
Signed between Statistics Portugal and the researcher

Figure 1

And the accreditation flow is described in Figure 2

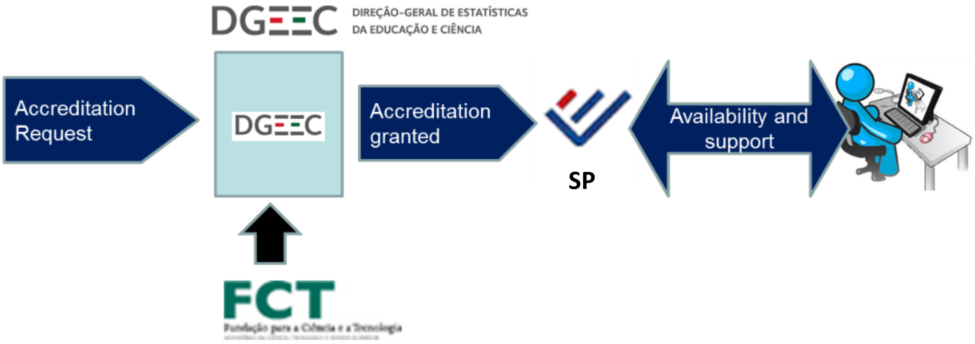

DGEEC DIREÇÃO-GERAL DE ESTATÍSTICAS DA EDUCAÇÃO E CIÊNCIA
Accreditation Request
DGEEC
Accreditation granted
SP
Availability and support
FCT
Fundação para a Ciência e a Tecnologia

Figure 2

Since the beginning of this service rendered to the academia SP deals with an average of 40 applications and 81 researchers per year. However, the trend is growing, which may be justified both by the increase of databases made available per various statistical domains and by the scientific community's knowledge of this service and availability of data access.

SP has presently a set of 47 microdata databases available, ready to be requested by researchers for scientific research. All databases have a brief description of their contents, the related methodological documents and files/databases registries description, in order to help researchers to make a better choice.

Those databases are organised in 15 domains:

Domain/Theme	N.º of databases	Domain/Theme	N.º of databases
Living Conditions	7	Agriculture	2
Enterprises	5	Industry and Energy	2
International Trade	1	Information Society	4
Construction and housing	2	Health	4
Demography	7	Education, training and learning	1
Labour market	3	Culture, sports and recreations	2
Short term indicators	3	Science and Technology	3
Transports and communication	1	**Total**	**47**

Table 1

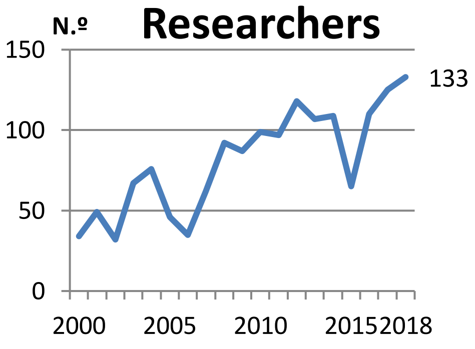


Figure 3

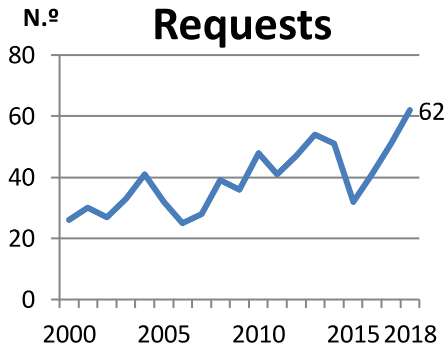


Figure 1

In 2018 we have received 62 applications (Figure 4) which correspond to 133 researchers (Figure 3). It is worth mention that between 2000 and 2018 SP has received a total of 744 requests for data, from which 421 were from researchers' teams and 323 were from individual researchers.

This solution, based on the above referred protocol, allows SP to provide the data to researchers in a fairly quick delivery period of time. Through a secure cloud, based at SP, access is granted in a controlled and totally safe environment to each researcher.

The average time for the availability of microdata databases to researchers was significantly reduced until 2013 and then stabilized around the average of one working day per answer. In 2018, this average time reached 0.8 working days, according to figure 5.

It is important to mention that with the beginning of the delivery of microdata databases to researchers, begins an intense and continuous relationship of cooperation and interaction.

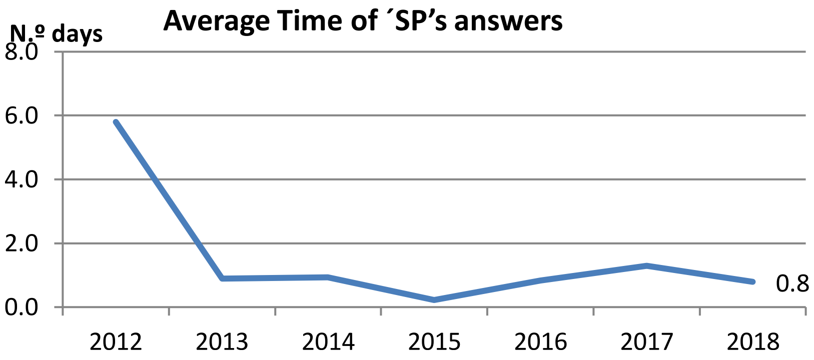


Figure 5

The starting of microdata analysis by the researcher often identifies new information needs, either because additional databases are needed, or because questions occur about the data accessing, metadata, etc. This situation is easily verified by the number of additional requests accessed to the microdata databases.

(in figure 6: 123 additional requests in 2018, are derived from 62 initial requests – figure 4).

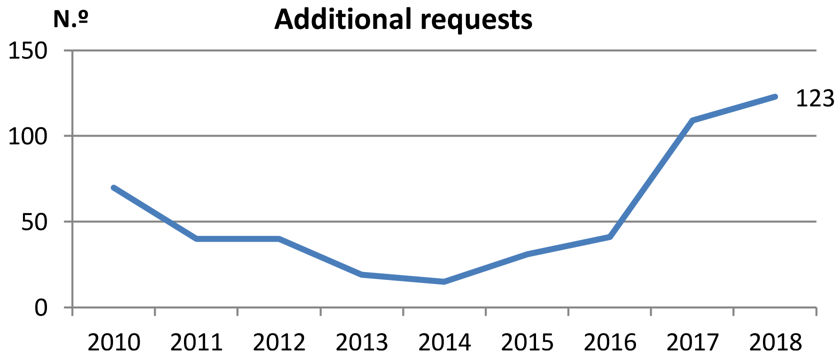


Figure 6

These additional interactions are a result of the increased demand and use of available information. The increase in research projects are the most relevant in number, when compared together with the requests for doctoral and master's degrees, as presented in figure 7.

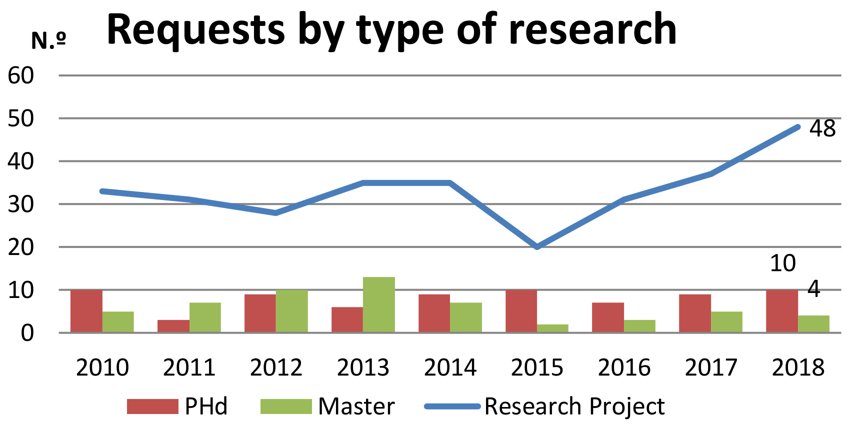


Figure 7

Since 2014 (Figure 6), SP has been receiving a growing number of accreditations requests to access microdata and additional requests of information. The advertising SP has been doing in universities and research centres are the main reason for that, particularly about the possibility of easy, free and fast access (three keywords) to statistical microdata for scientific research purposes, which contributes to increase the satisfaction levels of the national scientific community.

Meanwhile, SP has prepared files with information at observation unit level, the so-called Public Use Files (PUFs).

These files (data and metadata) contain anonymised records, processed and prepared in a way that the observation unit cannot be identified directly or indirectly, with the exception of statistical data on Public Administration.

Accesses to these files are free and comply with the principle of statistical confidentiality and personal data protection.

This offer helps to increase the interest and use of statistical data, mainly for university, or even high school, students. The present offers of PUFs are:

- Census 2001 and 2011 (5% sample on individuals and dwellings).
  The file corresponds to a 5% sample of residents. It includes two sampling tables, one for family and collective dwellings, containing some variables on the building; another table for resident individuals, both with 5% registers and a common bonding variable.
- Public Museums, from 2013 to 2016.
  The file contains annual data characterizing Museums, specifically, human and financial resources, store, collections and inventory, visitors' oriented activities, number of visitors (monthly and annual flows), and type of visitors (school groups and foreign visitors).

- Public Hospitals and Health Centres, from 2012 to 2016.

The file contains data on physical variables of public hospitals and Health Centers, in particular characteristics, equipment, facilities, human resources and demand and performance in hospitals.

4. The National Data Infrastructure

SP is the national statistical authority that is part of the European Statistical System and has specific and unique attributions that guarantee the independence and security of information. The evolution towards a National Data Infrastructure (NDI) is a logical and natural step for SP. This trend has not started now, but it is and will be gradual.

The increasing digitization generates a large volume of data. Their transformation into knowledge is only possible with adequate forms of storage, treatment and analysis, in a safe, consistent and reliable way.

SP already uses considerable amounts of administrative data in the production and statistical analysis process. As a national statistical authority, it has a structure that ensures the protection and integrity of the data, much like a safe-box. The new information requirements lead to the need to provide the State with the capacity to manage and analyze a large data set. Integrating fully the satisfaction of this need into SP, evolving into a national data infrastructure, has obvious gains in scale.

Taking advantage of SP competencies, tasks and mission, the objective is to adopt a more intensive and integrated use of data in the production of statistical information and to take advantage of the entire production chain, from the development of platforms, applications and algorithms to the collection and validation of data, until the analysis of the statistical information. It is an evolution of the SP and the development of new skills that will allow gaining resources, space to intensify innovation throughout the organic structure, through a greater return to society.

NDI will seek to respond to the need for SP scale up and gain critical mass in order to respond to an increasingly complex society that generates new expectations regarding statistics. New services and statistical products are sought, with new approaches, with a guarantee of quality.

The development of the NDI will ensure the SP the critical dimension to continue developing its skills and improving statistical production, benefiting the country by the increased processing and analysis capacity.

The creation of a national data infrastructure will take full advantage of the (growing) set of available data, without endangering their safety and privacy. Intensification of ownership and use of administrative data in the SP production process anticipates a large increase in the volume of data and a substantial broadening of the covered areas.

The national data infrastructure's main purposes and associated benefits, among others, provide a set of related data and resources from a single point of entry, regardless of where the data is kept or how data can be accessed (opened, protected or safe) to be guarantees of safety and quality data by providing integrated data services and metadata.

While on the one hand the whole process of producing official statistics greatly benefits from having access to a vast repository of administrative data gathered in a single access point, also the availability of information available at the level of anonymized microdata with multiple crossings and scientific research gains a new dimension and opens up numerous opportunities for partnerships and knowledge sharing, in the strictest compliance with the law of the statistical system, and in particular with respect to confidentiality.

In summary, we can say that the NDI will increase the economic and social impact of the public good that is the statistical information

5. Discussion

Access to confidential data for scientific and research purposes is a priority for SP and for the use of Official statistics because many questions – economic, social, and environmental and political sciences – can be answered adequately on the basis of relevant and detailed data allowing in-depth analyses. Proper delivery services to provide better access to data are also a main concern.

A wider access to confidential data for scientific work and research without compromising the high level of protection is beneficial to statistics and the experience and needs of researchers are relevant for improving accessibility of confidential data.

SP commitment is to provide good information and researchers' responsibility is to use data in full respect for the fundamental Principle of Statistical Confidentiality.

SP GOOD experience shows that COOPERATION for this specific purpose is an advantage: it allows, namely, split and share responsibilities' among parties with specific knowledge, and definitely rendering the system more efficient.

A comparative study of test statistics for testing homogeneity of variances in analysis of variance models

Oladugba Abimibola Victoria[1], Okiyi Bright Chiamaka[1], Caroline Ogbonne Odo[2]

[1] Department of Statistics, University of Nigeria, Nsukka

[2] Department of Agricultural Economics, Michael Okpara University of Agriculture Umudike, Abia State

Abstract

This study compared seven methods of testing homogeneity of variances in one-way and two-way analysis of variance models under the assumptions of normality and non-normality distributions when the sample sizes are equal and unequal using type-one-error and power of the test. The methods compared were: Bartlett test, Levene test, Brown-Forsythe test, O'Brien test, Z-variance test, Hartley's F-max test and Cochran's G-test. Monte Carlo simulation was used to generate response observations for normality and non-normality distributions (Chi-square). The result from the analysis showed that under normality and non-normality distributions, the Brown-Forsythe and O'Brien tests committed the least type-one-error while the Levene and Bartlett test maintained the highest power respectively with equal and unequal sample sizes in one-way analysis of variance. The Bartlett, Levene and Z-variance maintained the highest powers while the O'Brien committed the least type-one-error under non-normality with equal and unequal sample sizes in two-way analysis of variance.

Keywords

Type-one-error; Power; Bartlett; Levene test; Normality and Non-normality

1. Introduction

In many experimental data, the first thing one noticed in the data set is that the observed values are not all the same even under the same condition or subject. This shows that there is variability in the data set. The statistics that deals with variability in a data set is called variance. Variance is the expectation of the squared deviation of a random variable from its mean that is variance measures how far each observation in the data set was from their mean Vanhove (2018). When all the observed values in a data set are identical, the variance will be zero but when they are not all identical, the variance will be greater than zero; a large variance indicates that most of the observed values in the data set are far from each other while a small variance indicates the opposite Peter (2013).

The assumption of homogeneity of variance (HOV) also known as homoscedasticity implies that the variance within each of the population groups are equal. This is one of the assumptions of analysis of variance (ANOVA). When this assumption is violated, there is a greater probability of falsely rejecting the null hypothesis. Lack of homoscedasticity is known as heteroscedasticity. The statistical validity of many commonly used tests such as the t-test and ANOVA depend on the extent to which the data conform to the assumption of HOV. Accessing HOV is of paramount importance to many researchers as they are concerned with whether the dispersion of the dependent variable is similar across multiple groups. When comparing groups, their dispersion on the dependent variable should be relatively equal at each level of the independent (factor or grouping) variable (and neither should their sample sizes vary greatly across the groups), this implies that the dependent variable should exhibit equal levels of variance across the range of groups. Furthermore, violation of the assumption of HOV distort the F-distribution in ANOVA to such an extent that the critical F-value no longer corresponds to the chosen level of significance, this leads to a serious type-one-error Peter (2013). Heteroscedasticity may not only affect the validity test, it may also have a negative effect on the coverage of the confidence intervals and the accuracy of the estimator Koning (2014).

According to Parra-Frutos (2012) problems of heteroscedasticity of data set arise when the sample sizes are unequal. Usually unequal sized groups are common in research and may be as a result of simple randomization. In test like ANOVA, having both unequal sample sizes and variance dramatically affects statistical power and type-one-error rates Vanhove (2018). Test statistics is relatively insensitive to small departures from the assumption of equal variances for the treatments if the sample sizes are equal, this is not the case for unequal sample sizes; Also, power of the test is maximized if the samples are of equal sizes Montgomery (2013).

One of the basic assumptions to formulate classical tests for comparing variances is normal distribution of samples Garbunova & Lemeshko (2012). It is well known that classical tests are very sensitive to departures from normality. Therefore, the application of classical criteria always involves the question of how valid the obtained results are when the data are non-normal. Testing for the equality of variance is a difficult problem due to the fact that many tests are not robust to non-normality and are affected by sample sizes Vanhove (2018). Therefore, given that there are several methods available in testing for HOV, it would be of interest to determine which of them perform best under normality and non-normality when the sample sizes are equal and unequal. In this work, seven HOV tests using one-way and two-way ANOVA models were compared when the data are normal and non-normal for equal and unequal sample sizes. They are Bartlett test, Levene’s Test, Z-Variance test, Hartley F-max test, Cochran's G-test, Brown-Forsythe test and O'Brien Test. These tests are compared in order to find out which is more robust in terms of type-one-error rate and statistical power.

2. Methodology

The data used in this work were simulated from normal distribution and non-normal distribution (Chi-square distribution) based on an example from Montgomery (2013, p 134 and 177). The simulated data used represents the observed responses of an experiment adopted from Montgomery (2013, p 134) for the one-way ANOVA and Montgomery (2013, p 177) for the two-way ANOVA. For the one-way ANOVA, the data were generated from the experiments that studies the live of three different brands of battery (Brand 1, Brand 2 and Brand 3), each replicated five times. The lives of the batteries were measured in weeks. The experiment was designed to investigate if there was a difference between the lives of the three selected brands of battery. For the two-way ANOVA, the data were generated from an experiment that test for the effect of four chemical agents (factor A) on the strength of five bolts of cloth (factor B). The effects were measured in N/M^2 (Newton per square metre).

Three factors were manipulated in the course of the simulation: population distribution, sample size and population variance. the manipulation gave rise to eight configurations which are: normal distribution with equal sample size and equal variance, normal distribution with equal sample size and unequal variance, normal distribution with unequal sample size and equal variance, normal distribution with unequal sample size and unequal variance, non-normal distribution with equal sample size and equal variance, non-normal distribution with equal sample size and unequal variance, non-normal distribution with unequal sample size and equal variance and non-normal distribution with unequal sample size and unequal variance. Type-one-error was investigated for any configuration with equal variance, this is done by counting the number of rejection of H_0 (equal variance) when it is true. The power of the tests were investigated for any configuration with unequal variance, this is done by counting the number of rejection of H_0 (equal variance) when it is false. Unequal variance within the group samples was implemented by multiplying Brand 2 and Brand 3 group values by 2 and 3 respectively in the one-way ANOVA data and Chemical (2, 3 and 4) rows by 2, 3 and 4 respectively in the two-way ANOVA. A total of 2400 data points that is 20(iterations) * 8(configurations) * 3(groups) * 5(replications) were simulated for the one-way ANOVA and 1,600 data points, that is 10(iterations) * 8(configurations) * 4(groups) * 5(replications) were simulated for the two-way ANOVA.

Description of the tests

The Bartlett Test (B): The test statistics is given as follows

$$T = \frac{(N-k)\ln S_p^2 - \sum_{i=1}^{k}(N_i-1)\ln S_i^2}{1+\left(\frac{1}{3}(k-1)\right)\left(\left(\sum_{i=1}^{k}\frac{1}{N-1}\right)-\frac{1}{N-k}\right)} \qquad (1)$$

where S_i^2 is the variance of the i^{th} group, *N* is the total sample size, *N_i* is the sample size of the i^{th} group, k is the number of groups, and S_p^2 is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as:

$$S_p^2 = \frac{\sum_{i=1}^{k}(N_i-1)S_i^2}{N-k} \qquad (2)$$

The Levene Test (L): The test statistics, **W** of the absolute residuals are compared with the F critical value with k-1 and N-k degrees of freedom in the numerator and denominator respectively.

$$W = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k}N_i(\bar{Z}_{i.}-\bar{Z}_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij}-\bar{Z}_{i.})^2} \qquad (3)$$

where k **=** is the number of different groups to which the sampled cases belong

$$Z_{ij} = \left|Y_{ij} - \bar{Y}_i\right|$$

Ni = is the number of cases in the j^{th} group; *N* **=** is the total number of cases in all groups; Z_{ij} = is the value of the measured variable for the j^{th} case from the i^{th} group; Y_{ij} = original values of the dependent measures; $\bar{Y}$ = the mean of the i^{th} subgroup; $Z_{i.}$ = the mean of Z_{ij} for i^{th} sample; $Z_{..}$ = the overall mean.

Brown-Forsythe Test (BF): The Brown-Forsythe test is used for checking the equality of the variances among the groups of one variable. The base of the test is really similar to the Levene's test, but the Brown-Forsythe test uses the deviations from the group medians instead of the mean, that is $Z_{..}$ becomes the median Howard, et al (2010).

$$W = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k}N_i(\bar{Z}_{i.}-\bar{Z}_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij}-\bar{Z}_{i.})^2} \qquad (4)$$

O'Brien Test (OB): The test statistics is given below.

$$r = \frac{\left[(w+n_j-2)n_j(X_{ij}-\bar{X})^2 - w{s_j}^2(n_j-1)\right]}{[(n_j-1)(n_j-2)]} \qquad (5)$$

where w = weight (usually 0.5); n_j = size of the j^{th} group; s_j^2 = variance of the j^{th} group

X_{ij} = Score of the i^{th} person in the j^{th} group; $\bar{X}$ = mean of the j^{th} group. Every raw score r_{ij} is transformed using the above formula. The O'Brien test statistics will be the F-value computed on applying the usual ANOVA procedure on the transformed scores with $F_{k-1,n-k}$ degree of freedom.

Z-Variance Test (ZV): The test statistics is given as:

$$F = \frac{\sum_{i=1}^{k} Z_i{}^2}{k-1} \qquad (6)$$

Where

$$Z_i = \sqrt{\frac{c(n_i-1)s_i{}^2}{MSE}} - \sqrt{c(n_i-1) - \frac{c}{2}} \qquad (7)$$

$$c = 2 + \frac{1}{n_i} \qquad (8)$$

s_i^2 = the unbiased estimate of variance for the i^{th} subgroup; n_i = sample size of the i^{th} subgroup; MSE = pooled within group error variance

Hartley Test or F

Hartley Test or $\mathbf{F_{max}}$ Test (HF): The test statistic is just a simple ratio between the largest subgroup variance and the smallest:

$$F_{max} = \frac{S^2{}_{max}}{S^2{}_{min}} \qquad (9)$$

A table of values created by Hartley evaluates the test statistic with degrees of freedom of k and $\boldsymbol{n_i} - 1$ (if all the groups have the same size). When the groups have slightly different sample sizes, the harmonic mean may serve as the adjusted sample size *n* Zhang (1998).

$$\text{Harmonic mean} = \frac{k}{\sum n_j^{-1}} \qquad (10)$$

Cochran's G Test (CG): The test statistics is:

$$G = \frac{S^2{}_{max}}{kMS_{error}} \qquad (11)$$

$$MS_{error} = \frac{\sum(X_{ij}-\bar{X})^2}{N-k} \qquad (12)$$

$S^2{}_{max}$ = maximum variance; *k* = number of groups; *N* = number of all sample sizes

3. Result

The summary of the results from the analysis for one-way and two-way ANOVA are shown in Tables 1, 2, 3 and 4 for the seven HOV test considered.

Table 1: Type I error rates summary for One-way ANOVA: Minimum and maximum empirical Type I error rates and number of times the Type I error rates was rejected when it is true

Equal sample size	Normal Distribution			Non-normal Distribution		
Test	Minimum empirical Type I error	Maximum empirical Type I error	Number of times Type I error was rejected	Minimum empirical Type I error	Maximum empirical Type I error	Number of times Type I error was rejected
Bartlett	0.075	0.969	0	0.010	0.870	4
Levene	0.057	0.783	0	0.040	0.680	3
Brown-Forsythe	0.237	0.960	0	0.140	0.980	0
O'Brien	0.140	0.807	0	0.036	0.780	1
Z-variance	0.203	0.776	0	0.024	0.591	4
Hartley F-max	0.384	0.789	0	0.011	0.441	2
Cochran's G	0.731	0.750	0	0.009	0.485	3

Unequal sample size						
Bartlett	0.007	0.996	2	0.050	0.930	0
Levene	0.010	0.950	4	0.006	0.820	3
Brown-Forsythe	0.236	0.942	0	0.029	0.997	1
O'Brien	0.048	0.993	1	0.025	0.730	1
Z-variance	0.031	0.953	3	0.010	0.852	1
Hartley F-max	0.020	0.991	2	0.067	0755	0
Cochran's G	0.001	0.877	1	0.042	0.932	1

Table 2: Power rates summary for One-way ANOVA: Minimum and maximum empirical Power rates and number of times the Type I error rates was rejected when it is false

Equal sample size	**Normal Distribution**			**Non-normal Distribution**		
Test	Minimum empirical power	Maximum empirical power	Number of times Type I error was rejected	Minimum empirical power	Maximum empirical power	Number of times Type I error was rejected
Bartlett	0.002	0.890	8	0.001	0.660	9
Levene	0.014	0.730	11	0.005	0.340	4
Brown-Forsythe	0.020	0.640	1	0.020	0.817	1
O'Brien	0.011	0.740	3	0.035	0.478	3
Z-variance	0.034	0.670	10	0.022	0.321	9
Hartley F-max	0.013	0.580	8	0.031	0.551	9
Cochran's G	0.006	0.750	6	0.010	0.214	8
Unequal sample size						
Bartlett	0.025	0.880	2	0.000	0.950	7
Levene	0.015	0.820	4	0.015	0.710	7
Brown-Forsythe	0.008	0.980	1	0.240	0.820	0
O'Brien	0.037	0.705	1	0.118	0.805	0
Z-variance	0.011	0.797	2	0.028	0.464	7
Hartley F-max	0.023	0.637	2	0.033	0.221	4
Cochran's G	0.047	0.789	4	0.120	0.432	6

Table 3: Type I error rates summary for Two-way ANOVA: Minimum and maximum empirical Type I error rates and number of times the Type I error rates was rejected when it is true

Equal sample size	**Normal Distribution**			**Non-normal Distribution**		
Test	Minimum empirical Type I error	Maximum empirical Type I error	Number of times Type I error was rejected	Minimum empirical Type I error	Maximum empirical Type I error	Number of times Type I error was rejected
Factor	A(B)	A(B)	A(B)	A(B)	A(B)	A(B)

Bartlett	0.26(0.17)	0.97(0.80)	0(0)	0.02(0.03)	0.97(0.23)	3(4)
Levene	0.19(0.02)	0.91(0.94)	0(3)	0.00(0.02)	0.96(0.42)	1(3)
Brown-Forsythe	0.36(0.24)	0.94(0.99)	0(0)	0.32(0.04)	0.87(0.81)	0(1)
O'Brien	0.21(0.03)	0.84(0.96)	0(2)	0.01(0.12)	0.95(0.55)	1(0)
Z-variance	0.21(0.19)	0.78(0.88)	0(0)	0.03(0.00)	0.95(0.67)	3(4)
Hartley F-max	0.49(0.07)	0.67(0.77)	0(0)	0.01(0.04)	0.81(0.48)	3(2)
Cochran's G	0.31(0.02)	0.91(0.99)	0(1)	0.00(0.00)	0.88(0.38)	2(5)
Unequal sample size						
Bartlett	0.14(0.06)	0.94(0.96)	0(0)	0.02(0.02)	0.78(0.74)	1(2)
Levene	0.02(0.01)	0.88(0.68)	3(1)	0.07(0.00)	0.92(0.61)	0(6)
Brown-Forsythe	0.06(0.02)	0.97(0.75)	0(1)	0.13(0.01)	0.98(0.94)	0(4)
O'Brien	0.05(0.08)	0.93(0.90)	0(0)	0.20(0.07)	0.67(0.64)	0(0)
Z-variance	0.11(0.00)	0.73(0.91)	0(1)	0.01(0.03)	0.82(0.58)	1(4)
Hartley F-max	0.01(0.04)	0.64(0.83)	1(2)	0.09(0.02)	0.79(0.84)	0(1)
Cochran's G	0.21(0.07)	0.97(0.72)	0(0)	0.17(0.01)	0.99(0.69)	0(1)

Table 4: Power rates summary for Two-way ANOVA: Minimum and maximum empirical Power rates and number of times the Type I error rates was rejected when it is false

Equal sample size	**Normal Distribution**			**Non-normal Distribution**		
Test	Minimum empirical power	Maximum empirical power	Number of times Type I error was rejected	Minimum empirical power	Maximum empirical power	Number of times Type I error was rejected
Factor	A(B)	A(B)	A(B)	A(B)	A(B)	A(B)
Bartlett	0.97(0.00)	0.99(0.20)	0(7)	0.00(0.00)	0.54(0.15)	3(8)
Levene	0.95(0.00)	0.99(0.28)	0(6)	0.01(0.00)	1.00(0.09)	2(8)
Brown-Forsythe	0.96(0.01)	0.99(0.82)	0(2)	0.49(0.00)	0.95(0.46)	0(3)
O'Brien	0.91(0.11)	0.99(0.49)	0(0)	0.04(0.04)	0.68(0.43)	1(2)
Z-variance	0.93(0.00)	0.98(0.55)	0(6)	0.02(0.00)	0.76(0.37)	3(8)
Hartley F-max	0.67(0.01)	0.87(0.67)	0(5)	0.01(0.01)	0.72(0.59)	4(7)
Cochran's G	0.71(0.04)	0.91(0.59)	0(6)	0.03(0.00)	0.88(0.24)	2(6)
Unequal sample size						
Bartlett	0.70(0.01)	0.90(0.79)	0(3)	0.00(0.00)	0.93(0.38)	4(5)
Levene	0.36(0.00)	0.69(0.40)	0(4)	0.03(0.00)	0.67(0.23)	3(5)
Brown-Forsythe	0.50(0.01)	0.73(0.81)	0(1)	0.28(0.00)	0.76(0.90)	0(3)
O'Brien	0.52(0.16)	0.85(0.90)	0(0)	0.12(0.14)	0.57(0.95)	0(0)
Z-variance	0.08(0.04)	0.89(0.51)	0(3)	0.01(0.01)	0.71(0.57)	4(4)
Hartley F-max	0.09(0.03)	0.85(0.66)	0(1)	0.00(0.00)	0.61(0.74)	4(4)
Cochran's G	0.22(0.16)	0.71(0.47)	0(0)	0.03(0.01)	0.77(0.62)	4(2)

4. Discussion and Conclusion

Having observed the type-one-error rate and power of each HOV test, the following deductions were made from this study: All the HOV tests perform optimally under the condition of normality and equal sample sizes. This shows that population distribution and unequal sample sizes significantly affects type-one-error control and power of HOV tests. The Bartlett and Levene tests maintained the highest power for one-way ANOVA under normality and non-normality respectively with equal and unequal sample sizes. Under normality and non-normality with equal and unequal sample sizes, the O'Brien and Brown-Forsythe tests committed the least type I error for one-way ANOVA. Nevertheless, they also have the lowest powers for one-way ANOVA. Under normality with equal and unequal sample sizes, the Bartlett test committed no type-one-error for two-way ANOVA. It is very robust. Under non-normality with equal and unequal sample sizes, the O'Brien committed the least type-one-error for two-way ANOVA. Under normality and non-normality with equal and unequal sample sizes, the Bartlett, Levene and Z-variance maintained the highest power for two-way ANOVA. In order to achieve better results in terms of statistical power and type-one-error in testing for HOV when the data set are either normal or non-normal with equal and unequal sample sizes based on the results obtained from this study, we recommended that the Bartlett and O'Brien tests should be used to test for HOV under normality with equal and unequal sample sizes for one-way and two-way ANOVA. The Levene and Brown-Forsythe tests should be used to test for HOV under non-normality with equal and unequal sample sizes for one-way and two-way ANOVA.

References

1. Garbunova A. A. & Lemeshko B. Y. (2012). Application of variance homogeneity tests under violation of normality assumption. *Applied Methods of Statistical Analysis* 6, 28-36.
2. Howard B., Gary S., Kalz & Restori F. A. (2010). A Monte Carlo study of seven homogeneity of variance tests. *Journal of Mathematics and Statistics,* 6, 359-366.
3. Koning, A. J. (2014). Homogeneity of variances. In Wiley statsref: *Statistics Reference Online*. John Wiley & Sons Online library.
4. Parra-Frutos I. (2012). Testing homogeneity of variance with unequal sample sizes. *Computational Statistics*. 28, 1269-1297
5. Peter, S. (2013*).* Data assumption: Homogeneity of variance (univariate tests). Blog post:
6. Vanhove, J. (2018). Causes and consequences of unequal sample sizes. blog post: https://janhove.github.io/design/2015/11/02/unequal-sample-sized.

7. Zhang, S. (1998). Fourteen homogeneity of variance tests: When and how to use them. (Tech. Rep.). University of Hawaii at Manoa. (18p. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA).

On identification of influential observations in nonlinear regression models

T. von Rosen[1], K. Stal[2]

[1]Department of Statistics, Stockholm University, Stockholm, Sweden

[2]The Swedish National Agency for Education, Stockholm, Sweden

Abstract

New diagnostic measures based on the empirical influence function are developed for detecting influential observations in nonlinear regression models. The focus is on a regression model with a known mean function, which is nonlinear in its parameters and where the function is chosen according to the knowledge about the process generating the data. The error term in the regression model is assumed to be additive. Influential observations are detected using a stepwise procedure which also takes care about masking effect. The proposed methodology with some new graphical tools is illustrated by two examples.

Keywords

Added parameter plot; differentiation approach; influential observation; nonlinear models

1. Introduction

Nonlinear regression models found numerous applications in various disciplines, for example in pharmacokinetics and biological sciences, due to their ability to describe complicated, nonlinear relationships between variables. The Michaelis-Menten model (Michaelis and Menten, 1913) is a nonlinear regression model that is widely used to study enzymatic-catalyzed reactions, called enzyme kinetics. This model describes the behavior of the enzymatic reaction's velocity when adding different substrate concentrations to the process. Moreover, the parameters in the Michaelis-Menten model have chemically meaningful interpretations.

The parameter estimates and the results of hypothesis testing in statistical models are based on the observed data. It is well understood that some observations have a larger impact on the results of estimation and testing procedure than others. Small perturbations in the data causing significant changes in the results, indicate that the reliability of the conclusions and the inference can be compromized. Hence, an essential and inseparable part of statistical modeling is model diagnostic which can shed light on the presence of outliers and influential observations in the complex data sets.

The methods for identifying possibly influential observations have been extensively studied in various regression models. Local influence analysis introduced by Cook (1986) has found numerous applications and extensions (e.g. Beckman et al., 1987; Lawrance, 1988; Thomas and Cook, 1990; Wu and Luo, 1993; St. Laurent and Cook, 1993; Shi, 1997; Poon and Poon, 1999; Zhu and Lee, 2001; Zhu et al., 2007). The existing literature on influence analysis in nonlinear regression is not as extensive as for linear regression. One reason for this can be that there do not generally exist closed form estimators for the parameters in the nonlinear regression model.

Detection of influential observations on the fit of the nonlinear regression model is discussed by Cook and Weisberg (1982) and St. Laurent and Cook (1993). Cook and Weisberg (1982) developed a nonlinear version of Cook's distance, and St. Laurent and Cook (1993) proposed an approach for assessing the influence of the observations on the fitted values and on the estimate of the variance in a nonlinear regression model. Detection of multiple influential observations is in general a more difficult task due to masking and swamping effects (Atkinson, 1985; Rousseeuw & Leory, 1987; Chatterjee & Hadi, 1988; Lawrence, 1995). Masking occurs when an observation is not identified as influential unless another observation is deleted first. Swamping occurs when "good" observations are incorrectly identified as influential because of the presence in data of another observation. The available approaches to dealing with the problem of masking effects include the use of multiple 1 case deletions (Chatterjee and Hadi, 1988; Hadi and Simonoff, ; Lawrance, 1995) or stepwise procedures (Belsley et al., 1980; Bruce and Martin, 1989; Shi and Huang, 2011).

The aim of this article is to develop influence measures to assess the influence of a single observation as well as multiple influential observations on the parameter estimates in nonlinear regression models. The proposed measures allow to simultaneously assess the influence of several observations on the parameter estimates. This type of influence will be referred to as joint influence. Furthermore, it makes it possible to evaluate the influence that the kth observation has on the parameter estimates after another observation, say observation *i*, has been deleted. The type of influence that the kth observation has on the parameter estimates after the deletion of the ith observation is called conditional influence.

2. The influence measure *DIM*

Consider the following nonlinear model with an additive error term

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y}$ is the n-vector of responses, the known matrix $\boldsymbol{X} : n \times p$ comprises explanatory variables, $\boldsymbol{\theta}$ is a q-vector of unknown parameters, the vector of random errors $\epsilon \sim N(O_n, \sigma^2 I_n)$, O_n and I_n denote the n-vector with all elements equal to zero and the identity matrix of size n, respectively. The function f is assumed to be twice differentiable in $\boldsymbol{\theta}$, and

$$\boldsymbol{f}(\boldsymbol{X},\boldsymbol{\theta}) = (\boldsymbol{f}(\boldsymbol{X}_1,\boldsymbol{\theta}),\dots,\boldsymbol{f}(\boldsymbol{X}_n,\boldsymbol{\theta}))^T = (\boldsymbol{f}_1(\boldsymbol{\theta}),\dots,\boldsymbol{f}_n(\boldsymbol{\theta}))^T.$$

Influence analysis in nonlinear regression is not widely explored. We propose a new influence measure for assessing the influence of single and multiple observations on the parameter estimates in a nonlinear regression model using differentiation approach. The differentiation approach is used in linear regression to construct the influence measure $EIC_{\widehat{\beta},k}$ where the *EIC* stands for *Empirical Influence Curve* (see e.g. Belsley *et al.* (1980), Chatterjee and Hadi (1988) and Cook and Weisberg (1982)). Originally *EIC* measure was derived from the influence curve, a theoretical concept introduced by Hampel (1974).

We extend the ideas of using the differentiation approach for measuring the influence of an observation on the parameter estimate in nonlinear regression models. Two new influence measures for the parameter estimates in the nonlinear regression model (1) will be derived: the $DIM_{\widehat{\theta},k}$ and $DIM_{\widehat{\vartheta}j',k}$. The abbreviation stands for *Differentiation approach & Influence Measure*. The first diagnostic measure, $DIM_{\widehat{\theta},k}$, is used to assess the influence of a single observation on all parameter estimates in the model, simultaneously. It is constructed when all parameters are estimated from a perturbed model, presented in (2) later on, and it is referred to as the joint-parameter influence measure. The $DIM_{\widehat{\vartheta}j',k}$., on the other hand, is used to assess the influence of a single observation on the *j*th parameter estimate in the model. When constructing $DIM_{\widehat{\vartheta}j',k}$., only the *j*th parameter is estimated from the perturbed model, later defined in (2): the other parameters are estimated from an unper-b turbed model and regarded to be known. The $DIM_{\widehat{\theta}j,k,}$ is referred to as the marginal-parameter influence measure. We will now start with the definition of $DIM_{\widehat{\theta},k}$. Consider the following perturbed nonlinear model

$$\boldsymbol{y_\omega} = \boldsymbol{f}(\boldsymbol{X},\boldsymbol{\theta}) + \varepsilon_\omega, \tag{2}$$

where $\varepsilon_\omega \sim N_n\big(\boldsymbol{0},\sigma^2\boldsymbol{W}^{-1}(\omega_k)\big), \omega_k$ is the weight such that 0 $< \omega_k \leq$ 1 and the weight matrix $\boldsymbol{W}(\omega_k) = diag(1,\dots.,\omega_k,\dots,1)$.

Definition 2.1. *The influence measure for assessing the influence of the kth observation on $\widehat{\boldsymbol{\theta}}$ is defined as the following derivative*

$$DIM_{\widehat{\theta,k}} = \frac{d}{d\omega_k}\theta(\omega_k)\Big|_{\omega k=1,} \tag{3}$$

where $\widehat{\boldsymbol{\theta}}(\omega_k)$ is the weighted least squares estimate of $\boldsymbol{\theta}$ in the perturbed model (2).

Observe that, in Definition 2.1, if $\omega_k \to$ 1, then $\widehat{\boldsymbol{\theta}}(\omega_k) \to \widehat{\boldsymbol{\theta}}$, the unweighted least squares estimate.

To calculate the $DIM_{\hat{\theta},k}$ in (3) we need an estimator for $\boldsymbol{\theta}$ in the perturbed model (2). Using the method of weighted least squares, which is equivalent to the maximum likelihood approach, we have to find $\hat{\boldsymbol{\theta}}$ that minimizes the following

$$Q(\omega_k) = \left(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X},\boldsymbol{\theta})\right)^T \boldsymbol{W}(\omega_k)(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X},\boldsymbol{\theta}))$$

Differentiating $Q(\omega_k)$ with respect to $\boldsymbol{\theta}$ and solving normal equations using iterative methods (e.g. the GaussNewton method), the obtained least squares estimate of $\boldsymbol{\theta}$ is obtained as a function of ω_k. In the next theorem, the explicit expression of $DIM_{\widehat{\theta,k}}$, defined in (3), is presented.

Theorem 2.1. Let $DIM_{\widehat{\theta,k}}$ be given in Definition 2.1. Then

$$DIM_{\widehat{\theta,k}} = \boldsymbol{r_k F_k^T}(\hat{\boldsymbol{\theta}})\left((\boldsymbol{F}(\hat{\boldsymbol{\theta}})\boldsymbol{F^T}(\hat{\boldsymbol{\theta}}) - \boldsymbol{G}(\hat{\boldsymbol{\theta}})(\boldsymbol{r} \otimes \boldsymbol{I_q})\right)^{-\mathbf{1}},$$

provided that the inverse exists, where

$$r = (r_k) = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X},\hat{\boldsymbol{\theta}}), \tag{4}$$

$$\boldsymbol{F}(\hat{\boldsymbol{\theta}}) = \left(\boldsymbol{F}_1(\hat{\boldsymbol{\theta}}), \dots, \boldsymbol{F}_n(\hat{\boldsymbol{\theta}})\right) = \left.\frac{d\boldsymbol{f}(\boldsymbol{X},\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right|_{\theta=\hat{\theta}}, q \times n. \tag{5}$$

and

$$G(\hat{\boldsymbol{\theta}}) = \left(\frac{d}{d\theta}\left(\frac{df(\boldsymbol{X},\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right)\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{d\boldsymbol{F}(\widehat{\boldsymbol{\theta}})}{d\hat{\boldsymbol{\theta}}}, q \times nq \tag{6}$$

The $DIM_{\hat{\theta},k}$ derived in Theorem 2.1 measures the influence of the kth observation on all the parameter estimates in model (1) simultaneously. Therefore, $DIM_{\hat{\theta},k}$ is regarded to be a joint-parameter influence measure. However, it can be useful to measure the influence of the kth observation on a particular parameter estimate of the model. In order to assess the influence of the k observation on the jth parameter estimate, $\hat{\theta}_j$, a marginal-parameter influence measure will be defined, and its explicit expression will be derived.

Consider the perturbed model (2). Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_j)$ be a vector of parameter estimates, where $\hat{\boldsymbol{\theta}}_1 = (\hat{\theta}_1, \dots, \hat{\theta}_{j-1}, \hat{\theta}_{j+1}, \dots \hat{\theta}_q)^T$, are the maximum likelihood estimates in the unperturbed model (1) and $\hat{\theta}_j$ is estimated from the perturbed model (2), with parameter estimates $\hat{\boldsymbol{\theta}}_1$ inserted and regarded as known.

Definition 2.2. *The marginal influence measure for assessing the influence of the kth observation on the parameter estimate $\hat{\theta}_j$ is defined as the following derivative.*

$$DIM_{\widehat{\theta,k}} = \frac{d}{d\omega_k}\theta(\omega_k)\bigg|_{\omega k=1,}, \tag{7}$$

Where $\hat{\theta}_j(\omega_k)$ is the weighted least squares estimate of θ_j , given $\hat{\boldsymbol{\theta}}_1 = (\hat{\theta}_1, \dots, \hat{\theta}_{j-1}, \hat{\theta}_{j+1}, \dots \hat{\theta}_q)^T$.

Observe that, in Definition 2.2, if $\omega_k \to 1$, then $\hat{\boldsymbol{\theta}}(\omega_k) \to \hat{\boldsymbol{\theta}}$, the unweighted least squares estimate.

The weighted least squares criterion and the normal equation for the case when a single parameter is estimated from the perturbed model are given by

$$Q(\omega_k) = \big(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_1, \hat{\theta}_j)\big)^T \boldsymbol{W}(\omega_k)\big(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_1, \hat{\theta}_j)\big),$$

and

$$\frac{d\boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_1, \hat{\theta}_j)}{d\theta_j}\boldsymbol{W}(\omega_k)\left(y - \boldsymbol{f}\left(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_1, \hat{\theta}_j(\omega_k)\right)\right) = 0. \tag{8}$$

In the next theorem, an explicit expression of the marginal-parameter influence diagnostic $DIM_{\hat{\theta}\,j,k}$ defined in (7) will be provided.

Theorem 2.2. Let $DIM_{\hat{\theta}\,j,k}$be given in Definition 2.2. Then

$$DIM_{\hat{\theta}\,j,k} = r_k F_k(\hat{\theta}_j)(\boldsymbol{F}(\hat{\theta}_j)\boldsymbol{F}^T(\hat{\theta}_j) - \boldsymbol{G}(\hat{\theta}_j)\boldsymbol{r})^{-1},$$

provided that the inverse exists, where $\boldsymbol{r} = (r_k) = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_1, \hat{\theta}_j)$,

$$\boldsymbol{F}(\hat{\theta}_j) = \left(\boldsymbol{F}_1(\hat{\theta}_j), \dots, \boldsymbol{F}_n(\hat{\theta}_j)\right) = \frac{d\boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_1, \hat{\theta}_j)}{d\theta_j}\bigg|_{\theta j=\widehat{\theta J}}, 1 \times n, \tag{9}$$

$$\boldsymbol{G}(\hat{\theta}_j) = \frac{d\boldsymbol{F}(\hat{\theta}_j)}{d\hat{\theta}_j}\bigg|_{\theta_j=\hat{\theta}_j} = \frac{d^2\boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_1, \hat{\theta}_j)}{d\theta_j^2}\bigg|_{\theta_j=\hat{\theta}_j}, 1 \times n. \tag{10}$$

A note on $DIM_{\hat{\theta},k}$*and* $DIM_{\hat{\theta}\,j,k}$

When deriving the influence measures and studying the single observations' influence on the parameter estimates we observe some interesting aspects of influence analysis in nonlinear regression. A benefit of using the differentiation approach, where we compute derivatives of various quantities with respect to ω_k and evaluate the derivatives at $\omega_k = 1$, is that no additional iterations for computing the parameter estimates are needed. An alternative way of using the differentiation approach is to evaluate the same derivatives as $\omega_k \to 0$. If this approach were to be used instead, the explicit expressions of $DIM_{\widehat{\theta},k}$ and $DIM_{\widehat{\theta}j,k}$ would be functions of the parameter estimates with weights attached. As an example, consider the following derivative

$$\left.\frac{d\boldsymbol{f}(\boldsymbol{X},\widehat{\boldsymbol{\theta}}(\omega_k))}{d\omega_k}\right|_{\omega_k\to 0} = \boldsymbol{F}\big(\widehat{\boldsymbol{\theta}}(\omega_k)\big),$$

where $\omega_k \to 0$. This means that we would need to compute a parameter estimate for each k and additional iterations are needed. On the contrary, with the new proposed method in this thesis

$$\left.\frac{d\boldsymbol{f}(\boldsymbol{X},\widehat{\boldsymbol{\theta}}(\omega_k))}{d\omega_k}\right|_{\omega_k=1} = \boldsymbol{F}\big(\widehat{\boldsymbol{\theta}}\big),$$

which is the derivative of the expectation function from the unperturbed model (1) and hence, no additional iterations are needed.

We can further make a comparison between the proposed measure, $DIM_{\widehat{\theta},k}$, and the nonlinear version of Cook's distance and given by

$$\frac{\left(\widehat{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_{(k)}\right)^T \boldsymbol{F}(\widehat{\boldsymbol{\theta}})\boldsymbol{F}^T(\widehat{\boldsymbol{\theta}})\left(\widehat{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_{(k)}\right)}{q\widehat{\sigma}^2}, \quad k = 1,\ldots,n,$$

where q is the number of parameters in the model, $\widehat{\boldsymbol{\theta}}_{(k)}$ is the estimate of $\boldsymbol{\theta}$ when the kth observation is excluded from the calculations and $\boldsymbol{F}\big(\widehat{\boldsymbol{\theta}}\big)$ is defined in (5). The nonlinear version of Cook's distance is based on case-deletion. A consequence of this is that re-estimation of the parameters is needed for every observation we are interested in. Thus, the nonlinear version of Cook's distance demands additional iterations when estimating the parameters, which is avoided using our measure $DIM_{\widehat{\theta},k}$.

3. Assessment of influence of multiple observations

Thus far, we have discussed the differentiation approach to the detection of single influential observations. However, in practice it is likely that a data set contains more than one influential observation. Influence analysis concerning multiple observations is a more challenging problem since multiple influential observations can be more difficult to detect. We will borrow the idea of using the" directional" derivative and define the influence measure $DIM_{\widehat{\beta},k}$ for assessing the influence of multiple observations on $\hat{\beta}$.

a. Joint influence in nonlinear regression

Consider the following perturbed nonlinear model

$$\boldsymbol{y}_{\boldsymbol{\omega}} = \boldsymbol{f}(\boldsymbol{X},\boldsymbol{\theta}) + \ \varepsilon_{\omega}, \qquad (11)$$

where $\varepsilon_\omega \sim N_n\left(\mathbf{0}, \sigma^2 \boldsymbol{W}^{-\mathbf{1}}(\boldsymbol{\omega})\right), \boldsymbol{W}\,(\boldsymbol{\omega}): n \times n$ is a diagonal weight matrix, with diagonal elements $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$ and where $0 < \omega_k \leq 1$, for $k = 1, \ldots, n$. Also, let K be the subset containing the indices of the observations for which we would like to assess influence.

We present the following definition

Definition 3.1. *The diagnostic measure for assessing the influence of the observations with indices specified in the subset K, on the parameter estimate* $\widehat{\boldsymbol{\theta}}$*, is defined as the following derivative*

$$DIM_{\widehat{\boldsymbol{\theta}},K} = \boldsymbol{\ell}^T \left.\frac{d\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})}{d\boldsymbol{\omega}}\right|_{\boldsymbol{\omega}=\mathbf{1}_n}, \tag{12}$$

where $\boldsymbol{\ell} : n \times 1$ is a vector with nonzero entries in the rows with indices in K, where $||\boldsymbol{\ell}|| = \sqrt{\boldsymbol{\ell}^T \boldsymbol{\ell}} = 1$ and where $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$ is the weighted least squares estimate of $\boldsymbol{\theta}$, which is a function of the weight $\boldsymbol{\omega}$.

If $\boldsymbol{\omega} \rightarrow \mathbf{1}_n$, then $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}) \rightarrow \widehat{\boldsymbol{\theta}}$, the unweighted least squares estimate.

To derive the $DIM_{\widehat{\theta},K}$, for assessing the influence of multiple observations simultaneously on the parameter estimates, we need the weighted least squares estimate of $\boldsymbol{\theta}$ in (11). The weighted least squares criterion, which should be minimized, is given by

$$Q(\boldsymbol{\omega}) = (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}))^T\, \boldsymbol{W}(\boldsymbol{\omega})\big(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta})\big).$$

There is generally no explicit solution to the normal equations and iterative methods are used to find an estimate. The obtained estimate of $\boldsymbol{\theta}$ is a function of the weights, $\boldsymbol{\omega}$, and is denoted $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$. The next theorem provides an explicit expression of the $DIM_{\widehat{\theta},K}$ defined in (12).

Theorem 3.1. *Let* $DIM_{\widehat{\theta},K}$ *be given in Definition 3.1. Then*

$$DIM_{\widehat{\theta},K} = \boldsymbol{\ell}^T \boldsymbol{U}^* \left(\boldsymbol{r} \otimes \boldsymbol{F}^{\boldsymbol{T}}(\widehat{\boldsymbol{\theta}})\right)\left(\boldsymbol{F}(\widehat{\boldsymbol{\theta}})\boldsymbol{F}^{\boldsymbol{T}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{G}(\widehat{\boldsymbol{\theta}})(\boldsymbol{r} \otimes \boldsymbol{I}_q)\right)^{-1},$$

provided that the inverse exists.
In the expression above, $\boldsymbol{U}^*: n \times n^2$ *is a matrix with row vectors* $\boldsymbol{u}_i^T$,

$$\boldsymbol{u}_i = \boldsymbol{d}_i \otimes \boldsymbol{d}_i, for\ i = 1, \ldots, n, \tag{13}$$

where d_i *is the* i *th column of the identity matrix of size n. The quantities* $\boldsymbol{r}, \boldsymbol{F}(\widehat{\boldsymbol{\theta}})$ *and* $\boldsymbol{G}(\widehat{\boldsymbol{\theta}})$ *are defined in (4), (5) and (6), respectively.*

The $DIM_{\hat{\theta},K}$ is a diagnostic measure for assessing the simultaneous influence of several observations on the parameter estimates, $\hat{\boldsymbol{\theta}}$. Since all parameters in the model are estimated from the perturbed model, $DIM_{\hat{\theta},K}$ is regarded to be a joint-parameter influence measure. It can be of interest to assess the influence of multiple observations on a particular parameter estimate, $\hat{\theta}_j$, in model (1). If this is the case, we use the same methodology as above and obtain a marginal-parameter influence measure.

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\theta}_j)$be a vector of parameter estimates, where

$$\hat{\boldsymbol{\theta}}_1 = (\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_q),$$

are the maximum likelihood estimates from the unperturbed model (1), and $\hat{\theta}_j$ is estimated from the perturbed model (11).

References

1. R.D. Cook, *Assessment of local inuence*, J. R. Stat. S. Ser. B 48 (1986), pp. 133{169.
2. D.A. Belsley, E. Kuh, and R.E. Welsch, (1980), *Regression Diagnostics: Identifying Inuential Data and Sources of Collinearity*, New York: John Wiley.
3. N. Billor and R.M. Loynes *Local inuence: A new approach*, Comm. Statist-Theory Meth. 22 (1993), pp. 1595{1611.
4. R.D. Cook *Detection of inuential observations in linear regression*, Technometrics 19 (1977), pp. 15{18. R.D. Cook *Inuential observations in linear regression*, J. Amer. Statist. Assoc. 74 (1979), pp. 169{174.
5. R.D. Cook and S. Weisberg (1982). *Residuals and Inuence in Regression*. New York: Chapman and Hall.
6. B. Efron and R.J. Tibshirani (1993). *An Introduction to Bootstrap*. New York: Chapman and Hall.
7. FUNG, W. K. (1993). Unmasking outliers and leverage points: a con_rmation. Jour. Amer. Statist. Assoc. 88, 515-519. A.S. Hadi and J.S. Simonoff *Procedures for the identi_cation of multiple outliers in linear models*, Jour. Amer. Statist. Assoc 88 (1993), pp. 1264{1272.
8. A.J. Lawrance *Regression transformation diagnostic using local inuence*, Jour. Amer. Statist. Assoc. 83 (1990), pp. 1067{1072.
9. P. Prescott *An approximation test for outliers in linear models*, Technometrics 17 (1975), 129{132.
10. S. Weisberg (1985). *Applied Linear Regression*. 2nd. Ed. New York: Wiley

Comparative model analysis of customer retention drivers of universal banks in Ghana: structural equation model and ordinal logit model approaches

Abdul-Aziz A.Rahaman[1], Albert Luguterah[2], Bashiru I. I. Saeed[1]
[1]Department of Mathematics & Statistics, Kumasi Technical University, Ghana
[2]Navrongo Campus, University for Development Studies, Ghana

Abstract

The competitive power and survival of a bank lies in the degree of its customer retention. It depends on a myriad of factors and varies from product to product. This realization has made industry players and academics pay increasing attention to customer retention. This paper compared the Structural Equation Modelling (SEM) against the Ordinal logit model using primary data obtained from customers of selected universal banks in Ghana. The SEM results showed that loyalty influence to a great deal the trustworthiness and corporate image of universal banks in Ghana. Also, service quality is a relevant contributing factor to empathy, reliability and assurance of universal banks' services. Service quality has both direct and indirect effect on customer retention. However, satisfaction was deemed statistically insignificant in contributing to customer retention. On the other hand, the ordinal logit model results noted that service quality dimensions on tangible, responsiveness, empathy and trust are relevant contributing factors to customer satisfaction at universal banks. However, reliability and assurance, in terms of, service quality aspects are not really contributing marginally to customer satisfaction.

Keywords

Structural Equation Modelling; Ordinal Logit Model; Customer retention; Service Quality

1. Introduction

Customer retention is very significant in the creation and maintenance of competitive advantage in the service industry (Ndubisi, 2007). There are economic advantages associated with retaining loyal customers as opposed to recruiting new ones. This realization has made industry practitioners and academics pay increasing attention to customer retention studies (Ndubisi, 2007 & Zineldin, 2006). Furthermore, the longer a loyal customer stays with a firm, the more profitable it is to that firm (Kim and Cha, 2002). Loyalty in service businesses refer to the customer's commitment to do business with a particular organization, purchasing their products (Anderson and Jacobson (2000). The service sector has produced approximately two-thirds of worldwide GNP from twenty first century (Kara et al., 2005). Within the huge service sector, the banking sector is one of the most important entities. Quality in service can be determined by the extent to which customers' needs and expectations can be satisfied (Banerjee, 2012). Moreover, the competitive power and survival of a bank lies in the degree of its customer satisfaction (Titko and Lace 2010). Banks therefore pay particular attention to customer satisfaction, (Kattack and Rehman 2010). Similarly, Abdullah and Rozario (2009) posit that the level of satisfaction may be influenced by various internal and external factors.

Prabhakaran (2003) mentioned that the customer is the king. High customer satisfaction is important in maintaining a loyal customer base. To link the service quality, customer satisfaction and customer retention is important (Kumar et al., 2009). Saif (2009) found that customer satisfaction is the outcome of service quality. Researchers argued that service quality has influence on customer satisfaction and generates customer retention (Chang et al., 2009). Zeithaml et al., (2008) developed a conceptual model that correlates Service Quality, Customer Satisfaction and Customer retention in one frame. There is a positive relationship between the two constructs (Beerli et al., 2004). The relationship between customer satisfaction and service quality is debatable. Some researchers argued that service quality is the antecedent of customer satisfaction, while others argued the opposite relationship holds. This finding was further supported by Parasuraman et al., (1993). Most of the researchers found that service quality is the antecedent of customer satisfaction (Athanassopoulos and Iliakopoulos, 2003; Lee and Hwan, 2005; Naeem and Saif 2009; Balaji, 2009; Bedi, 2010; Kassim and Abdullah, 2010; Kumar et al., 2010). Yee et al (2010) found that service quality has a positive influence on customer satisfaction. A lot of factors that drive customer retention, based on service quality in particular, need to be examined in order to reliably measure it. Against this backdrop, this article seeks to analyse customer among universal banks in Ghana using a comparative analysis of both Structural Equation Modelling (SEM) and the Ordinal logit model approaches.

2. Methodology

2.1 Sampling Technique and Sample Size

In selecting the sample of customers, stratified random sampling technique was employed. The study sample consisted of 1,050 customers, of twenty-nine (29) universal banks, drawn from across five cities of the various regions comprising southern Ghana. Subsequently, a simple random sampling technique was used to select customers. The study employed self-administered questionnaires to collect data from the respondents. Moreover, the variables in the questionnaire relating customer retention were in five

Likert-scale form and rated as: Strongly Disagree=1, Disagree=2, Not sure=3, Agree=4 and Strongly Agree=5.

2.2 Model Specification, Estimations and Tests

2.2.1 Structural Equation Modelling versus Ordinal Logistic Regression

The four steps of SEM, specification, identification, estimation, and model evaluation, are examined and an example is introduced to clarify these concepts. In addition, model diagnostics that have been developed under the SEM framework are briefly discussed which can be viewed as part of model evaluation. However, in considering methods for Likert scale responses having more than two possible options, a number of methods have been developed for handling the various possibilities. The most appropriate method developed for this case is the *ordinal logit* concept (Agresti, 2002).

2.2.2 Latent Variable Model for Structural Equation

From the Fig. 1 below, considering the SEM framework, latent variables are considered to either be exogenous, such as ξ_1, as their causes lie outside the model, or endogenous, like η_1 and η_2, as their causes lie within the model. In Figure 1, it is hypothesized that ξ_1 is a cause of both η_1 and η_2 and that η_1 is a cause of η_2. The latent variable model for the hypothetical model in Figure 1 can be written in equation form as:

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1 \quad (1)$$

$$\eta_1 = \beta_{21}\eta_1 + \gamma_{21}\xi_2 + \zeta_2 \quad (2)$$

The random errors ζ_1 and ζ_2 are assumed to have an expected value of zero and homoskedastic variances and uncorrelated with ζ_1. Thus (1) and (2) can be written more compactly as

$$\eta = B\eta + \Gamma\xi + \zeta \quad (3)$$

2.2.3 Measurement Model for Structural Equation

The measurement model links the latent variables with observed variables (the terms observed variables, indicators, measures, and manifest variables are used interchangeably). The example in Figure 1 posits that each latent variable has three indicators, each of which is associated with only one factor. The indicators for η_1 are y_1 , y_2 and y_3 , the indicators for η_2 are y_4 , y_5 and y_6 , and the indicators for ξ_1 are x_1 , x_2 , x_3 , x_4 and x_5 . The measurement model associated with Figure 1 is written more compactly in matrix notation as:

$$x = \Lambda_x\xi + \delta \quad (4)$$

$$y = \Lambda_y\eta + \epsilon \quad (5)$$

2.2.4 The Ordinal Logit Model

The following notation would employed in the model. Let $\pi_{i2} = \Pr(Y_i = 2)$ denotes the probability that the ith individual's outcome belongs to the second class; More generally, $\pi_{ik} = \Pr(Y_i = k)$ denotes the probability that the ith individual's outcome belongs to the kth class. On the other hand when the categories are ordered to assume that the log odds of $Y \geq k$ is linearly linked with the predictor variables. The model is given by

$$\log\left(\frac{\pi_k + \cdots + \pi_k}{1 + \cdots + \pi_{k-1}}\right) = \beta_{0k} + X^T\beta \qquad (6)$$

Thus, we still have to estimate K −1 intercepts, but only p linear effects, where p is the number of explanatory variables (note that $K + p - 1 < (K - 1)(p + 1) if\ K > 2$.

2.2.5 Goodness of Fit Test for Structural Equation Model

A large class of omnibus tests exists for determining overall model fit. The χ2 statistic is often used, for which the null hypothesis indicates how close the default model is to the data set used.

2.2.6 Goodness-of-Fit Test for Ordinal Logit Model

From the observed and expected frequencies, the usual Pearson and Deviance goodness-of-fit measures can be computed. The Pearson goodness-of-fit statistic is

$$x^2 = \Sigma\Sigma\left(\frac{O_{ij} - E_{ij}}{E_{ij}}\right)^2 \qquad (7)$$

3. Result

Asymptotically distribution free method was adopted for parameter estimation to justify data set used. Measurement model and structural model test were used to test fitness of the model.

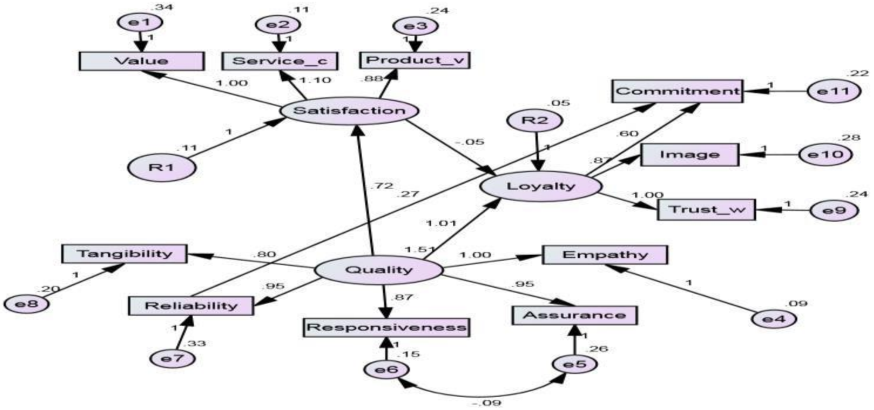


Figure 1: Outcome of Hypothesised Structural Model

Figure 1 depicted the empirical results of structural model by path analysis. The path coefficients of the latent constructs are visualized in Figure 1. The empirical results found significant positive relationship among service quality, customer satisfaction customer retention. It is notable that there is a direct effect of service quality on both customer satisfaction and loyalty. On the other hand, quality service has also significant direct but a statistically insignificant indirect effect on customer retention. The effect of Satisfaction of customers on Service charge is high for customers who strongly agree than their counterparts. Also, the effect of customer retention on Corporate Image is about 86.7% smaller for customers who strongly agree as compared to their counterparts. Again, ehe effect of the bank's Srtvie quality on Assurance is approximately 95% lower for those who strongly agree compared to their couterparts. However, there is a very weak negative relationship (-0.09) between the covaried error terms, 5 and 6, which is statistically significant at p<0.01

Table 2. The Ordinal Logistic Model

						95% Confidence *Interval*	
		Estimate	Std. Error	Wald	P-value	Lower Bound	Upper Bound
Threshold	[CL = 1]	-0.694	0.939	0.546	0.460	-2.534	1.146
	[CL = 2]	2.444	0.977	6.258	0.012	0.529	4.359
	[CL = 3]	4.527	1.111	16.595	0.000	2.349	6.705
	[CL = 4]	9.972	1.878	28.190	0.000	6.291	13.653
Location	[TA=1]	3.650	1.837	3.947	0.996	0.049	7.252
	[TA=2]	-0.636	1.036	0.377	0.794	-2.668	1.395
	[TA=3]	-0.783	0.793	0.974	0.003	-2.337	0.772
	[TA=4]	-1.657	0.679	5.959	0.036	-2.987	-0.327
	[RL=1]	0.810	2.643	0.094	0.759	-4.369	5.990
	[RL=2]	2.371	1.396	2.885	0.089	-.365	5.106
	[RL=3]	0.257	0.883	0.085	0.771	-1.473	1.988
	[RL=4]	-0.302	0.758	0.159	0.690	-1.787	1.183
	[RS=1]	-0.219	1.594	0.019	0.467	-3.343	2.906
	[RS=2]	-0.080	1.083	0.005	0.317	-2.203	2.043
	[RS=3]	-0.262	0.937	0.078	0.044	-2.099	1.574
	[RS=4]	0.706	0.584	1.458	0.034	-0.440	1.851
	[AS=1]	-9.541	4.546	4.405	0.036	-18.452	-0.631
	[AS=2]	-0.033	1.188	0.001	0.978	-2.362	2.296
	[AS=3]	1.165	0.897	1.689	0.194	-0.592	2.922
	[AS=4]	-0.329	0.765	0.185	0.667	-1.828	1.170
	[EM=1]	5.044	2.577	3.831	0.049	-0.007	10.094
	[EM=2]	0.200	1.185	0.028	0.866	-2.124	2.523
	[EM=3]	0.035	0.915	0.001	0.970	-1.758	1.828
	[EM=4]	-0.327	0.698	0.219	0.639	-1.696	1.042
	[TR=1]	6.672	3.451	3.738	0.053	-0.092	13.436
	[TR=2]	7.367	1.885	15.281	0.000	3.673	11.061
	[TR=3]	1.917	0.841	5.193	0.023	0.268	3.566
	[TR=4]	0.830	0.724	1.314	0.001	-0.589	2.248

From the observed significance levels (p<0.05) in Table 2 below, it can be seen that four factors out of the six service quality dimensions were statistically significant in influencing a customer retention. These dimensions include; tangibility, responsiveness, empathy and trust. Meanwhile, customers who agree to tangibility are more likely to assign higher ratings on loyalty than their counterparts who do not disagree. Also, universal banks customers who agree on the *bank's responsiveness* are more likely to assign higher ratings for loyalty than customers who think otherwise. Interestingly, customers who disagree on the dimension of *empathy* are more likely to assign high ratings for loyalty than those who agree. Moreover, customers who agree on the dimension of *trust* in the bank are more likely to assign higher ratings for loyalty than their counterparts who just disagree.

However, service quality dimensions including reliability and assurance were each not statistically significant. This means that each of these service quality dimension marginally influence customer retention. To have a more rigorous interpretation for the customer retention with the mediation of customer Satisfaction, the Goodness of fit indices need to be assessed. Also the GFI = 0.963, NFI = 0.934, CFI = 0.941, and IFI = 0.941. All the incremental fit measures fulfil the cut-off values (suggested values). Therefore, the model can be said to be a good fit model. However, the χ^2 statistic of 788.084 (df=39) is large. The χ^2statistic for model fit is still significant, meaning that the null hypothesis of a good fit to the data can be rejected. This could be due the large sample size used here since the χ^2 test is widely recognized to be problematic. It is sensitive to sample size, and it becomes more and more difficult to retain the null as the number of cases increases, which may lead to the rejection of a good model or the retention of bad ones. The RMSEA likewise suggests that the fit of the model is just about tolerable. The value of 0.083 exceeds the 0.05 cut-off value for accepting the model fit.

Table 3. Model Fit

Model	-2LogLikelihood	Chi-Square	df	P-value
Intercept Only	243.057			
Final	125.468	117.589	24	0.000

From Table 3 above, it can be noted that the difference between the two log-likelihoods with Chi-square distribution has a p-value less than the significance level 0.05 (i.e. p<0.05). This indicates that there is sufficient basis to reject the null hypothesis and therefore conclude that the final model gives a significant improvement over the baseline intercept-only model. Hence the final model gives better predictions than if you just guessed based on the marginal probabilities for the outcome categories.

It was also observed that the p-value (0.640) is greater than the significance level (0.05). This means that we fail to reject the null hypothesis that the fitted model is consistent with the observed data. Thus we conclude that the data and the model predictions are similar at 95% confidence level which implies a good model. The Pseudo R-square (Nagelkerke=76.8%) indicates that the predictor variables explains most of the proportions of variation between customer satisfaction (response). There is however about 23.2% of the variability which is unaccounted for, which may be due to research related errors. Additionally, in the SEM framework, Service Quality accounted for about 83.9%, 82.8% and 80.5% of the variability recorded in Assurance, Tangibility and Reliability respectively. Meanwhile, Loyalty explained about 86.2%, 83% and 80.1% for the variance in Trustworthiness, Commitment and Corporate Image respectively. Customer satisfaction recorded 90.6% for Service Charge. This implies that Customer satisfaction accounts for majority of the variation in the bank's Service Charge.

4. Discussion

This study has established that there is a link between service quality and customer retention at Universal banks' in Ghana. This study finds service quality impacts on customer satisfaction and customer retention at Universal banks' in Ghana. This result is consistent with finding of other scholars (Ndubisi, 2007 & Titko and Lace 2010). Usually, customer satisfaction is the important predictor of customer retention, but this study establishes service quality has great impact on customer retention simultaneously with customer satisfaction. Again, the empirical results show customer satisfaction has the mediating role between service quality and customer retention. It implies that quality has a direct impact on customer satisfaction and indirect impact on customer retention through satisfaction, which is at variant to other studies (Zineldin, 2006). Meanwhile, previous researches (Jamal and Nasr, 2003 and Parasuraman *et al.,* 1993) found that there is no important relationship between customer satisfaction and tangible aspects of service quality, in contrast, this study noted that tangible significantly influence customer satisfaction. The study here asserted that satisfaction is strongly influenced by service quality dimension on responsiveness which contradicts a previous study Banergee (2012). There are overwhelming arguments that it is more expensive to win new customers than to keep existing ones (Hormozi and Giles, 2004).

In conclusion, the ordinal logit model fit the data set adequately and provides a better model fit indices compared to the SEM, where some of its model fit indices are off the threshold. Also, the parameter estimates and fit indices in the case of the SEM appear to have exaggerated compared to the ordinal logit model, which could be attributed to the sample size. Again, the SEM could utilise a mediation variable but the ordinal logit could not incorporate such variable. Meanwhile, the results from the SEM showed that loyalty influence to a great deal the trustworthiness and corporate image of universal banks in Ghana. Also, service quality is a relevant contributing factor to empathy, reliability and assurance of universal banks' services. However, reliability and assurance, in terms of, service quality aspects are not really contributing significantly to customer satisfaction. Service quality have both direct and indirect effect on customer retention. However, satisfaction was deemed statistically insignificant in contributing to customer retention. However, the results of the ordinal logit model showed that service quality dimensions in terms tangible, responsiveness, empathy and trust are relevant contributing factors to customer retention at Ghana commercial bank.

References

1. Agresti, A. (2002) Categorical Data Analysis, Second Edition. Hoboken, New Jersey: John Wiley & Sons, Inc.
2. Parasuraman, A. (2004) Assessing and Improving Service Performance for Maximum Impact: Insight from a two-decade-long research. Journal of performance measurement and metrics, 5(2), 45-52.
3. Titko, J and Lace, N. (2010) Customer satisfaction and loyalty in Latvian retail banking. Journal of economics and management, 15, 1031-1038.
4. Khattak, N. A. and Rehman, K. U. (2010), Customer satisfaction and awareness of islam banking system in Pakistan, African Journal of Business Management, 4(5), 662-671.
5. Prabhakaran, S., and Satya, S. (2003). An insight into Service Attributes in Banking Sector. *Journal of Services Research*, 3(1), 157-169.
6. Kumar, M., Kee, F. T., and Manshor, A. T. (2009). Determining the relative importance of critical factors in delivering service quality of banks: an application of dominance analysis in SERVQUAL model. *Managing Service Quality*, 19(2), 211-228.
7. Chang, H. H., Wang, Y. A., and Yang, W. Y. (2009). The impact of e-service quality, customer satisfaction and loyalty on e-marketing: Moderating effect of perceived value. *Total Quality Management and Business Excellence*, 20(4), 423.
8. Zeithaml, V. A., Wilson, A., and Bitner, M. J. (2008). *Services Marketing.* 4th ed. New Delhi: The McGraw-Hill Companies.
9. Bedi, M. (2010). An integrated framework for service quality, customer satisfaction and behavioural responses in Indian Banking industry: a

comparison of public and private sector banks. *Journal of Services Research,* 10(1), 157-172.

10. Kassim, N., and Abdullah, N. A. (2010). The effect of perceived service quality dimensions on customer satisfaction, trust, and loyalty in e-commerce settings: a cross cultural analysis. *Asia Pacific Journal of Marketing and Logistics*, 22(3), 351-371.
11. Kumar, S. A., Mani, B. T., Mahalingam, S., and Vanjikovan, M. (2010). Influence of Service Quality on Attitudinal Loyalty in Private Retail Banking: an empirical study. *IUP Journal of Management Research*, 9(4), 21-38.
12. Naeem, H., and Saif, I. (2009). Service Quality and its impact on Customer Satisfaction: An empirical evidence from the Pakistani banking sector. *The International Business and Economics Research Journal,* 8(12), 99.
13. Balaji, M. (2009). Customer Satisfaction with Indian Mobile Services. *IUP Journal of Management Research,* 8(10), 52-62.
14. Lee, M. C., and Hwan, I. S. (2005). Relationships among service quality, customer satisfaction and profitability in the Taiwanese banking industry. *International Journal of Management*, 22(4), 635-648.
15. Athanassopoulos, A., and Iliakopoulos, A. (2003). Modeling customer satisfaction in telecommunications: assessing the effects of multiple transaction points on the perceived overall performance of the provider. *Production and Operation Management,* 12(2), 224-245.
16. Yee, R. Yeung, A., and Cheng, T. (2010). An empirical study of employee loyalty, service quality and firm performance in the service industry. *International Journal of Production Economics,* 124(1), 109.
17. Banerjee, N. (2012). A Comparative Study of Customers’ Perceptions of Service Quality Dimensions between Public and Private Banks in India. *International Journal of Business Administration.* 3 (5): 34.
18. Abdullah, D.N.M.A. and Rozario, F. (2009), Influence of service and product quanlity towards customer satisfaction: A case study at the staff cafeteria in the Hotel industry, World Academy of Science, Engineering and Technology, 53, 185-190.
19. Hormozi AM, Giles S (2004). Data Mining: A Competitive Weapon for Banking and Retail Industries. Information System Management.
20. Anderson H, & Jacobsen PQ (2000). Creating Loyalty: Its strategic importance in your customer strategy. S.A. pp.55-67.
21. Ndubisi NO (2007). Relationship marketing and customer retention. Mark. Intell. 25(1):98106.
22. Zineldin M (2006). "The royalty of loyalty: CRM, quality and retention". J. Consum. Mark. 23(7):430-437.

23. Kim W.G., and Cha Y (2002). Antecedents and consequences of relationship qualityin hotel industry. Hosp. Manag. 21:321-338.

Single and two-population mortality rate modeling for selected CEE countries

Justyna Majewska, Grażyna Trzpiot
Department of Demography and Economic Statistics, University of Economics in Katowice,Poland, 40-881 Katowice

Abstract

Multi-population models for modeling and forecasting mortality rates have been the major focus of many authors since the seminal work by Lee and Li (2005). Models are typically based on the assumption that the forecasted mortality experiences of two or more related populations converge in the long run. We compare two-population mortality model by Li and Lee (2005) (for pairs of populations) and Lee-Carter model (1992) for each population independently. The aim of the paper is to derive evidence for similarity of some populations from Central and Eastern Europe in order to model and forecast mortality rates using multi-population models.

Keywords

mortality; multi-population; common stochastic trends

1. Introduction

Recent decades have seen significant improvements in mortality in most developed countries. A variety of different models have been considered in the literature and practice. For many years, the Lee-Carter mortality projection methodology (Lee and Carter 1992) has been the benchmark extrapolative mortality forecasting method (Booth and Tickle 2008; Shang et al. 2011; Stoeldraijer et al. 2013). Lee-Carter model has been supplemented by a variety of alternatives that might be considered improvements on the single-factor LC model according to a variety of criteria (e.g. Brouhns et al. 2002; Currie et al. 2004; Renshaw and Haberman 2006; Cairns et al. 2006; Hyndman and Ullah 2007).

A number of authors have considered multi-country and methods comparisons. Janssen (2018) provides a shortly overview of important recent advances in mortality forecasting and the current available advanced mortality forecasting approaches. Booth and Tickle (2008) reviews the main methodological developments in mortality modelling and forecasting since 1980 under three broad approaches: expectation, extrapolation and explanation. Cairns et al. (2009) compare quantitatively different stochastic models explaining improvements in mortality rates in selected countries. Macdonald et al. (1998), Tuljapurkar et al. (2000) and Booth et al. (2006) make some qualitative comparisons of various countries using single-population models. Numerous studies do not provide a conclusion which model – in general – is the best in order to project mortality rates. The best model for one country does not mean that this model will be the best for the other.

Most work has focused on stochastic mortality models for single populations, however, it is important to be able to model two or more populations simultaneously. As Li and Lee (2005) indicated the convergence in mortality levels for closely related populations can lead to unsuitable mortality projections, if the projections for individual populations are obtained in isolation from one another. Similar historical trends in long-run life expectancy patterns can be useful for countries. Besides, analyses of the main determinants of life expectancy (the socio-economic, environmental or behavioural factors) of associated populations are crucial. Knowledge of existence of some common stochastic trends in mortality rate in cluster of European countries can be used for projections mortality rates and life expectancy (Majewska 2017; Lazar et al. 2016).

2. Methodology

Many models have been proposed in the literature to represent the mortality evolution of two or more related populations. The majority of such models extend known single population models by specifying the correlation and interaction between the involved populations (Villegas et al., 2017). The Lee-Carter model (1992) was originally developed for a single country, and is defined as follows:

$$\log m(x,t) = a(x) + b(x)k(t) + e(x,t)$$

The country specific *$\alpha(x)$* determines the baseline shape of the mortality curve in a country, *$\beta(x)$* (age-specific component) tells us which rates decline rapidly and which rates decline slowly in response to changes in *$\kappa(t)$* (time-varying mortality index). *$\varepsilon(x,t)$* is the error term of Lee-Carter model with mean zero and variance σ_{δ}. Mortality index *$\kappa(t)$* is used to forecast the series. Since parameters *$\beta(x)$* and *$\kappa(t)$* are unobserved variables, the least square estimates can be found by using the Singular Value Decomposition method.

We adopt an extension of the Lee-Carter method suggested by Li and Lee (2005), the so-called augumented common factor model. Model is defined as follows:

$$\log m(x,t,i) = a(x,i) + b(x)k(t,i) + e(x,t,i)$$

Li and Lee (2005) proposed that the parameters should be estimated using two-step singular value decomposition, firstly estimating the common parameters from the combined data for all countries, and secondly estimating the rest of the country specific parameters. Model takes into account the fact that mortality patterns for closely related populations are expected to be similar.

where *C1* and *C2* donotes populations from two different countries. This ensures that the rates of change of the future mortality rates are the same for the two populations, and thus avoids crossovers.

The crucial issue was to derive similar countries to Poland. Thus, the analysis was preceded by an idenfitifaction of homogenous spatial clusters of EU countries according to the following demographic and economic variables (details can be found in Majewska and Trzpiot, 2019):

- Human Development Index – developed by the United Nations to measure and rank countries' levels of social and economic *development,*
- Air pollution – greenhouse gas emissions in tons per capita,
- Social protection expenditures measured as percentage of GDP,
- Doctors providing direct care to patients per 1000 inhabitants,
- Alcohol – annual sales of pure alcohol in liters per person aged 15 years and older,
- Cigarettes – a percentage of daily smokers of the population aged 15 years and over,
- Obesity – a percentage of obese inhabitants in population; obesity is measured by the body mass index.

Cluster with Poland contains also the following countries: Czech Republic, Malta, Latvia, Lithuania, Slovakia, Croatia, Hungary, Romania, Bulgaria, Greece, Cyprus. For comparison Slovakia, Czech Republic, Lithuania and Hungary are selected.

The dataset comprises the number of deaths and the number of exposures for male and female in above-mentioned countries since the beginning of 1950 until 2015.

3. Results

Trends in mortality for the countries grouped in a spatial cluster are presented in figure 1. A visual inspection of figure 1 suggests the possibility of common stochastic trends in mortality.

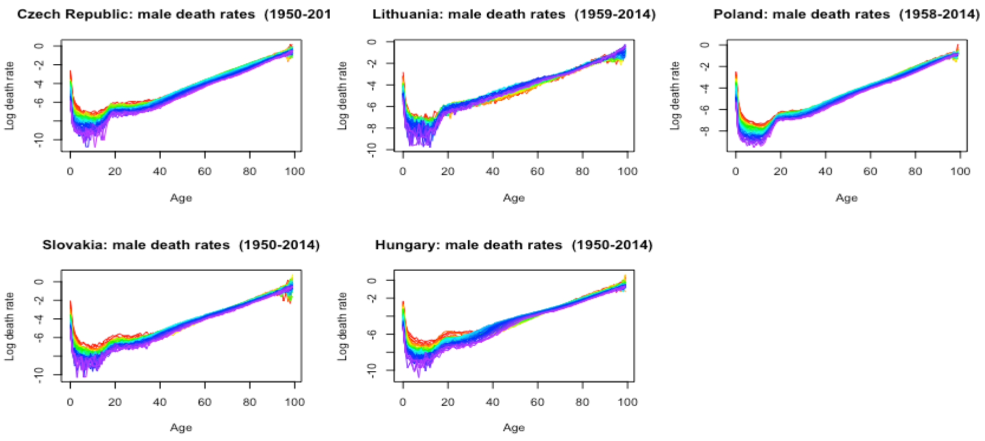
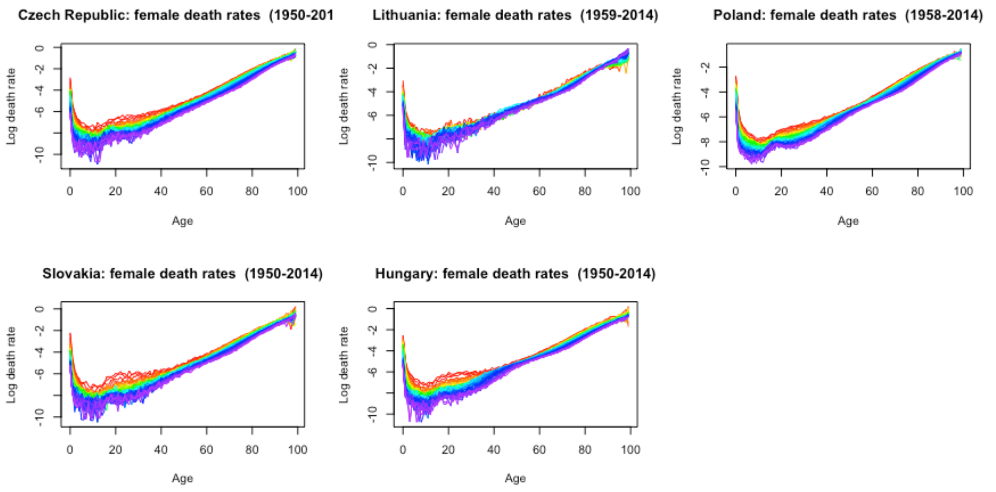


Fig. 1. Log mortality rates for male in selected European countries



In a two-population model, we expect certain behaviour from its parameters. In a perfect situation the common parameters should be able to capture the true global mortality trend, in both age and time, amongst the populations as a whole (Enchev et al., 2015). If the underlying philosophy of the model is correct, then we would expect that the country specific period effects all fluctuate around some constant level in the long term (Enchev et al., 2015). Significant differences from this level, in either range or shape would mean that this particular population is somehow different from the other populations.

Presented results are limited to the age-specific component, because parameters *α(x)* for each pair of countries exhibit the same values and manifest the expected general shape of mortality schedule across ages. Besides, almost a similar downward trend is observed in the *κ(t)* values. Figures 3-6 present values of estimated parameters *β(x)* for the independent LC model (each country separetly) and for augmented common factor model. Parameter *β(x)* reflects the relative change of mortality rate.

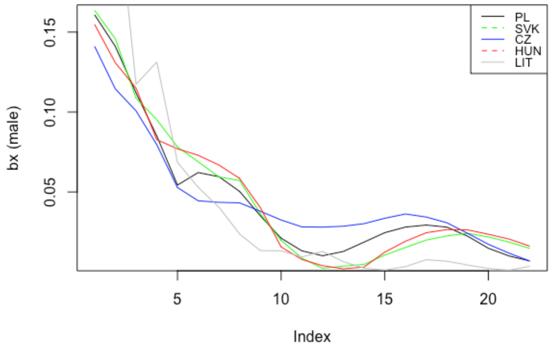


Fig. 3. Age-specific parameter bx for individual LC model fitted to pairs of male populations for ages 0-100 (5-year age groups) and the period 1950-2015

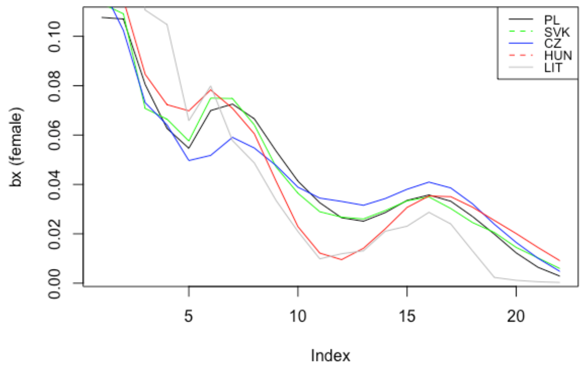


Fig. 4. Age-specific parameter bx for individual LC model fitted to pairs of female populations for ages 0-100 (5-year age groups) and the period 1950-2015

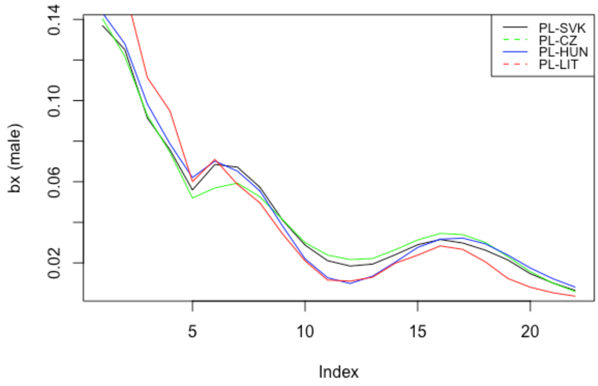


Fig. 5. Coherent parameter bx for the augumented common factor model fitted to pairs of male population for ages 0-100 (5-year age groups) and the period 1950-2015

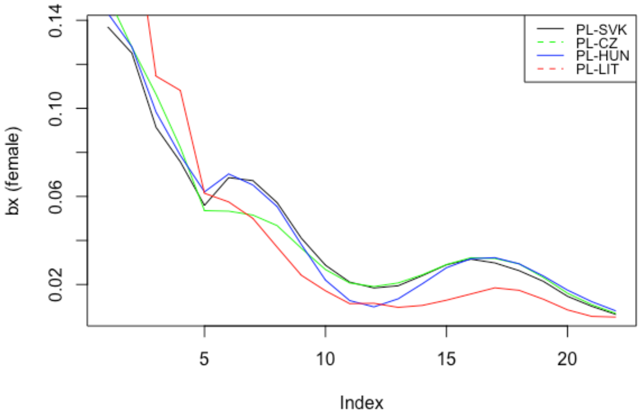


Fig. 6. Coherent parameter bx for the augumented common factor model fitted to pairs of female population for ages 0-100 (5-year age groups) and the period 1950-2015

Table 1 presents several methods to compare the models' goodness-of-fit for in-sample fitting by using explanation ratio (ER), mean absolute error (MAE) and mean absolute percentage error (MAPE). Li and Hardy (2011) and Li and

Lee (2005) noted that a model with the highest value of ER signifies the best fit to the data. The lowest MAE and MAPE indicate for a better fit to historical data as well.

Table 1. In-sample goodness-of-fit measures

	LC for each country separately	Two-population models
Explanation ratio	0.8538	0.9251
Mean absolute error	0.0061	0.0049
Mean absolute percentage error	1.361	1.054

4. Discussion and Conclusion

We compared the two-population mortality model by Li and Lee (2005) and Lee-Carter model for ech country independently. We notice that the parameters for two-population model for each pair of populations look similar. There are some differences between the age-specific component in model Poland-Lithuania and rest of models. As expected, all of the common parameters behave similarly, which is an indication that the models capture the common trend between Poland and other countries.

The historical period for in-sample fitting is ranged from the year 1950 until the year 2015. The results in Table 1 suggest that the augmented common factor model shows the best in-sample error performances as compared to the independent model.

References

1. Booth, H., Hyndman, R.J., Tickle, L. & De Jong, P. (2006) Lee-Carter mortality forecasting: A multi-country comparison of variants and extensions. *Demographic Research*, 15: 289-310.
2. Brouhns, N., Denuit, M., & Vermunt, J. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. Insurance: Mathematics and Economics, 31(3), 373–393.
3. Cairns, A. J. G., Blake, D., & Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. Journal of Risk and Insurance, 73(4), 687–718.
4. Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., & Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. North American Actuarial Journal, 13(1), 1–35.
5. Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., & Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. North American Actuarial Journal, 13(1), 1–35.
6. Currie, I. D., Durban, M., & Eilers, P. H. C. (2004). Smoothing and forecasting mortality rates. Statistical Modelling, 4(4), 279–298.

7. Hyndman, R. J., & Ullah, M. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. Computational Statistics & Data Analysis, 51(10), 4942–4956.
8. Janssen F. (2018). Advances in mortality forecasting: introduction. Genus Journal of Population Sciences, **74**:21.
9. Lazar, D., Buiga, A. & Deaconu A. (2016). Common stochastic trends in European mortality levels: testing and consequences for modeling longevity risk in insurance. Romanian Journal for Economic Forecasting 2, 152–168
10. Lee, R. D., & Carter, L. (1992). Modeling and forecasting the time series of US mortality. *Journal of the American Statistical Association, 87*, 659–671.
11. Li, N., & Lee, R. D. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography, 42*, 575–594.
12. Macdonald, A.S., Cairns, A.J.G., Gwilt, P.L. & Miller, K.A., (1998) An international comparison of recent trends in population mortality. *British Actuarial Journal* 4: 3-141.
13. Majewska J. (2017). An EU cross-country comparison study of life expectancy projection models, "Selected papers from the 2016 Conference of European Statistics Stakeholders. Special issue", Luxembourg: Publications Office of the European Union, 83-93.
14. Renshaw, A. E., & Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. Insurance: Mathematics and Economics, 38(3), 556–570.
15. Tuljapurkar, S., Li, N. & Boe, C. (2000). A universal pattern of mortality change in G7 countries. *Nature* 405: 789-792.
16. Villegas, A. M. & Haberman, S. & Kaishev, V. K. & Millossovich, P. (2017). A Comparative Study Of Two-Population Models For The Assessment Of Basis Risk In Longevity Hedges, ASTIN Bulletin: The Journal of the International Actuarial Association, Cambridge University Press, vol. 47(03), 631-679, September.

Distribution of working population and economic activities in the northern region, Malaysia

Nur Azmina Ahmad Zuhkhori, Zainuddin Ahmad, Suzana Abu Bakar
Department of Statistics Malaysia

Abstract

The distribution of working population and economic activities in a province or state are usually influenced by its levels of development. This study focuses on the distribution of working population in the states of the northern region of Malaysia. The technique used for the study is the location quotient (LQ) analysis on 2010 to 2013 data of the working population based on the economic activities in the agriculture, mining and quarrying, manufacturing, construction and services sectors. The analysis involved data of the four states in the northern region namely Kedah, Perlis, Pulau Pinang and Perak. Based on LQ analysis, the employment convergence in Kedah, Perlis and Perak were high in the agriculture sector, while in Pulau Pinang, the employment convergence was high in the manufacturing sector.

Keywords

Location quotient; Northern Region; Working Population

1. Introduction

Economic analysis in Malaysia showed changes from agriculture and mining sectors to manufacturing and services sectors. In 1987, agriculture and mining sectors contributed 32.6% to Malaysia GDP as compared to services sector, 45.3%. In 2014, the agriculture and mining sector contribution dropped to 18.2% while manufacturing and services sectors increased to 23.0% and 53.5% respectively (Department of Statistics Malaysia, 2014).

Pulau Pinang recorded the highest GDP among other states in the Northern Region, 7.0% in 2013. It was contributed by the manufacturing sector, 47.8% and services sector, 47.0%. Perak showed the second highest contributor to GDP, 5.3% with the biggest contribution was the services sector, 63.2%. Kedah with 3.4% contribution to GDP showed that services and manufacturing sectors recorded 54.9% and 30.4% respectively. Perlis posted the lowest contribution to GDP in the Northern Region with a share of 0.5% and the main contributor to Perlis GDP was the services sector (Department of Statistics Malaysia, 2014).

Industrial sector growth was estimated to be 12% a year, although in 1994-1995, this sector grew at 16.3%. Meanwhile the agriculture sector was estimated to expand at 2.5% annually due to the domination of paddy and rubber, while, services sector was estimated to grow at 9.4% a year (Hashim, 1996).

However, despite of the increased in the GDP per capita and the life quality of the population, there were unbalanced development among the states in Malaysia (Hasnah, Noraziah and Sanep Ahmad, 2011) and the main focused in this study was for the states in the Northern Region of Malaysia that were Kedah, Perlis, Pulau Pinang and Perak.

In spite of that, this study was taken to see the concentration of labour force by economic activities in the agriculture, mining and quarrying, manufacturing, construction and services sectors in 2010 to 2013 in the Northern Region of Malaysia.

2. Methodology

Data used in this study was the distributive data on working population by the economic activities in 2010 to 2013. The source of this study was based on the publication which was published by Department of Statistics Malaysia, Labour Force Survey Report.

The technique used for this study is the location quotient (LQ) analysis. LQ is an index to compare the sharing of activity of an area with other regions in aggregate. It can show whether the distribution of a particular activity is concentrated in an area or it is distributed equally. In general, LQ is the ratio of the region's labour force to the ratio of labour force to a particular industry or region (Anuar Ali, 1983 and Bendavid, 1974).

$$LQ = \frac{(ai/bi)}{(Ai/Bi)}$$

ai = Labour Force in sector i in state S
bi = Labour Force in state S
Ai = Labour Force in sector i in country n
Bi = Labour Force in economy n

3. Results

This analysis was conducted to compare the distribution between the states in the Northern Region, the entire Northern Region itself and the comparative between the studies of 2003 to 2006 with 2010 to 2013 studies. The analysis of comparisons of all the states in the Northern Region is as Table 1.

For Kedah, the agriculture and manufacturing sectors recorded LQ index more than 1 point which indicated high employment concentration in both sectors as compared to the same sectors in Malaysia. LQ for the agriculture sector recorded an increase from 1.17 point in 2010 to 1.59 point in 2013. The index recorded the highest figure in 2012 with 1.63 point. The mining and quarrying sectors recorded the lowest index but with an upward trend.

Meanwhile, construction and services sectors recorded LQ index nearly to 1 but the trend was declining.

Perlis showed that the employment concentrations were in the agriculture, construction and services sectors which registered the LQ index above 1 point. The agriculture sector recorded the highest LQ index in 2011 but showed a downward trend by recording the LQ index 1.12 point in 2013. Construction sector showed an upward trend from 2011 (LQ index 1.01 point) to 1.08 point in 2013. However, in 2010, it recorded the highest LQ index of 1.17 point. LQ index of manufacturing sector registered 1.06 point in 2013 as compared to 1.09 point in 2012. The manufacturing sector and the mining and quarrying sector consistently recorded a relatively low LQ index during the review period.

Table 1: Location Quotient Index by States in the Northern Region with Malaysia, 2010 to 2013

State	Economic Activities	2010	2011	2012	2013
Kedah	Agriculture	1.17	1.37	1.63	1.59
	Mining and Quarrying	0.20	0.20	0.48	0.49
	Manufacturing	1.23	1.26	1.15	1.26
	Construction	0.92	0.71	0.76	0.78
	Services	0.91	0.90	0.87	0.84
Perlis	Agriculture	1.14	1.38	1.21	1.12
	Mining and Quarrying	0.51	0.39	0.54	0.64
	Manufacturing	0.48	0.62	0.56	0.66
	Construction	1.17	1.01	1.02	1.08
	Services	1.10	1.04	1.09	1.06
Pulau Pinang	Agriculture	0.14	0.18	0.24	0.23
	Mining and Quarrying	0.06	-	0.10	0.27
	Manufacturing	1.89	1.87	1.78	1.84
	Construction	0.73	0.69	0.82	0.80
	Services	0.98	0.95	0.97	0.97
Perak	Agriculture	1.08	1.12	1.21	1.25
	Mining and Quarrying	0.64	0.88	0.63	0.89
	Manufacturing	1.05	1.06	0.99	1.06
	Construction	0.97	0.87	0.82	0.82
	Services	0.97	0.98	0.99	0.96

Note. "-" no labour force at that sector in related states

The highest employment concentration in Pulau Pinang was in the manufacturing sector. The LQ index consistently recorded not less than 1.78

x`x`point in the review period. The highest was recorded at 1.89 point in 2010. The second highest to employment concentration was in the services sector that recorded the LQ index approaching to 1 point. The least concentrated sector in Pulau Pinang were agriculture and mining and quarrying which registered LQ index 0.23 and 0.27 points respectively in 2013.

In Perak, the employment concentration is fairly balanced. The LQ index range in 2013 for construction sector registered 0.82 point while the agriculture sector recorded 1.25 point. The agriculture sector consistently recorded the LQ index above 1 point in the review period and showed an upward trend. Manufacturing sector recorded the LQ index above 1 point except in 2012, with 0.99 point. In 2013, construction and mining and quarrying sectors recorded 0.82 and 0.89 points respectively.

This study also analysed comparison by state and economic sectors as shown in Table 2. Based on the table, all states in the Northern Region recorded LQ index exceeding 1 point for agricultural sector except Pulau Pinang. This proves that employment was focused in the agriculture sector for the Northern Region. Kedah recorded the highest LQ index for 2013 at 1.59 point, followed by Perak (1.25 point) and Perlis (1.12 point). Employment concentration for the agricultural sector in Pulau Pinang was at 0.23 point.

Mining and quarrying sectors showed that all states registered LQ index less than 1 point. In 2013, Perak recorded the highest LQ index (0.89 point). During the review period, the LQ index in Perak showed a range between 0.63 point and 0.89 point. Perlis recorded relatively consistent LQ (0.39 to 0.64 points) while Pulau Pinang recorded the lowest LQ of between 0.06 point and 0.27 point. For Kedah, although the LQ index in 2010 recorded 0.20 point, it showed an upward trend.

Table 2: Location Quotient Index by Economic Activities and State, 2010 to 2013

Economic Activities	State	2010	2011	2012	2013
Agriculture	Kedah	1.17	1.37	1.63	1.59
	Perlis	1.14	1.38	1.21	1.12
	Pulau Pinang	0.14	0.18	0.24	0.23
	Perak	1.08	1.12	1.21	1.25
Mining and Quarrying	Kedah	0.20	0.20	0.48	0.49
	Perlis	0.51	0.39	0.54	0.64
	Pulau Pinang	0.06	-	0.10	0.27
	Perak	0.64	0.88	0.63	0.89
Manufacturing	Kedah	1.23	1.26	1.15	1.26

	Perlis	0.48	0.62	0.56	0.66
	Pulau Pinang	1.89	1.87	1.78	1.84
	Perak	1.05	1.06	0.99	1.06
Construction	Kedah	0.92	0.71	0.76	0.78
	Perlis	1.17	1.01	1.02	1.08
	Pulau Pinang	0.73	0.69	0.82	0.80
	Perak	0.97	0.87	0.82	0.82
Services	Kedah	0.91	0.90	0.87	0.84
	Perlis	1.10	1.04	1.09	1.06
	Pulau Pinang	0.98	0.95	0.97	0.97
	Perak	0.97	0.98	0.99	0.96

Note. "-"no labour force at that sector in related states

In the construction sector, Perlis recorded the highest employment convergence with LQ index exceeding 1 point which was 1.08 point in 2013. It also consistently recorded the LQ index above 1 point in the review period. Perak had the second highest LQ index of 0.82 point followed by Pulau Pinang, 0.80 point. Overall, this sector's employment focus in the Northern Region was relatively well-balanced, ranging from 0.69 point to 1.17 point during the review period. In 2013, the range was between 0.78 point and 1.08 point.

As for the services sector, the focus of employment between the states in the Northern Region was well balanced. The lowest index recorded was 0.84 point in Kedah in 2013 and the highest index was 1.10 point recorded in Perlis (2010). Perlis also recorded LQ index more than 1 point in the review period. Pulau Pinang recorded a range between 0.95 point and 0.98 point while Perak recorded a range between 0.96 point and 0.99 point. Kedah showed a declining trend from 0.91 point in 2010 to 0.84 point in 2013.

Table 3: Location Quotient by Economic Activities in the Northern Region, 2003 to 2006

State	Economic Activities	2003	2004	2005	2006
Northern Region	Agriculture	0.73	0.66	0.83	0.86
	Mining and Quarrying	0.82	0.43	0.67	0.78
	Manufacturing	1.32	1.08	1.31	1.29
	Construction	0.86	2.83	0.88	0.83
	Services	0.97	0.77	0.96	0.96

Source: Study by Hasnah, Noraziah and Sanep Ahmad (2011)

Table 3 and 4 showed the LQ index for the Northern Region for 2003 to 2006, and 2010 to 2013. During 2003 to 2006, the LQ index for manufacturing exceeded 1 point. This illustrated the high employment convergence in the manufacturing sector as compared to Malaysia's manufacturing sector in that period. For the period of 2010 to 2013, the same trend was shown for the manufacturing sector in the Northern Region. The manufacturing sector consistently recorded the LQ index above 1 point. For the agricultural sector, high employment concentrations were recorded in 2012, with the LQ index of 1.04 point. The mining and quarrying sector showed less employment concentration as compared to other sectors in the Northern Region. The LQ index for the sector was the lowest at 0.33 point in 2010 and 0.57 point in 2013

Table 4: Location Quotient by Economic Activities in the Northern Region, 2010 to 2013

State	Economic Activities	2010	2011	2012	2013
Northern Region	Agriculture	0.82	0.92	1.04	1.05
	Mining and Quarrying	0.33	0.39	0.42	0.57
	Manufacturing	1.34	1.35	1.26	1.34
	Construction	0.88	0.77	0.80	0.81
	Services	0.96	0.95	0.95	0.93

4. Discussion and Conclusion

The study shows that the highest employment convergence in the Northern Region was in the manufacturing sector. The sector which recorded less employment was the mining and quarrying sector. Agriculture sector showed a growing trend starting from 2012. For state level, the highest employment concentration in Pulau Pinang was in the manufacturing sector, while in Kedah, agriculture sector recorded the highest LQ index of 1.59 point in 2013. Perak was also relatively balanced, and the highest LQ index was in the manufacturing sector. Meanwhile for Perlis, employment concentrations were in agriculture, construction and services sectors which registered LQ index above 1 point.

The result of the study showed that the manufacturing sector had a high concentration of employment in the Northern Region of Malaysia. However, this study has to be done for all the states in Malaysia so that comparison of employment convergence can be seen for the whole of Malaysia. Further studies on these topics should also be undertaken to see the impact on employment convergence for the economy of a state.

References

1. Asan Ali Golam Hassan. (1998). Ketidakseimbangan agihan industri dan migrasi penduduk di Negeri Kedah. Universiti Utara Malaysia.
2. Hasnah Ali, Noraziah Ali, & Sanep Ahmad, (2011). Ketidakseimbangan wilayah dan sektor berpotensi mentransformasikan sosioekonomi penduduk menggunakan pendekatan location quotient. Geografia: Malaysian Journal of Society and Space, 7 (5 (spe). pp. 190-201. ISSN 21802491.
3. Bendavid, A. (1974). Regional Economic Analysis for Practitioners: An Introduction to Common Descriptive Methods. Revised ed (1st ed.). New York, etc.: Praeger.
4. Hashim Ismail. (1996). Pelan Tindakan Pembangunan Negeri Kedah, Kertas Kerja Seminar Ekonomi Peringkat Kebangsaan 96. Universiti Utara Malaysia.
5. Department of Statistics Malaysia. (2013). Labour Force Statistics, Malaysia.
6. Department of Statistics Malaysia. (2014). KDNK by State 2005-2013.

The evolution of foreign direct investment in Malaysia

Syamimi Shahbudin, Yusnita Mohd Yusof
Department of Statistics Malaysia

Abstract

The inflows of Foreign Direct Investments (FDI) brought many changes to structure of economy and spatial development in Malaysia. This development has been attributed by investment inflows into Malaysia since 1970s by foreign investors especially in financial transactions besides trading activities. The evolution of investment in Malaysia was portrayed through the economic transition from an agriculture-based economy into an industrialised economy largely driven by the FDI inflows into Malaysia which also plays a prominent role in the regional and global trade value chain. The success of inflows in industry brings positive effects to the host country in terms of new investments, economics growth and foreign exchange earnings. The inflows of FDI in Malaysia have increased from RM23.9 billion in 2008 to RM41.0 billion in 2017. Therefore, the purpose of this paper is to analysis Sector Industry of FDI in Malaysia and to examine the nature of FDI that was channelled into economy activities in Malaysia as well as comparative analysis of trends FDI in Malaysia, which help to determine strategies by industry that investment and stimulate economic growth in Malaysia.

Keywords

Investments; foreign direct investment; evolution investment

1. Introduction

Investment is one of the engines for performance of economy in the past few decades and it was increasing rapidly since 1990s. The inflows were particularly notable for developing countries for which Foreign Direct Investment (FDI) has become the most stable for capital flows for country. The contributions of these investments to economy provide positive impact on the balance of payments, creating new job, modern technologies, and etc. Most of the studies found a positive impact of FDI on economic growth in developed and developing countries. While inflows by economics sectors have different effects to economic growth. Thus, when understanding the impact of FDI, it is important to understand what attracts FDI, how this has changed over time and how these changes in determine economic growth.

Foreign Direct Investment

Foreign direct investment is a form of inflows from a foreign company into another country. It is the establishment of an enterprise by a foreigner. Foreign Direct Investment (FDI) in Malaysia is set up following the holding of at least 10% of the total equity in a resident company by a non-resident investor. Consequent transactions in financial assets and liabilities between resident companies and non-resident direct investors linked by a foreign direct investment relationship (FDIR) can also be known as FDI. The transactions could be between Malaysian companies and with its immediate or ultimate parent or fellow companies. More specifically, foreign direct investment is a cross-border corporate governance mechanism through which a company obtains productive assets in another country. Its definition can be extended to include investments made to acquire lasting interest in enterprises operating outside of the economy of the investor.

Trend of FDI

The level inflow of investment into Malaysia has changed throughout the year 2008 to 2017. Nevertheless, Malaysia's economic transition from agrarian activity in 1970s to manufacturing in 1990s and subsequently to services in 2000s was largely influenced by the economic policy implemented by the government. One of the most significant policies implemented by Malaysia was Look East Policy introduced in 1981. This reaffirmed the Malaysia's status as an open economy. Malaysia has become the major beneficiaries of trade openness and globalisation.

In the early 1970s, Malaysia was known as a major destination for FDI mostly in the manufacturing sector. One of the FDI strategy adopted by Government of Malaysia was to develop FDI location for specific industry. As a result, many cities in Malaysia have been transformed into a thriving Manufacturing hub. Over the years, foreign investments were more diversified into other industries precisely in Agriculture, Mining & quarrying, Construction and Services.

2. Literature Review

The economic effects of FDI are very difficult to measure accurately given that the performance depends on many factors, which FDI also affects several of these sectors. Findlay (1978) and Wang (1992) on their study find that FDI are importance in transferring technology which was relate to the foreign investment inflows if manufacturing or service sectors rather than to the primary sectors. Besides, Hirschman (1958) also emphasized that not all sectors have the same potential to absorb foreign technology or to create linkages with the rest of the economy. From that, he also highlights weak linkages between agriculture and mining. The weak linkages between sectors will have limited effect in spurring growth in an economy. Previous study by De Gregogio, and Lee (1998) and Carkovic and Levine (2002), find little support for FDI or positive effects in the primary sector, a positive effect of FDI in manufacturing on growth and ambiguous evidence from the service sector.

3. Methodology

The compilation of FDI is based on the guidelines recommended in Balance of Payments and International Investment Position Manual, Sixth Edition (BPM6) of the International Monetary Fund that provides a comprehensive methodology framework for collection and compilation of direct investments statistics. For countries to implement TSA, this manual aid providing related recommendations such as basic standard classifications, guidelines, definitions & concepts for compilation of FDI. While, economic sector is classified according to Malaysia Standard Industrial Classification 2008 ver. 1 related to classifications activity industries that fit in with FDI in Malaysia. The data for this compilation of FDI obtained from Quarterly Survey on IIP, jointly conducted between Department of Statistics Malaysia and Central Bank of Malaysia, which records transaction flows and position of Malaysia's external financial assets and liabilities with the rest of the world.

The purposed of this study is to measure the evolution sectors of industry in Malaysia on FDI performance. Descriptive analysis whish comparative analysis of trend is be using in this study. Due to limited data availability, the trend analysis period is from 2008 – 2017, while comparative analysis is between performance of FDI in year 2010 and 2017. The gap between the years was 7 years. Year of 2010 are chosen for this study due to stability of economy in Malaysia after global financial crisis.

4. Analysis and Findings

Performance of FDI in Malaysia, 2008-2017

In Malaysia, data on FDI in terms of flows and stocks are recorded in a statistics of Financial Account of the Balance of Payments (BOP), International Investment Position (IIP) and annual publication Statistics of Foreign Direct Investment in Malaysia. From the **Figure 1**, show the FDI in Malaysia has been on the upward trend since 2001 with an exception of 2009. The lower FDI flows in 2009 were largely due to global financial crisis. After the financial crisis in 2009, global economy has gradually recovered from the lower inflows of FDI to upward trend of inflows over 10 years. The FDI flows reached a new high in 2016 with a value of RM47.0 billion in 2016 and dropped to RM41.0 billion in 2017, partly reflecting the subdued global growth. The higher net inflow in 2016 was supported by higher net inflow in equity and investment fund shares. A large portion of these investments were directed into the Services sector, mainly for the acquisition of power assets by a foreign entity. As at end of

2017, Malaysia's FDI position stood RM570.3 billion and FDI flows on that year was largely contributed in equity & investment fund shares.

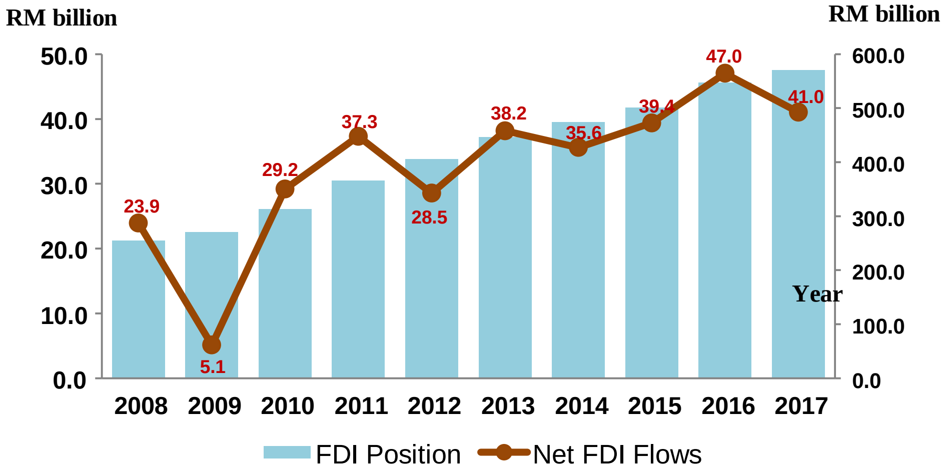


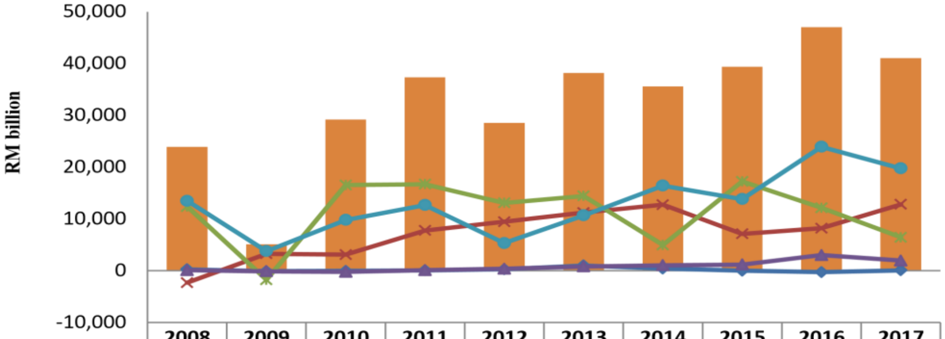
Figure 1: Performance of FDI in Malaysia, 2008- 2017

(Source: Department of Statistics Malaysia)

Performance of FDI in Malaysia by Economy Sectors, 2008-2017

The period 2008 – 2017 can be divided in terms of sectors economy of FDI in Agricultures, Mining & Quarrying, Manufacturing, Construction and Services. For the period 2008 to 2017, we can find that Services showed increasing trend over the year compared to Agriculture, Mining Construction and Manufacturing sector are showed slow increasing trend **(Figure 2)**.

The sectors with the highest FDI inflows were Services and Manufacturing which mean given for both sectors were RM13.0 billion and RM11.2 billion respectively. Services sector record the highest inflows of sector of economy in 2016 with value of RM23.9 billion and the lowest was from Mining sector in 2008 due to outflows to the parent company in abroad.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Total FDI	23,925	5,121	29,183	37,325	28,537	38,175	35,600	39,377	47,025	41,041
Agriculture	265	-88	-47	82	291	1,001	416	0	-268	87
Mining	-2,296	3,270	3,108	7,777	9,446	11,223	12,737	7,108	8,209	12,806
Manufacturing	12,368	-1,742	16,545	16,704	13,096	14,445	4,969	17,238	12,140	6,442
Construction	114	-103	-223	98	379	817	1,053	1,190	3,025	1,945
Services	13,473	3,786	9,800	12,663	5,325	10,688	16,425	13,841	23,919	19,762

Figure 2: Performance of FDI in Malaysia by Sectors, 2008- 2017

(Source: Department of Statistics Malaysia)

Performance of FDI by Sectors for year 2010 and 2017

From **figure 3,** we see that inflows in sector Agriculture, Mining and Construction shows increased in the period 2010 to 2017. While Manufacturing show decreased of inflows in FDI for year 2010 to 2017. In the sector of Agriculture and Construction, a higher inflow of RM 0.9billion and RM 1.9billion was recorded from an outflow of –RM 0.5billion and –RM 0.2billion in 2010 respectively. In the meantime, Mining and services are also recorded a higher inflow in year 2017 compared to year 2010. However, in the Manufacturing sector, a lower inflow was registered in 2017 compared to 2010. We see that among the sectors, activity in economy of FDI is mainly channelled to the services sectors. However, funds into the Manufacturing are become lower and decreased trend.

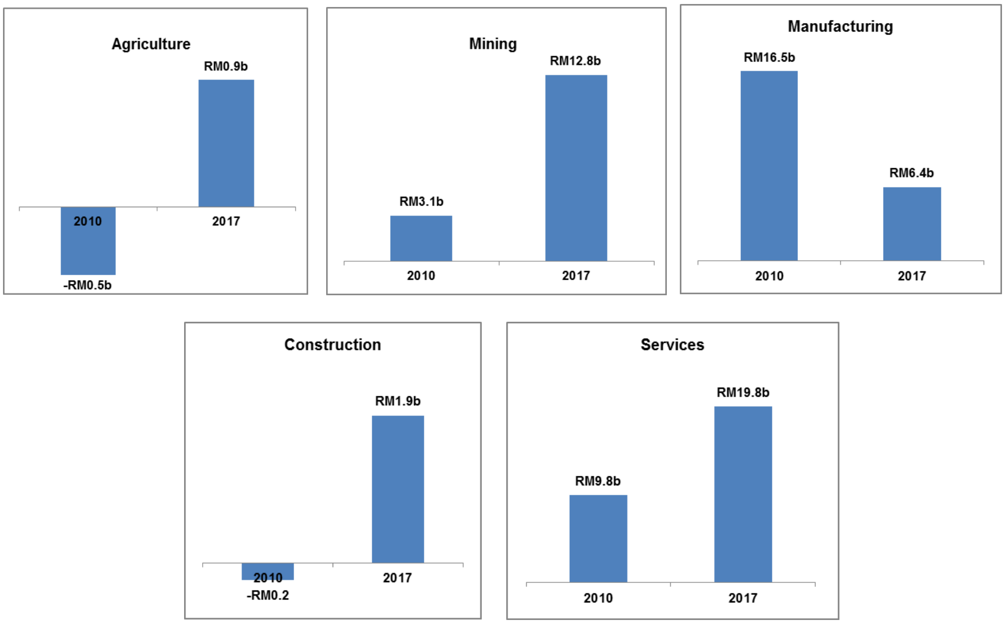


Figure 3: Performance of FDI by Sectors for year 2010 and 2017

(Source: Department of Statistics Malaysia)

5. Discussion and Conclusion

Targeting the economics sector for investment is determine by the higher inflows in sectors of industry, the dynamic reforming of inflows may improving business environment to attract a larger volume of FDI in the economy and also developing country. The analysis carried out to identify which sectors of industry in Malaysia do affect the FDI inflows. Based on the study, Services shows an improved in attracting a larger volume of FDI in economy. In the period of study 2008 to 2017, Services shows an increasing trend with higher inflows of FDI. However, Manufacturing shows a reverse trend compared to Services which a lower inflow received in Manufacturing sector. The lower contribution of the Manufacturing sector was partly due to the change in Government's focus to attract quality investment with high value added, high technology and strong linkages with domestic industries. With this strategy, Services sector have risen in term of new inflows in Malaysia.

Besides that, with the implementation policy such as Economic Transformation Plan (ETP), more investments have been channelled into higher value added activity with less labour-intensive.

References

1. Department of Statistics, Malaysia, *Statistics of Foreign Direct Investment in Malaysia (FDI) 2017* (2018), Putrajaya.
2. Marinela G, 2015. *Analysis of the Evolution of Foreign Direct Investment In The European Union, Amid The Global Economic Crisis.* University Of Bucharest, Romania.
3. Laura A, 2003. *Foreign Direct Investment and Growth: Does the Sector Matter?* Harvard Business School.
4. Oluchukwu A, 2013. *Foreign Direct Investment and Manufacturing Sector Growth In Nigeria*. Godfrey Okoye University.
5. Azeroual, 2016. *The Impact of Foreign Direct Investment on the Productivity Growth in the Moroccan Manufacturing Sector: Is Source of FDI important?* Rabat, Morocco.
6. Dirk W, 2006. *Foreign Direct Investment and Development: An Historical Perspective*. Overseas Development Institute.
7. Noor H, Paul C.W, Tim J.C, Euan F, 2003. *Foreign Direct Investment in Manufacturing Sector in Malaysia.* University Of New England.
8. J Bitzer, Holger G, 2005. *The Impact of FDI on Industry Performance*. University Berlin, University Of Nottingham.

A theoretical framework for analyzing instructors' beliefs, attitudes, and practices in the context of quantitative reasoning

Gerald Iacullo
Berkeley College

Abstract

College mathematics and statistics courses have been undergoing a transformation from a model focused on the acquisition of mathematical concepts, algorithms, formulas, and procedures to a model centered on students' ability to apply mathematics and statistics to a wide variety of authentic contexts, which is an essential characteristic of quantitative reasoning. To effectively achieve this change, instructors are being asked to rethink their instructional practices and move to reform-oriented instructional strategies that foster conceptual understanding and students' problem-solving skills. Research suggests that among the most significant predictors of instructional behaviour are the personal beliefs that instructors hold regarding the teaching and learning experience, including their own personal efficacy and the perceived usefulness of the instructional practices. This paper will explore four theories regarding the relationship between beliefs and behaviour as well as present an integrated model, especially with reference to the roles of personal teaching efficacy and perceived usefulness in determining instructional behaviour: the theory of reasoned action, the theory of planned behaviour, the technology acceptance model, and self-efficacy theory.

Keywords

Quantitative Reasoning; Instructional Behavior; Self-Efficacy; Perceived Usefulness

1. Introduction

College mathematics and statistics courses have been undergoing a transformation from a model focused solely on the acquisition of mathematical concepts, algorithms, formulas, and procedures to a model that emphasizes students' ability to apply mathematics to a wide variety of authentic contexts, which is an essential characteristic of quantitative reasoning and statistical literacy (Hughes-Hallett, 2003). To effectively achieve this change, mathematics and statistics instructors are being asked to rethink their instructional practices, which some have already done. Instead of strategies that concentrate on computational and procedural fluency in a traditional lecture, practice and drill format, there has been a movement toward reform-oriented instructional strategies that foster conceptual understanding and students' problem-solving skills. However, any attempt at a more comprehensive transformation of instructional practices depends on the beliefs of mathematics and statistics instructors.

2. Teachers'Belief

Among the most significant facilitators of and barriers to instructional behavior are the personal beliefs that instructors hold regarding the teaching and learning experience (Brownell & Pajares, 1999; Handal, 2003; Hassad, 2011, 2013). Such beliefs can be wide-ranging, encompassing the convictions, thoughts, and values that teachers may hold regarding the curriculum, teaching strategies, and styles of learning, as well as the beliefs students and teachers hold about students' abilities (Ernest, 1989; Handal, 2003). Moreover, teachers' beliefs, in particular self-efficacy and perceived usefulness, have been shown to be key determinants of instructors' decision to adopt innovative instructional practices (Hassad, 2011) and, in this regard, are more salient than an instructor's knowledge of the content or methods of instruction (Ernest, 1989). In this regard, four theories support this understanding of the relationship between beliefs and behavior, especially with reference to the roles of personal teaching efficacy and perceived usefulness in determining instructional behavior:

a) the theory of reasoned action (Ajzen & Fishbein, 1980),
b) the theory of planned behavior (Ajzen, 1991),
c) the technology acceptance model (Davis, 1986),
d) self-efficacy theory (Bandura, 1993, 1997).

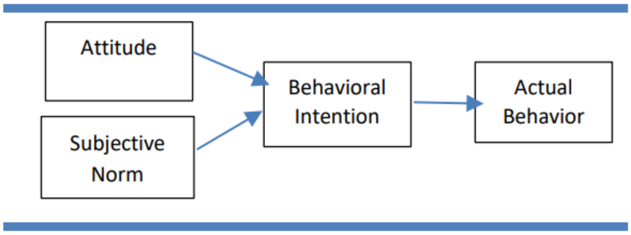
3. Theoretical Framework

Theory of Self-Efficacy. Personal teaching efficacy can be traced to Bandura's (1997) work on social cognitive theory, in which he identified self-efficacy as a prime motivational factor in determining human behavior and behavioral change. Generally, self-efficacy has been described as an individual's belief about his or her capability to successfully implement and accomplish a task in a particular situation or context (Bandura, 1993, 1997; Brownell & Pajares, 1999). This construct has been applied to many domains, including the field of teaching and learning. In this field, personal teaching efficacy is described as a "teacher's judgment of his or her capabilities to bring about desired outcomes of student engagement and learning" (Tschannen-Moran & Woolfolk-Hoy, 2001, p. 783).

The role of personal teaching efficacy as a determinant of instructor behavior has been well-established. Notably, an instructor who believes he or she has limited ability to affect the success of students will likely not put much effort into trying to improve student achievement. In contrast, instructors with high positive perceptions of their ability will tend to use a broad range of educational strategies to improve student performance (de la Torre Cruz & Arias, 2007; Swars, 2005; Tschannen-Moran & Woolfolk-Hoy, 2001). Moreover, high self-efficacy has been linked to instructors' willingness to overcome classroom problems through inventiveness and extra efforts (Narvaez, Khemelkov, Vaydich, & Turner, 2008, p. 5) as well as their openness to the use of innovative and reform-based instructional practices (Hassad, 2011, 2013; Pajares & Urdan, 2006).

Theory of Reasoned Action (Figure 1). With the theory of reasoned action, Ajzen and Fishbein (1980) attempted to describe the relationship between beliefs, attitudes, intentions, and behavior. The theory of reasoned action (TRA) addresses the idea that a person's performance of a particular behavior is determined by an individual's attitude toward performing the behavior (Ajzen & Fishbein, 1980). In addition, attitude toward performing the behavior is determined by an individual's beliefs regarding the consequences of performing the behavior (Pryor & Pryor, 2009). The path from beliefs and attitudes to behavior is mediated by behavioral intentions (Montano & Kasprzyk, 2002).

Figure 1. Theory of Reasoned Action

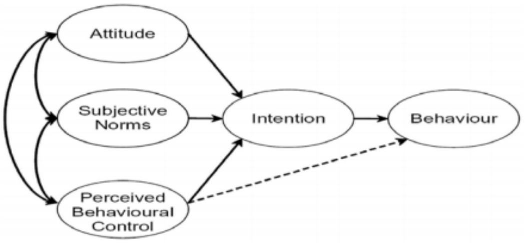


The theory of reasoned action (TRA) also posits that a person's attitudes toward a behavior is not the only predictor of behavioral intentions. Specifically, intent is caused by a person's attitudes and subjective norms, which have been described as the expectations of other people. As such, the influence that an individual's attitude has behavioral intention and actual behavior may be weakened by subjective norms (Ajzen, 1991). Since behaviors are not under complete volitional control, it is not always possible to predict behavior from intention (Armitage & Conner, 2001).

In terms of instructional practice, TRA has been used to suggest that an instructor's intention to adopt an instructional practice or approach will be consistent with his or her attitudes toward the practice. Moreover, these attitudes are rooted in the instructor's beliefs, in particular personal teaching efficacy and perceived usefulness. Nonetheless, an instructor may have a positive attitude toward the use of reformed-based instructional practices, but this may not be enough to predict the intention and (subsequently) the use of these instructional practices since relevant norms may suggest otherwise.

Theory of Planned Behavior (Figure 2). In the theory of planned behavior (TPB), Ajzen (1991) extended TRA by adding perceived behavioral control (PBC) as an additional determinant of behavior when behaviors are not under complete volitional control. That is, there may be both internal and external factors, like skills and resources, which may limit an individual's complete control over his or her behavior. In these situations, PBC, defined as "the ease or difficulty of performing the behavior of interest" (Ajzen, 1991, p. 183), affects the intention to perform the behavior and provides "useful information about the actual control a person can exercise...and can therefore be used as an additional direct predictor of behavior" (Ajzen, 2002, p. 184).

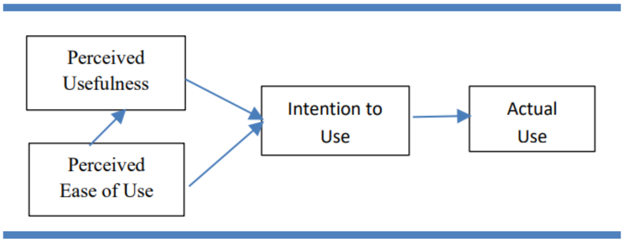
Figure 2. Theory of Planned Behavior



Self-efficacy, as articulated by Bandura (Ajzen, 1991; Montano & Kasprzyk, 2002), has been closely associated with perceived behavioral control. For instance, Fishbein and Cappella (2006) identified the two elements as the same construct, while Rodgers, Conner, and Murray (2008) noted a great deal of overlap in the way that self-efficacy and PBC were operationalized. Furthermore, Ajzen (2002) recognized self-efficacy (i.e., an individual's belief in his or her ability to perform a particular behavior), together with controllability (i.e., how much control individuals believe they have over a specific behavior), are both dimensions of perceived behavioral control (Ajzen, 2002). After an analysis of the research, Ajzen (2002) concluded that the overarching construct of perceived behavioral control and each of its two dimensions of controllability and self-efficacy are determinants of intention and behavior.

Technology acceptance model (Figure 3). This technology acceptance model (TAM) (Davis, 1986) was derived from the TRA in order to explain and predict the acceptance of information system technology. TAM was used to identify two main elements: (a) perceived usefulness, or the degree to which an individual believes that using a particular technology would enhance job performance (Davis, Bagozzi, & Warshaw, 1989); and (b) perceived ease of use, or the degree to which an individual believes that using a particular system would be free of effort (Davis et al., 1989). TAM relates to an individual's intention to perform the behavior and to actual usage. Subsequent research has demonstrated the importance of perceived usefulness on behavioral intention and behavior (Davis et al., 1989; Hassad, 2011; Roca, Chiu, & Martinez, 2006). As Hassad (2007) noted, perceived usefulness "has been theoretically and empirically established as…a strong and significant predictor of intention and behavior" (p. 98).

Figure 3. Technology Acceptance Model



Integrated model. This model explains the relationship among personal teaching efficacy, perceived usefulness and instructional behavior, consistent with the Theory of Reasoned Action (Ajzen & Fishbein, 1980) and the Theory of Planned Behavior (Ajzen, 1991). That is, an instructor's decision to perform a specific behavior (instructional practice) is based on the beliefs and attitudes the instructor has regarding the instructional practice. In particular, self-efficacy theory (Bandura, 1997) and the technology assistance model (Davis, 1986) identified personal teaching efficacy and perceived usefulness as proximal antecedents of both the intention to adopt and the adoption of these instructional practices.

4. Conclusion

Prior research has been largely consistent in demonstrating the relationship between self-efficacy and other determinants of performance like perceived usefulness (Hassad, 2011; Jung-Wan, & Mendlinger, 2011; Phan, 2013, Iacullo, 2015). In particular, Hassad (2011) found a moderately positive correlation between perceived usefulness and personal teaching efficacy in the use of reform-based instructional practices among instructors of introductory statistics courses for the behavioral sciences. Similar results were found among instructors of developmental math (Iacullo, 2016). Nonetheless, more research is needed on the theoretical framework described above as well as other theoretical approaches used to explain the relationship between instructors' beliefs and instructional behavior.

References

1. Ajzen, I. (1991). The theory of planned behaviour. *Organizational Behaviour and Human Decision Processes*, 50, 179–211.
2. Ajzen (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology*, 32, 665– 683.
3. Ajzen, I., & Fishbein. M. (1980). *Understanding attitudes and predicting social behaviour*. Englewood Cliffs, NJ: Prentice-Hall.
4. Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology*, 40(4), 471–499.
5. Bandura, A. (1997). Self-efficacy: The exercise of control. New York: W. H. Freeman.
6. Battista, M. T. (1999). The mathematical miseducation of America's youth. Phi Delta Kappan, 80(6), 424–433.
7. Brownell, M., & Pajares, F. (1999). Teacher efficacy and perceived success in mainstreaming students with learning and behavioral problems. *Teacher Education and Special Education,* 22(3), 154–164.
8. Davis, F. D. (1986). *A technology acceptance model for empirically testing new enduser information systems*: *Theory and results.* (Unpublished doctoral dissertation). Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.
9. Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models, *Management Science 35*, 982–1003.
10. de la Torre Cruz, M., & Arias, P. (2007). Comparative analysis of expectancies of efficacy in in-service and prospective teachers. *Teaching and Teacher Education*, 23(5), 641–652.
11. Handal, B. (2003). Teacher’s mathematical beliefs: a review. *The Mathematics Educator* 13(2), 47–57
12. Hassad, R. A. (2007*). Development and validation of a scale for measuring instructors' attitudes toward concept-based or reform-oriented teaching of introductory statistics in the health and behavioral sciences*. (Doctoral dissertation). Retrieved from Dissertations & Theses: A&I. (Publication No. AAT 3281778)
13. Hassad, R. A. (2011). Constructivist and behaviorist approaches: Development and initial evaluation of teaching practice scale for introductory statistics at the college level. *Numeracy, 4*(2): Article 7. DOI: http://dx.doi.org/10.5038/19364660.4.2.7
14. Hassad, R. A. (2013). Faculty attitude toward technology-assisted instruction for introductory statistics in the context of educational reform. Paper presented at the conference of the International

Association for Statistical Education, Cebu City, The Philippines. Retrieved from http://icots.net/roundtable/docs/Thursday/IASE2012_Hassad.pdf

15. Hughes-Hallett, D. (2003). The role of mathematics courses in the development of quantitative literacy. In B. L. Madison & L. A. Steen (Eds.), *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges* (p. 91−98). Princeton: National Council on Education and the Disciplines.
16. Iacullo, G. (2016) The evaluation of a pedagogical tool for quantitative literacy. In JSM Proceedings, Statistical Education Section. Alexandria, VA: American Statistical Association. 607 - 615.
17. Montano, D. E., & D. Kasprzyk. 2002. The Theory of Reasoned Action and the Theory of Planned Behavior. In K. Glanz, B. K. Rimer, and F. M. Lewis (Eds.). *Health \behavior and health education: Theory, research, and practice*. San Francisco: Jossey-Bass,
18. Narvaez, D., Khmelkov, V., Vaydich, J., & Turner, J. (2008). Teacher self-efficacy for moral education: Measuring teacher self-efficacy for moral education. *Journal of Research in Character Education,* 6(2), 3–15
19. Pajares, F., & Urdan, T. (2006). Self-efficacy beliefs of adolescents. Greenwich, CT: Information Age Publishing.
20. Phan, H. P. (2013). The capitalization of personal self-efficacy: Yields for practices and research development*. Journal of Educational and Developmental Psychology*, 3(1), 72.
21. Pryor, B. W., & Pryor, C. R. (2009). What will teachers do to involve parents in education? Using a theory of reasoned action. Journal of Educational Research & Policy Studies, 9(1), 45-59.
22. Roca, J. C., Chiu, C. M., & Martínez, F. J. (2006). Understanding e-learning continuance intention: An extension of the technology acceptance model. *International Journal of Human-Computer Studies, 64*(8), 683–696.
23. Swars, S. (2005). Examining perceptions of mathematics teaching effectiveness among elementary preservice teachers with differing levels of mathematics teacher efficacy. *Journal of Instructional Psychology,* 32(2), 139–146.
24. Tschannen-Moran, M., & Woolfolk-Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17, 783–805.

A study on the training effectiveness in school (cluster) of methodology, research and quality, ILSM

Norbani Yacob
Department of Statistics Malaysia

Abstract

Training and development is becoming an increasingly important function in human resource management. In statistical organization, staff's training is one of the main pillars of ensuring higher quality statistics product. To achieve this, ILSM is responsible to produce knowledgeable staff in the field of statistics. The purpose of this study is to determine the satisfaction level of participants in the training program organised by the School of Methodology, Research and Quality (MPK); to evaluate the effectiveness of the training programs as perceived by the participants and to evaluate the effectiveness of the training programs as perceived by the participant's supervisor. Two sets of questionnaires were used to gather all data related to the study. The results revealed that the participants were satisfy with the training programs organised by MPK. Also, overall, the trainings were effective as perceived by the participants and the participant's supervisor.

Keywords

Training programs; Training effectiveness; Satisfaction level

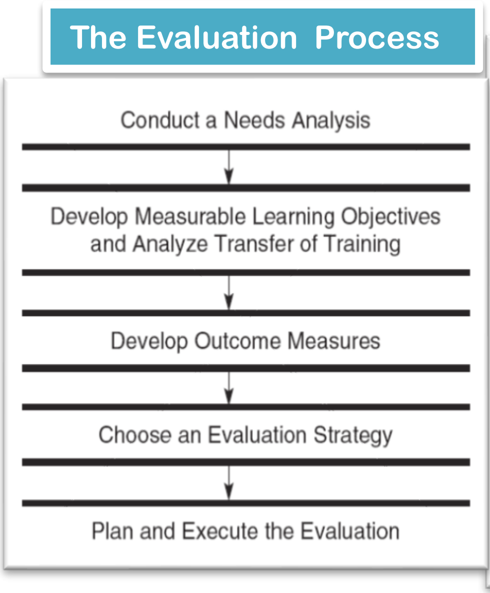
1. Introduction

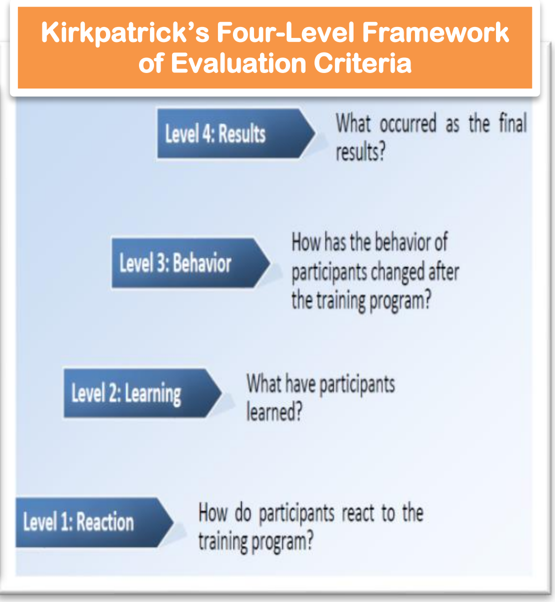
Training is said to be the acquisition of knowledge of skills, and the competencies. It has specific goals of improving one's knowledge, skills and their capacity, capability, performance and their productivity (Ganesh M. & Indradevi R.Dr., 2015). Training and development is becoming an increasingly important function in human resource management. Enrico Giovannini, UNECE (2013) stated that human resources are the most important asset of statistical offices. Appropriate and skilled human resources are essential to ensure the production of high quality statistics and to implement more efficient and effective production processes based on new technologies. Proactive human resources management is essential to achieve the above mentioned change and to allow statistical offices to meet the challenges today and in future.

Riikka Mäkinen, Statistics Finland, UNECE (2013) addressed that statistical work and professionalism have many aspects, in which you can only become skilled by means of on-the-job learning or specific training provided by the statistics sector. For this reason, personnel training is of key importance to statistical institutes. Professional skills in statistical work include know how associated with producing statistics (for example methodology), knowledge of the phenomena on which statistics are compiled, as well as competence relevant to needs for and presentation of statistical information. These aspects comprise the central learning objects of the Training Programme in Statistical Skills.

A training program should be evaluated as to identify the program's strengths and weaknesses, to assess whether content, organisation, and administration of the program contribute to learning and the use of training content on the job, to identify which participants benefited most or leased from the program, to gather data to assist in marketing training programs, to determine the financial benefits and costs of the program and to compare the costs and benefits of training versus non-training investments and the different training programs to choose the best program.

Training evaluation is often defined as the systematic process of collection to determine if training is effective (Goldstein & Ford, 2002). For training initiative to be effective, organisation need to examine the extent to which training and HRD system closely connected with the organisational strategy, and more important, the measure to ensure effectiveness of training and development activities (Haslinda & Mahyidin, 2009).


The Evaluation Process
Conduct a Needs Analysis
Develop Measurable Learning Objectives and Analyze Transfer of Training
Develop Outcome Measures
Choose an Evaluation Strategy
Plan and Execute the Evaluation


Kirkpatrick's Four-Level Framework of Evaluation Criteria
Level 4: Results
What occurred as the final results?
Level 3: Behavior
How has the behavior of participants changed after the training program?
Level 2: Learning
What have participants learned?
Level 1: Reaction
How do participants react to the training program?

Donald L. Kirkpatrick introduced a four-step approach to training evaluation in 1959 (Shelton & Alliger, 1993). He describes his approach in a chapter titled 'Evaluation' in the three editions of the Training and Development Handbook; (1987, 1976, 1967). In these chapters, Kirkpatrick states, 'nearly everyone would agree that a definition of evaluation would be the determination of the effectiveness of a training programme' (1987, p.302). His four steps have become commonly known in the training field as: Level One, Level Two, Level Three, and Level Four Evaluation. The order of the levels is reaction, learning, behavior and results.

2. Methodology

2.1 Definition of Training Program

The training program organised by School of Methodology, Research and Quality (MPK), ILSM during the year 2012 to 2013 were studied. The training program were aimed to provide training in fundamental statistical techniques and operations, statistical frameworks and methodologies on the fields of official statistics.

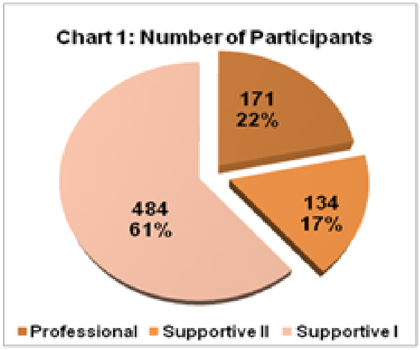
2.2 Scope of the Study

In this research, 39 training programs (N=789) were analysed to determine the level of satisfaction and effectiveness of the training perceived by the participant. While 33 training programs (N=575) were analysed to evaluate the effectiveness of the training perceived by the participant's supervisor.

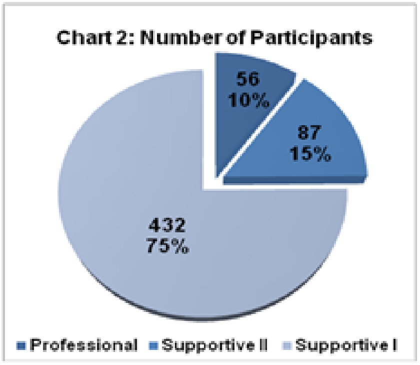
3. Results

The distribution of the participant was higher for supportive I with a total of 484 participants (61%) as shown in Chart 1 and about 75% of supportive I with a total of 432 participants have been evaluated (Chart 2).

Evaluation by participants



Evaluation by participant's supervisor



Most of the participant (80.5%) rated that the training programs are good as shown in Table 2.

Table 1: Statistics

Statistics	Value
N	789
Mean	2.80
Median	3.00
Mode	3
Std. Deviation	0.397

Table 2: Overall level of satisfaction

	Frequency	Percent	Valid Percent	Cumulative Percent
Satisfy	154	19.5	19.5	19.5
Good	635	80.5	80.5	100.0
Total	**789**	**100.0**	**100.0**	

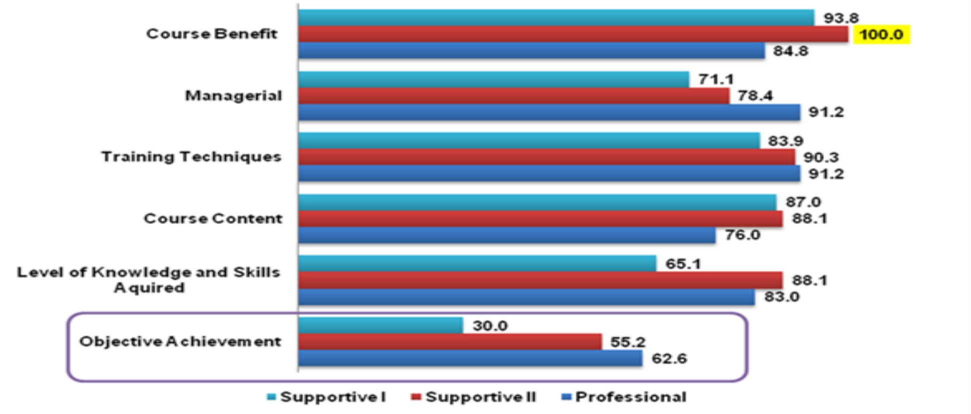
Table 3: Comparison of ratings between designation of the participants

Designation	Mean	Median	Std.Deviation
Professional	2.91	3.00	0.284
Supportive II	2.78	3.00	0.413
Supportive I	2.77	3.00	0.420
Total	**2.80**	**3.00**	**0.397**

On the average as shown in Table 3, a group of Professional has a higher mean overall rating (2.91) compared to Supportive II (2.78) and Supportive I (2.77). There is a higher variability (s = 0.420) in the ratings for Supportive I compared with the others. Overall, the majority of participants in the group are satisfied with the training programs that they participate (median score=3.0=good).

What aspects are the most satisfy the participant?

Chart 3: Distribution of Participants by Component of Training



The supportive II staff stated that all of the courses are benefit to them as they rated the higher score (good). The objective achievement aspects indicate lower percentage rated in the group, as this result should be further investigated in depth analysis.

Does the satisfaction level of participants is differed for the group?

Table 5: Chi-Square Tests

Designation	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.129[a]	2	.000
Likelihood Ratio	18.536	2	.000
Linear-by-Linear Association	13.770	1	.000
N of Valid Cases	789		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 26.15.

There is insufficient evidence to indicate that the satisfaction level of participants distribution is different for the three groups.

Does the training program affect the work outcome, self quality, knowledge and skills?

The results of the pair t-test as shown in table 7, indicated a significant p value p < 0.05 (0.000), therefore it can be concluded that the three aspects has significant influence on training effectiveness perceived by the participant's supervisor.

Table 6: Paired Samples Tests

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Work Outcome Before	7.000	575	1.0763	0.0449
	Work Outcome After	8.278	575	0.7195	0.0300
Pair 2	Self Quality Before	7.292	575	0.9321	0.0389
	Self Quality After	8.368	575	0.6590	0.0275
Pair 3	Knowledge and Skills Before	7.034	575	1.0983	0.0458
	Knowledge and Skills After	8.312	575	0.6943	0.0290

Table 7: Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Work Outcome Before - Work Outcome After	-1.2781	.7919	.0330	-1.3430	-1.2133	-38.702	574	.000
Pair 2	Self-Quality Before - Self Quality After	-1.0757	.6642	.0277	-1.1301	-1.0213	-38.834	574	.000
Pair 3	Knowledge and Skills Before - Knowledge and Skills After	-1.2780	.8066	.0336	-1.3440	-1.2119	-37.990	574	.000

4. Discussion and Conclusion

The score on the satisfaction level of participants indicate that the training program were effective. Although the satisfaction level was high as perceived by the participants, there is still a lot of evaluating task to be done in order to identify the program’s strengths and weaknesses.

This study revealed that our evaluation just fulfils the level one of the Kirkpatrick Four-Level Framework of Evaluation Criteria.

References

1. Ganesh M., & Indradevi R, Dr. (2015). Importance and Effectiveness of Training and Development*. Mediterranean Journal of Social Sciences*, 6(1), 334-338.
2. Goldestein, I. L. & Ford, J.K (2002). Training in organization: Need Assessment Development and Evaluation. (4th, Eid) WARDSWORTH.
3. Haslinda, A., & Mahyuddin, M.Y. (2009). The effectiveness of training in the public service. *American Journal of Scientific Research,* 6(2009), 39-51.
4. Neeraj S. Borate, Gopalkrishna and Sanjay L.Borate (2014). *A Case Study Approach for Evaluation of Employee Training Effectiveness and Development Program.* GB14 Second International Conference on Global Business, Economics, Finance and Social Sciences.
5. Shelton, S., & Alliger, G. M. (1993). Who's afraid of level 4 evaluation? Apractical approach. *Training and Development Journal*, 47, 43–46.
6. United Nations Economic Commission for Europe, (2003). *Human Resources Management and Training: Compilation of good practices in statistical offices*. United Nations.

Addressing the contribution of foreign investment to Malaysia's economy

Mohd Harith Faiz Md Saad
Department of Statistics Malaysia

Abstract

In the era of globalisation and technological progression, multinational enterprises are engaged globally by establishing their operations abroad to expand their businesses. This expansion initiated the countries to assess the contribution of the foreign companies to their economy. Hence, the Manual on Statistics of International Trade in Services 2010 (MSITS 2010) recommended the standards to gauge the contribution in the economy through the Foreign Affiliate's Statistics (FATS). Conceptually, the Inward FATS refers to the activity of foreign affiliates' resident in that controlled by parent companies outside of the country which owns more than 50 per cent of the equity. Malaysia developed the first Inward FATS in 2009 which was published in 2010 for internal circulation. Over the years, the number of foreign affiliates in Malaysia showed an increasing trend, whereby there were 2,964 inward affiliates recorded in 2016 compared to 2,742 affiliates in 2010. These affiliates mainly involved in the manufacturing and services sectors and the highest numbers of foreign affiliates were from Asia region. The main purpose of this paper is to present the growing role of the foreign affiliates to Malaysia's economy.

Keywords

Inward; Foreign Affiliates

1. Introduction

Inward FATS refer to economic statistics relating to the operation of foreign affiliates in Malaysia. Operation of foreign affiliates in Malaysia expanded for the past few years which may also contributed to the economy. This can be measured by quantifying the companies' cross-border investment, either by foreign enterprise investment inside the country or vice versa and other related information. This phenomenon is closely connected with the issue of economic globalisation and the displacement of productive resources.

Hence, this paper focuses on the represents of the information on Foreign Affiliates Statistics in Malaysia (Inward FATS) in 2016. Inward FATS reflect the performance of foreign-controlled companies in Malaysia which hold more than 50 per cent of the equity. This statistic measures the commercial presence of foreign affiliates at Malaysia's market level based on the International Trade in Services, particularly for the mode of supply three (Mode 3). This statistic can be used by government agencies, economists, academicians as well as individuals for planning and formulations policies, economics analysis, projections and to assist in business development planning.

2. Literature Review

Tattasawart (2011) stated that the Inward FATS depicted a depth understanding related to the foreign investment's activities to the economics of the host country. The result of these economic activities created the jobs formation and the exchangeable of trade and investment, technological and managerial skill transfers. In addition, the inward became the most important factor in expanding the size of Thailand's economy through the investment promotion package that been encouraged by the government. The foreign affiliates activities in Thailand focusing more on the manufacturing sector.

According to Statistics New Zealand (2014), the companies that have located enterprises at a foreign country can reach the capital sources to their country by hiring more workers and also these foreign affiliates involved mainly in manufacturing and wholesale trade recorded highest sales of revenue.

3. Methodology

The compilation is based on the guidelines recommended in Manual of International Trade in Services (MSITS) 2010 of United Nations and Balance of Payments and International Investment Position Manual, Sixth Edition (BPM6). Moreover, the Malaysia Standard Industrial Classification (MSIC) 2008 Ver.1.0 is used to classify the inward statistics by economic sector.

In this paper, the data have been retrieved mainly from the Economic Census, Census of Distributive Trade and other economic surveys conducted by DOSM. Within the year which the EC is not conducted, the data retrieved from the Annual Economic Survey (AES) which conducted by the respective divisions in DOSM. The census and the surveys contained the question regarding the ownership of residents or non-residents in Malaysia. Those establishments that stated the equity ownership by non-resident of more than 50 per cent and above are classified as foreign affiliates. The important information that acquired from EC is on the country of an ultimate parent company of the establishments whereas most of the companies tend to report the immediate parent company.

4. Geographical Distribution of Affiliates in 2016

For the year 2016, a total of 2,964 foreign affiliates were located in Malaysia, an increase of 8.1 per cent (2010: 2,742 affiliates) which controlled by enterprise group with a decision centre (global headquarter) located outside Malaysia. Among the economic sectors, foreign affiliates were largely involved in the manufacturing sector which contributed 57.1 per cent. The services sector was the second place contributed about 34.5 per cent **(Figure 1).** The remaining affiliates with the share of 8.4 per cent were involved in the construction, agriculture and mining sectors.

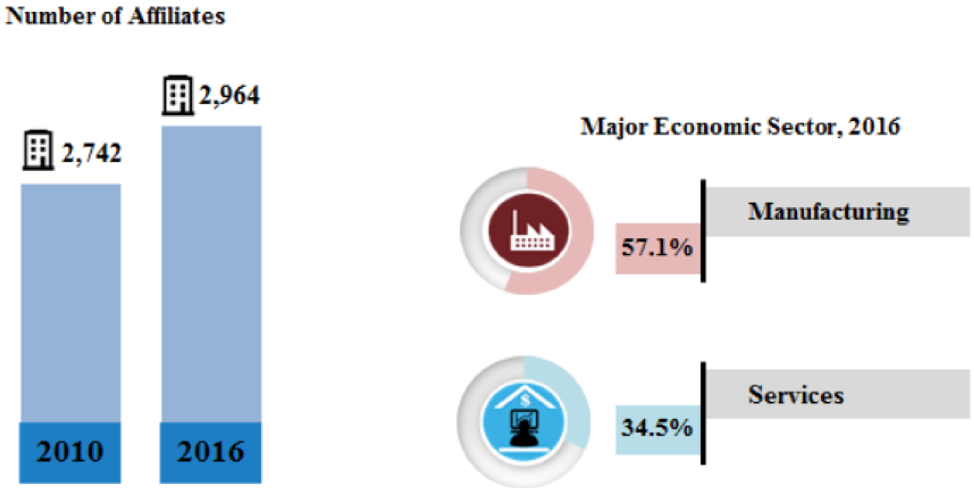


Figure 1: Number of affiliates and percentage share by major economic sector, 2016

Malaysia is one of the attractive investment destinations for Asia companies whereby 64.6 per cent or 1,914 affiliates were from this region (Figure 2). This was followed by companies from Europe and America which contributed 21.0 per cent and 12.2 per cent respectively. The remaining percentage of foreign affiliates with the share of 2.2 per cent was Oceania and Africa region.

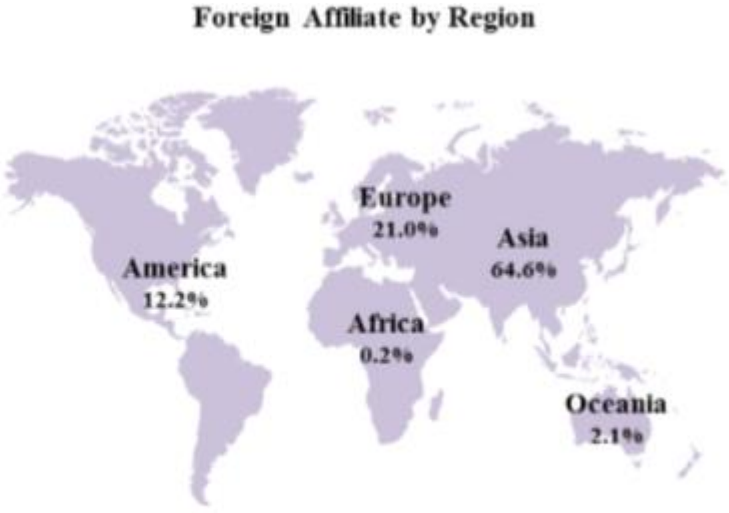


Figure 2: Percentage share of foreign affiliates by region

5. Contribution of Inward FATS

5.1 Economic Value

The foreign affiliates generated an economic value of RM222.7 billion compared to RM159.8 billion in 2010. The economic value as a percentage of GDP climbed up to 18.1 per cent in 2016. Within sectors, the manufacturing sector represented the largest contributor of 47.2 per cent (Figure 3) primarily in electrical and transport equipment. This was followed by the mining sector with 27.7 per cent.

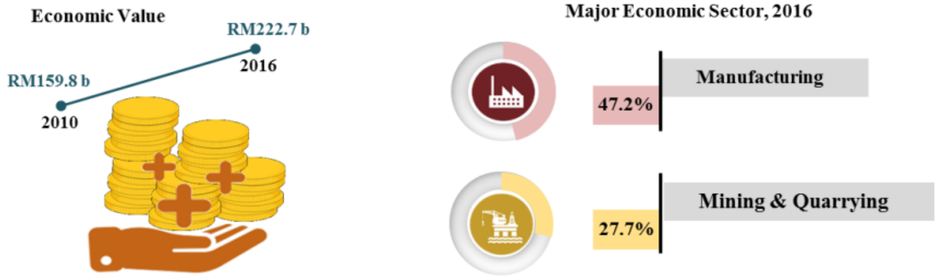


Figure 3: Economic value and percentage share by major economic sector, 2016

5.2 Gross Fixed Capital Formation (GFCF)

The Gross Fixed Capital Formation (GFCF) of foreign affiliates in Malaysia recorded RM63.8 billion in 2016 as compared RM41.3 billion in 2010. The highest shares were recorded in manufacturing sector constituted 42.4 per cent followed by mining sector with 39.3 per cent respectively **(Figure 4).**

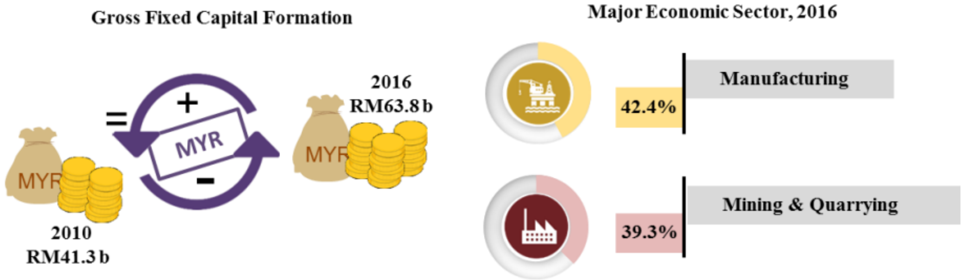


Figure 4: Gross Fixed Capital Formation and percentage share by major economic sector, 2016

5.3 Job Creation

Job creation is also an indicator to further measure the effectiveness of inward FATS in Malaysia. In this context, foreign affiliates hired 847,269 employees in 2016 compared to 798,895 persons in five years back (Figure 5). More than 70 per cent jobs were created in manufacturing sector which recorded 640,760 employees, largely in electrical and transport equipment.

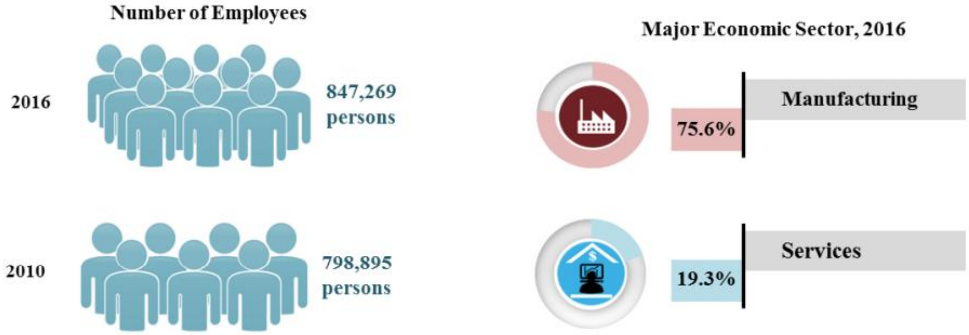


Figure 5: Number of employees and percentage share by major economic sector, 2016

5.4 Imports and Exports

Foreign affiliates trading activities expanded from RM255.3 billion in 2016 (2014: RM238.0 billion) for exports while imports registered marginally decrease from RM296.8 billion in 2014 to RM292.8 billion which amounted for 34.0 per cent and 35.1 per cent to Malaysia's total imports and exports respectively. The highest share was contributed by affiliates in manufacturing sector as depicted in **(Figure 6).**

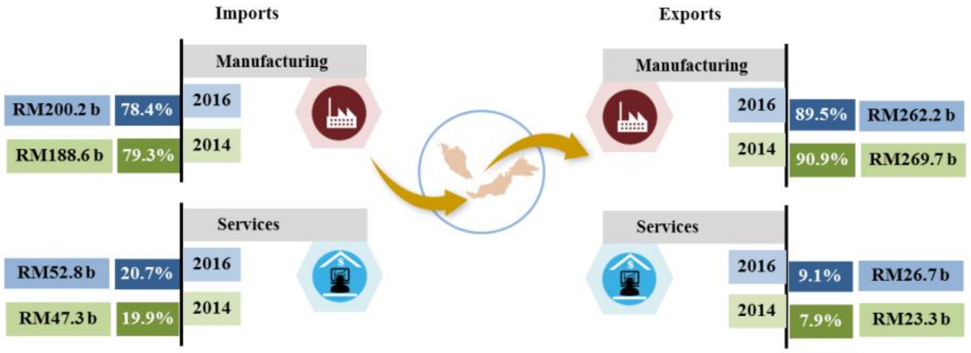


Figure 6: Top economic sectors based on imports and exports and percentage share, 2016

6. Discussion and Conclusion

In Malaysia, foreign-controlled enterprises concentrated mostly in the manufacturing and services sectors. Indicators such as economic value, GFCF and job creation showed an increasing trend for the past five years suggesting the growing role of foreign affiliates to Malaysia's economy. Such information is crucial for stakeholders and policymakers to assess the impact of foreign controlled enterprises on economy and understand the direction of foreign trading industries. It also used to monitor the effectiveness of internal market and the gradual integration of economies within the context of globalisation. DOSM is taking initiative to enhance the Inward FATS compilation by taking into consideration the varying degree of capacity to measure the FATS. As a path forward, DOSM is upgrading Inward FATS by expanding new variables such as compensation of employees and output to sit the comparison with the rest of the world.

References

1. Department of Statistics, Malaysia, *Statistics on Foreign Affiliates* in Malaysia 2016 (2018), Putrajaya.
2. Manual on Statistics of International Trade in Services 2010. United Nations Publication.
3. Tattawasart, O. (2011). Towards FATS and Beyond: The case of Thailand. Data Management Department, Bank of Thailand.
4. Statistics New Zealand (2014). New Zealand's inward foreign affiliate statistics.
5. Essays, UK. (2013). Analysis Of Foreign Direct Investment In Malaysia Economics Essay.
6. Kornecki, Lucyna. (2013). Performance of Inward and Outward U.S. Foreign Direct Investment during Recent Financial Crises. Managerial Issues in Finance and Banking: A Strategic Approach to Competitiveness.

Malaysian economic indicator: How well does it provide signal on economic crisis?

Siti Nuraini Rusli, Nur Hidaah Mahamad Rappek
Department of Statistics Malaysia

Abstract

The ability to predict the future economic situation especially slowdown is an advantage as it could give early signals and thus, measures could be taken to mitigate the effects to the nation. Malaysia has experienced economic recessions with the latest occurred in 2008. Malaysian Economic Indicator consisting Leading, Coincident and Lagging Indexes are used to monitor the Malaysia's near-term economic direction on a monthly basis. It is useful in assisting the policy makers, investors, researchers and the public. As such, it is important to demonstrate the capability of the existing indicators in providing reliable direction especially in economic crisis for decision makers. In other words, could the indicators provide signals in real-time? The paper answers the question using the Three D's Method by the Conference Board.

Keywords

Leading Economic Index; Business Cycle; Three D's Method

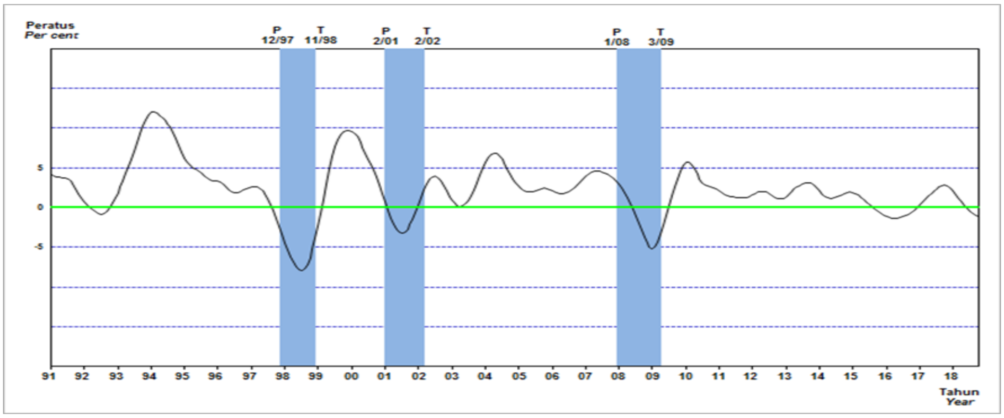
1. Introduction

Malaysian Economic Indicator consisting Leading, Coincident and Lagging Indexes are used to monitor the Malaysia's near-term economic direction on a monthly basis. It is useful in assisting the policy makers, investors, researchers and the public. As such, it is important to demonstrate the capability of the existing indicators in providing reliable direction especially in economic crisis for decision makers.

Coincident indicators (CI) which comprises of employment, income, production, capital utilisation and retail trade, comprehensively measure the overall current economic performance. Hence, they define the business cycle. Leading indicators (LI) consists of money supply, industrial index, imports, housing permits, sales value and new companies registered, consistently lead the CI. In other words, the series tend to shift the direction in advance of the business cycle. On the other hand, the Lagging indicators (LG) validate the signal given by LI and CI specially to confirm the turning points (peaks and troughs) of the LI and CI.

Based on the LI historical data, the signal provides reached 80.0 per cent of accuracy with the short-term signal between four to six months in advance as against Malaysia Gross Domestic Product (GDP). The LI is able to provide early signal of turning points for peak or trough as shown in Chart 1.

Chart 1: Annual Growth Rate of Leading Index (Smoothed) and Business Cycle



Source: Malaysian Economic Indicators: Leading, Coincident & Lagging Indexes

The line in Chart 1 shows the smoothed annual growth rate of LI from January 1991 to October 2018. Meanwhile, the blue shaded regions refer to the peak and trough of business cycle which implies the real recession period in Malaysia. There are three recession period from January 1991 to October 2018. The LI provides early signal of recession as follows:

- Lead 3 months ahead for Asian Financial Crises in November 1998.
- Lead 10 months ahead for Global Economic Slowdown in February 2002.
- Lead 2 months ahead for US Debt Crisis/Euro Zone Crises in March 2009.

The composite index is the combination of individual indicators which measures the economic cycles behaviour. The advantage of composite index compared to individual analysis is the tendency to smooth out some of the volatility of the series. The composite index is generally more reliable in generating clear and consistent turning points than individual indicators. Table 1 shows components of Leading, Coincident and Lagging composite indexes.

Table 1: Malaysia's Composite Index

LI	CI	LG
1) Real Money Supply, M1	1) Total Employment, Manufacturing	1) Unit Labour Cost, Manufacturing
2) Bursa Malaysia Industrial Index	2) Real Salaries & Wages, Manufacturing	2) Number of Investment Projects Approved Number of New Vehicles Registered
3) Real Imports of Semi-Conductors	3) Industrial Production Index	

4) Real Imports of Other Basic Precious & Other Non-ferrous Metal
5) Number of Housing Units Approved
6) Expected Sales Value, Manufacturing
7) Number of New Companies Registered

4) Real Contribution, EPF
5) Capacity Utilisation, Manufacturing
6) Volume Index of Retail Trade

3) Exports of Natural Gas & Crude Oil
4) CPI for Services

Another important source of information that can be used to confirm the business cycle is diffusion indexes. According to Meşter (2007), diffusion indexes are not redundant even though they are based on the same set of data as the composite indexes. Diffusion indexes used to measure how widespread a particular business movement (expansion or contraction) has become and measure the width of that movement. Meanwhile, the composite indexes, differentiates between small and large overall movements in the component's series. Occasionally, these two indexes move in different directions and the dissimilarity can be used to confirm or anticipate cyclical turning points.

The objective of this paper is to examine the accuracy of time series forecasts of recessions for the Malaysian economy, using composite and diffusion of LI for the past recession period from January 1991 to October 2018 using the "Three Ds" approach.

2. Methodology

Identifying the turning points is not an easy task even for the expert analysts. In practice, the economist and analysts apply rules of thumb to help identify recent turning points and coming recession (Meşter, 2007). One of them is using the "Three D's" – the duration, depth and diffusion method.

Based on the "Three Ds" approach, the longer the weakness continues, the deeper it becomes; and the more widespread it turns out to be, the more likely recession will occur. According to this approach, a recession usually follow when the (annualized) six-months decline in the LI reaches 4.0-4.5 per cent and the six month diffusion index falls below 50.0 per cent (The Conference Board, February 2010).

The LI does not increase or decrease in long continuous movements. Expansions are spread with a few months of decline, and recessions include months of increase. Interpreting declines in the LI using duration eases the emergence of short-term patterns or trends. Meanwhile, the depth and diffusion of those declines help to define how likely a short-term fluctuation is to be a recession warning. This motivates the use of the Three D's in conjunction with one another.

The duration interval of a decline is perhaps the most obvious signal of imbalances in the economy, which might eventually enter a recession as a result. However, for reliable interpretation of these declines, most economists also require a significant downward movement in the index, as well as declines in the majority of the component series. These are the second and third aspects of the Three D's - depth and diffusion. In brief, the greater the decline (depth), the more likely it is that a serious economic downturn will occur and the more likely that the decline is not a random fluctuation. The seriousness of the decline can be assessed by calculating the per cent change of the decline over a given span of months.

Source of data for this study is composite and diffusion of LI for the period of January 1991 to October 2018 which was taken from Malaysian Economic Indicator: Leading, Coincident & Lagging Indexes.

3. Result

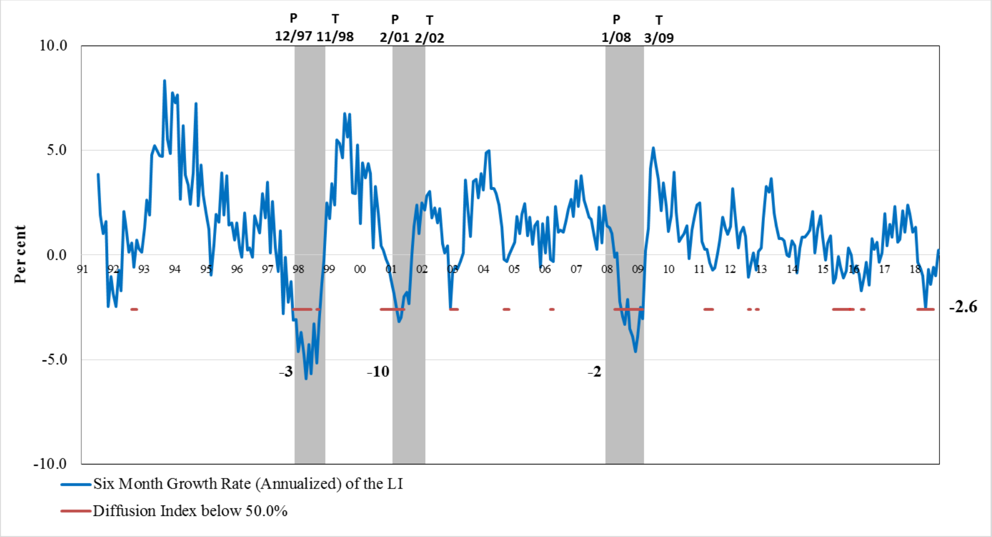
Chart 2 shows the annualized six-month changes and the durations in which the diffusion index falls below 50 per cent. Numbers next to the grey shaded region denote the lead times of each of the past peak for three recessions since 1991. The average lead is five months.

Looking at data month on month, it is clear that the LI has many brief declines that have nothing to do with cyclical downturn in the economy. For example, in April 2002 reading for LI is 3.0 per cent and declined to -2.8 in December 2002. It started to rebound in January 2003 which indicating LI gave a false alarm of recession for this period.

Generally, a recession started when two criteria of "Three D's" approach are met simultaneously across a six-month period. Based on this study, the cut of point of annualized changes to ensure that it is recession is below -2.6 per cent for data series January 1991 to October 2018 and the diffusion index is below 50 per cent.

As of April 1998, the annualized changes rate is -5.9 per cent which is below -2.6 per cent, and the diffusion index was below 50.0 per cent. Concurrently, the annualized changes rate in December 2008 is -4.6 per cent and the diffusion index was below 50.0 per cent. Hence, LI was managed to give an early signal of the real recession for both periods. Recently, the six-month growth rate of LI in May 2018 reached -2.6 per cent. However, the growth rate starts to pick-up again. So, it is a false signal of recession.

Chart 2: Six-Month Growth Rate (Annualized) of the LI:
January 1991 to October 2018



4. Discussion and Conclusion

The duration of a decline is perhaps the most obvious indication of imbalances in the economy, which might eventually enter a recession as a result. According to the result of Three D's approach conducted by the Conference Board using the Composite Index of Leading Economic Indicators for the period of 1959 to 2000, they found a false signal. The false signal should also be recognized that these predictions of recessions that did not materialize are not necessarily flaws. Sometimes false signals are quite insightful because the LI is sensitive enough to point to imbalances in the economy that could result in a recession. The LI turned down significantly, even though a recession did not follow. Because economic growth weakened slightly thereafter, many economists believe that the index warned appropriately that the risk of a recession had increased.

The false signals occur because of reliance on a rule-based, naive reading of the LI. If all available indicators are interpreted thoroughly, individually as well as in combination, the risks of the economy entering a recession can be evaluated more realistically. To increase the chances of getting true signals and reduce those of getting false ones, it is advisable to rely on all such potentially useful indicators as a group." (Business Conditions Digest, May 1975).

For the Malaysia Case, Three D's approach is able to give an early signal of recession based on LI time series data from January 1991 to October 2018. However, it is suggested to use other approach such as Markov-Switching model developed by Hamilton (1996) to examine the accuracy of time series forecasts of recessions and expansions of the Malaysian economy.

References

1. Burns A.F., a. M. (1946). Measuring Business Cycles. National Bureau of Economic Research, New York.
2. Business Conditions Digest. (May 1975).
3. MEŞTER, I. T. (2007). Indicator Approach to Business Cycle Analysis. 8.
4. Proietti, T. (2004). New Algorithms for Dating the Business. *Elsevier Science*, 28.
5. The Conference Board. (2000). Business Cycle Indicators Handbook. 156.

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

Updating views on black litterman model

Retno Subekti, Abdurrahman, Dedi Rosadi
Gadjah Mada University, Indonesia

Abstract

The Black Litterman model is known as a model in financial industry for improving equilibrium return with the specific view from investor. In this research we propose the procedure to gain the outperform result based on renewed weight of Black Litterman model. Since we update the view, so we need to determine the time to conduct the renewing calculation. This conditional view is depending on time for investment and portfolio return. We still treat a simple procedure with Moving Average as time series method in predicting views. The result show that we derive a better performance when we work on dynamic portfolio with updating view.

Keywords

Black Litterman; views; time series method

1. Introduction

Modelling in mathematics is like how to catch the phenomenon in the world, what is the problem and how to solve it. In the portfolio management, the problem is how to put the proper allocation in each asset in order to gain the optimal portfolio. We have known the classical model, Mean Variance /MV Markowitz as the pioneer in terms of how to distribute the capital for getting portfolio more optimize. As a starting model, the emerged MV model has a great impact in the world investment. Many authors developed MV model into sophisticated optimization problems. One of the weakness of MV model is the input is only historical return and variance, we cannot entry the feeling or new information from manager/investor.

In 1992, Black Litterman has been developed with its aim is to cope the problem in classical MV model. Black and Litterman (1) refered into the bayes rule to combine the feeling and benchmark condition (equilibrium). The explanation of the BLM formula can be tracked from Satchell (2), Meucci (3) or in briefly, some of explanation from many authors were summarized in Walter (4).

The other question is how to get investor's feeling and put it into the computation in Black-Litterman practically. We can ask directly to the manager as a subjective input or in many references, we can approach it with time series method (5, 6). This research is continuing the previous work in developing of feeling restatement with dynamic approach. So, one of the keywords in BLM discussion is how to build the opinion in terms of incorporate with Capital Asset Price Model and gain Black-Litterman return as a new optimizer in portfolio.

2. Methodology

We begin this research with introduction the Black–Litterman return, views statement in many literatures in this section. Herafter, by incorporate the updating views for multi period, we present our numerical example and evaluation of portfolio performance in the result and discussion section.

2.1 Black Litterman Return

It is known that an emerging of a new model will have variety of response, including Black-Litterman Model is similar to the proposed Capital Asset Pricing Model when it became familiar from William Sharpe (1964), John Lintner (1965), Jan Mossin (1966) and Jack Treynor (1961). Black-Litterman model is a new formula that emerged in 1990 by Robert Litterman and Fischer Black in their article (1). This new model had a numerous explanation because in the original article is not clearly yet. (2) proposed a detailed bayes method for BLM construction and this model is convinced will help in financial portfolio. The idea is equilibrium condition and investor's feeling are blended to form a new posterior return as an expectation in the investment.

When tracking this model, everyone who has relation in the same topic will be inspired and try to learn, develop or make a new contribution on it. The growth of references for developing this model is very greatly from 2000 until now. Many researchers discuss about how to build the model from theoretical background and continue it with how to implement the model into reality. In other words, how to practice step by step this certain model, surely with many assumptions are required to limit the discussion of solving problem in the model.

The unique of this model is when we can put the opinion or feeling as a view in the future into the processing of optimization problem. Starting from implied return equilibrium, CAPM which is normally assumption then we focus on building views as a future return by Meucci (1) and Idzorek (2).

$$\mu_{BL} = \boldsymbol{\pi}' + (\boldsymbol{\tau\Sigma})\boldsymbol{P}'\,(\boldsymbol{P\tau\Sigma}\mathbf{P}')^{-\mathbf{1}}(\boldsymbol{Q} - \boldsymbol{P\pi}')$$

The detailed formula and its explanation can also be traced from Walter (3) who explain some of the difference from the authors investigating the formula of Black Litterman return. The outline is original version, alternative, theil mixed regression and sampling theory. In this research, we limited the discussion for the view development in BL formula.

2.2 Views Statement in Black Litterman

In many references, there are two types of views statement in BLM namely relative views and absolute views. Relative view is the statement when someone compare two or more assets related with the opinion of their return in the future while absolute view is the statement about return prediction for particular asset. It is clearly when we describe it in this following example.

A portfolio consists of 3 assets, namely A, B and C. Investors can state 3 views or more, but in this example only 2 views are used.

View 1: "I believe that asset A will return 1,75%"

View 2: "I convinced asset B will return 2% beyond C assets"

E(r) is an estimation of returns from investors, with these three assets, A, B and C while two views are the investor 's opinion which can be expressed in such linear combination of expected return.

$$\mathrm{E}(\mathrm{r}_\mathrm{A}) = 0{,}0175$$
$$\mathrm{E}(\mathrm{r}_\mathrm{B}) - \mathrm{E}(\mathrm{r}_\mathrm{C}) = 0{,}02$$

To transform it into linear combination, we need to define the link matrix or the pick matrix, P so that equation can be written as follows,

$$1.\mathrm{E}(\mathrm{r}_\mathrm{A}) + \ 0.\mathrm{E}(\mathrm{r}_\mathrm{B}) + 0.\mathrm{E}(\mathrm{r}_\mathrm{C}) = 0{,}0175$$

$$0.\mathrm{E}(\mathrm{r}_\mathrm{A}) + 1.\mathrm{E}(\mathrm{r}_\mathrm{B}) - 1.\mathrm{E}(\mathrm{r}_\mathrm{C}) = 0{,}02$$

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix} \ \boldsymbol{E}(\boldsymbol{r}) = \begin{bmatrix} E(r_A) \\ E(r_B) \\ E(r_C) \end{bmatrix} \boldsymbol{Q} = \begin{bmatrix} 0{,}0175 \\ 0{,}02 \end{bmatrix}$$

Sometimes, it is possible to have an opinion for some of assets simultaneously, such as when we believe that two assets will return outperformance two other assets. For example, the statement about asset C and D will predict 1,5 % outperform asset E and F. So, we can express

$$\boldsymbol{P} = [0.5 \quad 0.5 \quad -0.5 \quad -0.5] \ \boldsymbol{E}(\boldsymbol{r}) = \begin{bmatrix} E(r_C) \\ E(r_D) \\ E(r_E) \\ E(r_F) \end{bmatrix} \ \boldsymbol{Q} = [0.015]$$

In practice, we do not need to put the views in all assets in the portfolio, at least we have one statement about the future return to work with Black Litterman model. This is the unique of BL model because we can improve the starting point which is equilibrium point with views/opinion.

The next question is how to write it in statistical words to describe the combination process as same as fig.1. In (2), starting point for BL is neutral equilibrium market which is CAPM and it is stated as a prior distribution while views from investor is categorized as another information. The visualization of the combination process is described in fig.1 and showed unclear explanation about how and what method to ended up with posterior distribution. Idzorek (2) proposed a guide to involve with Black Litterman in practice.

On the other hand, according to (4) this view expressed by investor/manager is the prior distribution, Prob (E(r)). Through Bayes rules, we define Probability density function (Pdf) with notation Prob in order to distinguish with P as a Pick matrix.

$$Prob(E(r)|\pi) = \frac{Prob\ (\pi|E(r))\ \ Prob\ (E(r))}{Prob\ (\pi)}$$

Notice that the distribution of views return, (E(r)) is unknown, we assume it is normally distributed. Prob (π) represents the marginal probability of equilibrium returns and Prob $(\pi|E(r))$ is conditional probability of equilibrium return given return prediction from investor.

There is still uncertainty for this explanation, the Bayes theorm which is employed in the construction of Black-Litterman return is not similar with the discussion about Bayesian inference. The discussion about this confusion open the space to explore more the formula. It is familiar that in Bayesian, we determine the likelihood, information sample, and posterior. While in the original reference (5), the authors did not refer to likelihood but the authors mention about prior and posterior. Based on (4) it can be said that BLM is compatible with Bayes rule so that it is called BLM working on Bayesian environment. In the discussion of the view, there are many extensions to formulate it before the combination process with CAPM such as in (3) which clarify the Black-Litterman return via regression perspective

In this model, the future return is denoted by Q. We propose the updating views to renew the Q return based on data practically. Thus, BL return will change depend on the views dynamically. The important role for this experiment is determination the time when we can renew the views. In a simple practice, we can utilize plot of time series and interpret from the fluctuation of historical data. Even though we can determine it from the illustration of time series data for each asset at a glance, it is better to refer the time series method and to ensure the minimum error to get closer with the actual data.

3. Result

Numerical example

In this research, we will use a portfolio of 7 top stocks from the LQ45 stock index in Indonesia. The LQ45 stock index consists of the 45 most traded stocks of the Indonesian Market main stock index JKSX. This portfolio is based on weekly data from February 22, 2016 to February 3, 2017. A market portfolio is required to calculate equilibrium return so we will use the JKSX as the market portfolio in applying BLM. We will list the shares by their ticker from www.yahoofinance.com, namely WIKA, TLKM, SMGR, SMRA, UNVR, UNTR and SSMS with their respectively sector and its summary statistics.

Table 1. Description Data

No.	Stock	Description	Average return	Std dev return	skewness
1	WIKA	construction	-0,0015	0,041063	0,068558
2	UNTR	heavy equipment products	0,003025	0,034093	0,098283
3	UNVR	consumer goods	0,00654	0,048661	-0,23986
4	TLKM	telecommunication network and service provider	-0,00109	0,040671	0,666554
5	SMGR	cement producer	0,006251	0,030345	0,14745
6	SMRA	real estate development company	-0,00263	0,047904	0,389205
7	SMSS	palm oil company	0,000342	0,045838	-0,48442
8	JKSX	market portfolio	-0,0015	0,041063	0,068558

Based on summary statistics in Table 1, we select 3 stocks such as UNTR, UNVR and SMGR which are have positive return and more than 0,1 % return on average.

In order to apply the BLM we need to prepare with all components such as views, market return and free risk return to build the CAPM. We propose an updated view to gain more profitable result rather than fix views.

In this preliminary research we focus on simple time series method such as Moving average. We then experiment with all assets and calculate the profit in two periods a head. As an illustration, we investigate a weekly report value of portfolio. Regarding with BL formula, the views is built from MA (2) until MA (6) and we describe in fig. 2.

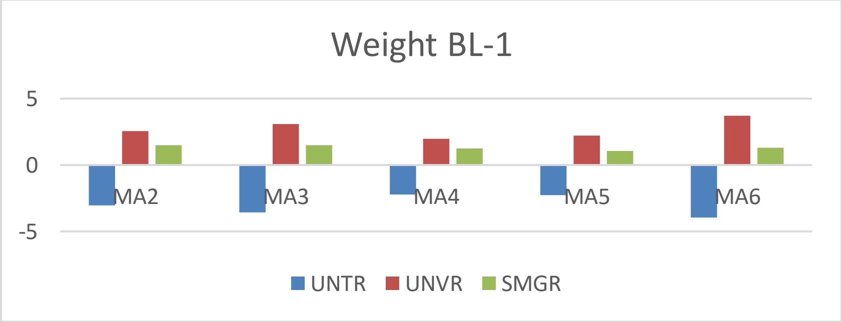


Figure 2. Weight BL initial.

The bar chart illustrated the weight in one period, in this first investment, the allocation for UNTR is negative, it means, we did not set the portfolio without short sale. We observed all the portfolio value in next period and described it in the following graph to see when we adjust the view.

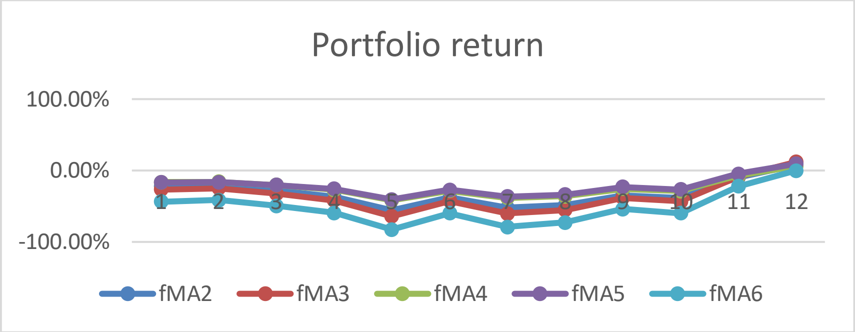


Figure 3. Porfolio Return Observed

For almost entire investment, the portfolio in underperform and in loss condition for all views until time t+11. In this situation, we still wait and hold the portfolio and in t=12 the portfolio can be executed because we will gain the positive return. Based on the graph, we get the best estimate view for the next period is resulted from MA-3. This portfolio is continued to execute in the next second period (May 8, 2017- July 17, 2017), for that reason, we update the views based on each MA, weight BL-2 in Fig.3.

The bar chart in Fig.3 provide the allocation for those assets after the adjusted views. There is a change for each asset extremely for UNTR and UNVR, while SMGR has a less proportion.

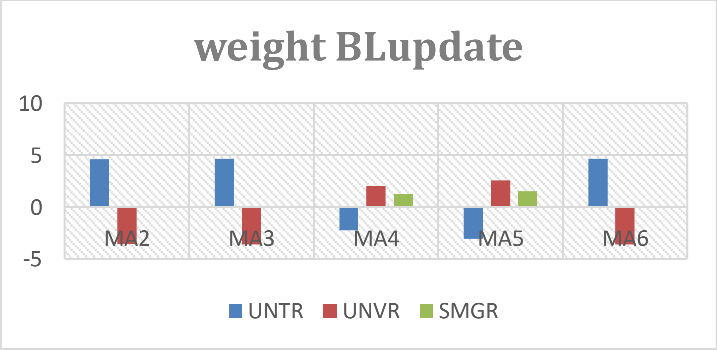


Figure 4. Renew weight BL based on updating views.

We observe then the following value of portfolio based on the new views in Fig 4. In order to find the best estimate for view, we compare the result of portfolio return with the fix view from the beginning in the Fig. 5.

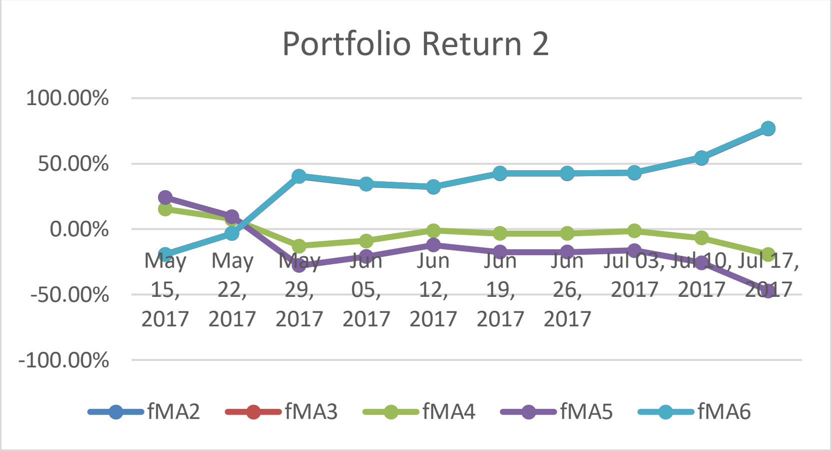


Figure 5. Portfolio Return Observed 2

The portfolio returns in second observation for three portfolios in Fig. 4 which is resulted from MA-2, MA-3 and MA-6 respectively are likely similar and become the best estimate for this portfolio. Meanwhile, we still observe the portfolio with the initial views and display the return to compare with the updated one.

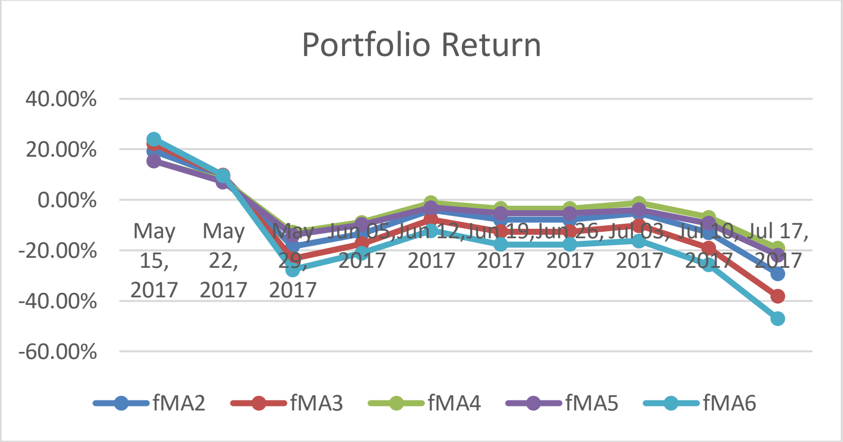


Figure 6. Portfolio return with initial weight.

The Fig. 5 reported that the value decreased into almost 20-40% after t+2. The impact is the portfolio in loss condition for many times. The profit condition only survived in short term t+12 until t+14 according to Fig. 2 and Fig 5.

4. Discussion and Conclusion

In this research, we develop model from the empirical data. Beginning with observation and creating views from simple method such as Moving Average (MA). We tried several MA with different k from 2 to 6 then we predict a view in implementing the Black Litterman model. We use unconditional views to build portfolio with Black Litterman for the first assumption so that we do not reproduce a new view in holding period until the certain t (time) when we excute the portfolio. Then, we apply updating views to renew the portfolio in t and we observe the result. Comparing both Fig. 4 and 5 show that updating view through MA-3 as a portfolio 2 give the best estimate and result rather than a portfolio without renewing weight. The portfolio return is underperformed for almost entire investment horizon time, it only outperforms in short term which is t+11 until t+14 according to Fig. 2 and Fig 5. This is relevant with (6) when view is predicted with RBFNN it yields a best performance based views for 10 period ahead. We cannot hold with views for long term periods because the error will be increase.

Let say, $\mu_{t|t-1}$ as predicted mean conditional to the posterior return in t-1 and treat it as a renew Q. Then we described it as a procedure in building the portfolio BLM with updating views. The following step by step is a procedure incorporating updating weight in BLM:

1. Create an initial view as the difference between predicted mean $\hat{\mu}_{t+1}$ from time series method, such as MA and the return at t, r_t. We denote as $q_0 = \Delta_{\hat{\mu}_{t+1,r_t}}$
2. Determine the t for execute the portfolio when it reaches the expected return. When it does not gain the expected then hold the portfolio without revision of the Q. Let say t as a time to execute then we create new prediction based on $\mu_{t|t-1}$
3. Determine the posterior return as a new result on portfolio return

5. Conclusion

Based on the result, we investigate only two periods in evaluation and show that investor's view is no longer to be put in BL portfolio for the entire investment horizon time. We demonstrate how to renew the weight based on views updating in BLM. The research is still limited on practical for the building new procedure and we ignore the component of error to be listed in this stage. We will continue for the further research to be matched with mathematical modelling.

References

1. Meucci A. 2010. The Black-Litterman Approach: Original Model and Extensions. *http://ssrn.com/abstract=1117574*
2. Idzorek TM. 2005. A STEP-BY-STEP GUIDE TO THE BLACK-LITTERMAN MODEL Incorporating user-specified confidence levels
3. Walters J, Estimation M. 2014. The Black-Litterman Model In Detail. , pp. 1–65
4. Satchell S, Scowcroft A. 2007. A demystification of the Black-Litterman model: Managing quantitative and traditional portfolio construction. *Forecast. Expect. Returns Financ. Mark.*, pp. 39–53
5. Black F, Litterman R. 1992. Global Portfolio Optimization. *Financ. Anal. J.* 48(5):28–43
6. Wutsqa DU, Subekti R, Kusumawati R. 2016. Radial Basis Function Neural Network for Views Prediction on Black-Litterman Model. . 3(1):71–78

How will we know when we have achieved our dream: a world free of poverty?

Jonathan Haughton
Suffolk University

Abstract

In this paper we examine the sensitivity of measures of poverty to the choices made by analysts about "internal" decisions, mainly using survey data from Rwanda, but with appropriate reference to the experience of other countries. We focus on four (of the many possible) "internal" issues where the assumptions made by the analyst may matter: valuing auto consumption, adjusting for prices over time and space, specifying adult equivalents, and establishing a poverty line. Of these, the most difficult is getting the prices right. Unless there is some consistency, or code of best practice, in the methods used, the results of analysts, even using the same underlying data, will vary widely and may tell different stories.

Keywords

Poverty; Rwanda; adult equivalents; auto consumption

1. Introduction

Less than two decades ago the World Bank published a book by Sandra Granzow (2000) entitled: *Our Dream: A World Free of Poverty*. This has now been firmly established as the first and most prominent of the UN's Sustainable Development Goals, which is to "end poverty in all its forms everywhere", or more concretely, "by 2030, eradicate extreme poverty for all people everywhere."

The goal is noble but determining when it has been achieved will be exceptionally difficult. This is because of the serious problems that arise in measuring poverty. As recently as 2016, Angus Deaton wrote,

Among the most difficult and pressing problems with household surveys is the quality of the data; in some cases, the problems are severe enough to threaten even the most basic understanding of growth, poverty, and inequality (Deaton 2016, p. 1223).

He notes that the problem is especially serious in Africa, where poverty is widespread but household surveys "are often weak, often outdated, … sometimes inconsistent over time within countries, have nonmatching definitions … so that it is extremely difficult to assess progress over time, or to make comparisons of poverty or inequality between countries" (Deaton 2016, p.1224).

Even when survey data are of good quality, the steps required to arrive at reliable and valid measures of poverty are sufficiently intricate that researchers may draw very different conclusions from the same data. A dramatic illustration of this is found in the case of Rwanda, whose GDP has grown by xxx% per year over the past decade. According to the National Institute of Statistics of Rwanda (NISR), the proportion of Rwandans in poverty fell from 45% in 2011 to 39% in 2014 and 38% in 2017, although the drop over the latter period was not statistically significant (NISR 2018). A recent study that appeared in the Review of African Political Economy used the same survey data to conclude that the poverty rate in Rwanda rose from 52% in 2014 to 58% in 2017 and was higher in 2017 than in 2001 (ROAPC 2019).

This provides support for the contention of Pogge and Wisor (2016 pp. 4-5) that "The result of various internal challenges … the methods used for setting poverty lines and calculating individual achievements against this standard … is that the scope, distribution, and trend of poverty within and across countries varies greatly depending on which assumptions are used." If poverty reduction were closely linked to GDP growth – with a stable income elasticity of poverty – then it would be practicable to predict poverty rates based on anticipated economic growth as done by Chandy et al. (2013), and this may be reasonable over long intervals (Dollar and Kraay 2001), but the relationship is not close enough in the short-run for this to be compelling. We thus need to rely on survey data to track the evolution of poverty over time.

In this paper we examine the sensitivity of measures of poverty to the choices made by analysts about these "internal" decisions, mainly using survey data from Rwanda, but with appropriate reference to the experience of other countries. We focus on four (of the many possible) "internal" issues where the assumptions made by the analyst may matter: valuing auto consumption, adjusting for prices over time and space, specifying adult equivalents, and establishing a poverty line. Of these, the most difficult is getting the prices right.

In what follows we set out each issue briefly and indicate our preliminary findings.

2. Valuing autoconsumption

In poor countries, a significant amount of household consumption comes from home production. Since this autoconsumption is not sold, the question arises of how best to value it. Many surveys ask households how much they could get if they were to sell the good or service in the marketplace. A potential difficulty here is that there is some evidence that households tend to understate the selling price. Also, it may not be ideal to value (say) sweet potatoes at the buying price for some households (if they buy the good), and at a different price (the selling price, if they consume their own production) for others.

One approach would be to value autoconsumption using some cluster-level median household- reported price, which should reduce idiosyncratic variation in prices, but at the risk of overlooking variations in quality. Another possibility would be to collect prices in a separate survey, with the express purpose of using them to value household consumption. And a third possibility, used in Rwanda, is to use prices collected by the NISR for constructing the consumer price index.

Using data from 2017, we found that there is no systematic difference between the median prices reported by households, and the prices collected in the clusters where the households reside, although the household-reported prices have a very wide variance.

3. Adjusting for prices over time and space

The household survey data needed to measure poverty are collected in different areas of the country at different times of the year. There is geographical and temporal variation in prices, and the value of household spending needs to be deflated in order to take this variation into account.

Some, but certainly not all, household surveys collect information on both values and quantities of items consumed, and so one can infer unit costs, which are similar to prices. In principle these could be used to deflate household expenditure, although in practice they may not be very accurate, given variations in the quality of products. In a few cases, a complementary survey collects price data from the areas in which households are surveyed, but the quality of these data often suffers, since the enumerators may not be skilled at correcting for differences in quality or type. Sometimes the price data collected for the consumer price index (CPI) are used, although in some countries these numbers are only available for urban areas, which makes them poor guides to the prices faced by poor, largely rural, households.

Given a set of prices, it is possible to deflate consumption for each household, as done by Kenya (2007) and recommended by Deaton and Zaidi (2002). Since these indexes in effect use the consumption weights in the end period, they are Paasche indexes, and tend to understate inflation.

It appears to be more common to create a “poverty price index” that uses commodity weights that reflect the experience of households close to the poverty line – perhaps the poorest (say) 40% of the population (as in Rwanda 2018), or those whose consumption is expected to be close to the poverty line (e.g. Vietnam). More often than not these indexes use the end weights, and so are also Paasche indexes, although if there are panel data, more satisfactory indexes may be applied (see Haughton and Khandker 2009).

The conflicting conclusions about the evolution of poverty in Rwanda arise in large part because of differences in the handling of prices: the NISR (2018) creates a poor-price index using detailed price data, while ROAPE (2019) uses the published consumer price index data. A spike in (relative) food prices, which coincided almost exactly with the survey period of the 2016/17 household survey (EICV5), makes the issue more complex.

4. Specifying adult equivalents

Consumption per households is clearly not an adequate measure of wellbeing, because households differ in size and composition. But what should be used instead?

The commonest solution is to use consumption per capita, which gives a measure that is easy to calculate and explain. However, it does not take into account the different needs (food, clothing, etc.) of family members of different ages and gender; nor does it make any adjustment for economies of scale.

There is a large literature on how best to choose an adult equivalence scale. In the U.S. and OECD, some researchers divide by the square of the number of household members. For a while the "OECD scale" was popular; it may be written as AE = 1 + 0.7(A-1) + 0.5C, where AE is adult equivalents, A is the number of adults, and C is the number of children. One might also use a general scale, along the lines of $AE = (A + \alpha C)^{\gamma}$, where the α measures the weight of children relative to adults, and γ measures the extent of economies of scale. There is no best way to choose a scale (see Bellù and Liberati 2005a and 2005b).

Among poor households, there are probably minimal economies of scale in food consumption (which usually accounts for about two-thirds of all spending), but modest economies of scale in non-food items such as shelter. We explore the implications of using a scale that separates the two effects (our "suggested scale"), like this:

AE = (2/3) (Calorie-based equivalence scale) + (1/3) (Non-food based equivalence scale).

Various caloric scales are possible, but one that is based on FAO data that relates caloric intake to age and gender be as shown here in the top panel, shown alongside the weights currently used for all consumption by Rwanda ("NISR weights") and Uganda.

Suggested caloric weights

	<1	1-2	3-4	5-7	8-9	10-11	12-13	14-15	16-17	18-29	30-59	60+
M	0.3	0.4	0.5	0.6	0.7	0.8	1.0	1.1	1.2	1.0	1.0	0.8
F							0.9	0.9	0.9	0.8	0.8	0.7

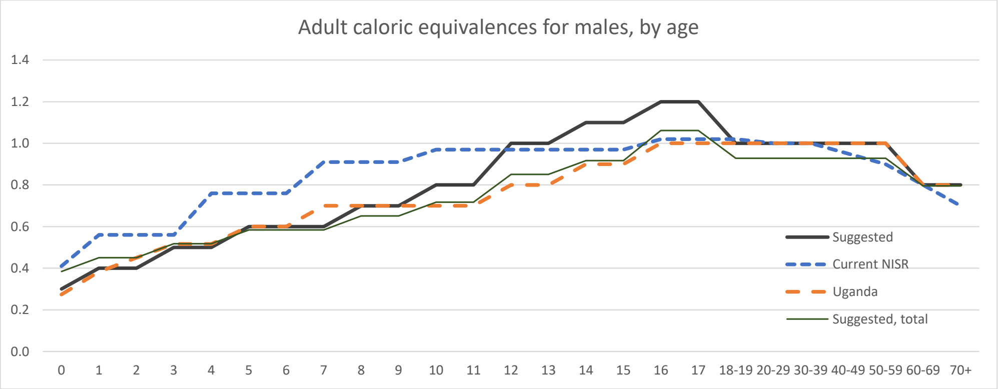
Current NISR weights

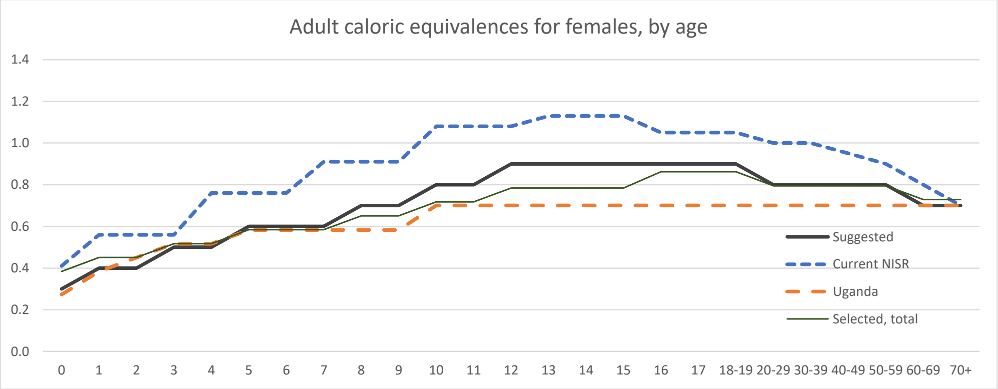
	<1	1-3	4-6	7-9	10-12	13-15	16-19	20-39	40-49	50-59	60-69	70+
M	0.41	0.56	0.76	0.91	0.97	0.97	1.02	1.00	0.95	0.90	0.80	0.70
F					1.08	1.13	1.05					

Uganda weights

	<1	1	2	3-4	5-6	7-9	10-11	12-13	14-15	16-17	18-29	30-59	60+
M	0.3	0.4	0.5	0.5	0.6	0.7	0.7	0.8	0.9	1.0	1.0	1.0	0.8
F						0.6		0.7	0.7	0.7	0.7	0.7	0.7

The choice of scale makes some difference, as the figure below shows. The suggested scale (thin black line) would give somewhat less weight to children and to women than the current NISR scale. The poverty line would also have to be adjusted, if a different scale is used, and explorations of the implications of this for the pattern of poverty are on-going.





5. Establishing a poverty line

Like many countries, Rwanda uses a cost-of-basic-needs approach to establish a poverty line. It begins with the

In setting its poverty line, Rwanda uses a cost-of-basic-needs approach, which first defines the number of calories that are needed for an adequate diet for an adult (2,500 kcals per day), and then adds a non-food component. Household consumption per adult equivalent is then compared to this poverty line, where the adult equivalents are defined as follows:

<1	1-3	4-6	7-9	10-12	13-15	16-19	20-39	40-49	50-59	60-69	70+
0.41	0.56	0.76	0.91	M: 0.97 F: 1.08	M: 0.97 F: 1.13	M: 1.02 F: 1.05	1.00	0.95	0.90	0.80	0.70

The main issue of controversy here is whether the food threshold of 2,500 kcals/day is appropriate. For instance, a recent World Bank mission to Rwanda (May 4-12, 2016) recommended lowering the calorie threshold to 2,150 kcals per day, presumably on the grounds that this is more in line with practice elsewhere, particularly in the region.

The WHO (1985) finds that the daily energy requirement for a moderately active man (1.6 times basal metabolic rate) weighing 65 kg is about 2,700 kcals per adult, or possibly slightly less in the tropics. The FAO sets a lower limit for those who are undernourished – about 2,270 kcals. Given the current age and demographic structure, and the NISR weights shown above, this implies an average consumption of about 2,066 kcals per capita, in line with practice elsewhere (e.g. 2,100 kcals in Vietnam).

Caloric thresholds, and adult equivalence scales, vary widely from country to country, making comparisons across countries difficult.

6. Conclusion

It is clear that the measurement of (monetary) policy is a technical intricate process, and different assumptions can lead to somewhat different conclusions about the pattern and changes of poverty rates.

Moreover, there are other very different ways to measure poverty, including versions of the increasingly popular multidimensional poverty index (e.g. Wang et al. 2016; Rwanda 2018b), or indexes based on a stripped-down set of indicators (e.g. Pogge and Wisor 2016). And surveys themselves are rarely very accurate (see Beegle et al. 2016).

Given how hard it is to measure poverty, it should be no surprise that it will be difficult to determine when we have indeed achieved a world without poverty, but the wider use of best practices in measuring autoconsumption, deflating, applying adult equivalences, and setting poverty lines would certainly be welcome.

References

1. Beegle, Kathleen, Luc Christiaensen, Andrew Dabalen, and Isis Gaddis. 2016. Poverty in a Rising Africa. World Bank, Washington DC.
2. Bellù, Lorenzo Giovanni, and Paolo Liberati. 2005a. *Equivalence Scales: Objective Methods*, FAO and UN, Rome.
3. Bellù, Lorenzo Giovanni, and Paolo Liberati. 2005a. *Equivalence Scales: Subjective Methods*, FAO and UN, Rome.
4. Chandy, Laurence, Natasha Ledlie, and Veronika Penciakova. 2013. *The Final Countdown: Prospects for Ending Extreme Poverty by 2030*. Brookings Institute, Washington DC.
5. Deaton, Angus. 2016. Measuring and Understanding Behavior, Welfare, and Poverty. *American Economic Review*, 106(6): 1221-1243.
6. Deaton, Angus and Salman Zaidi. 2002. Guidelines for Constructing Consumption Aggregates for Welfare Analysis. LSMS Working Paper No. 135, World Bank, Washington DC.
7. Dollar, David, and Aart Kraay. 2001. Growth is Good for the Poor. World Bank, Washington DC.
8. Granzow, Sandra. 2000. Our Dream : A World Free of Poverty. World Bank. https://openknowledge.worldbank.org/handle/10986/2411 License: CC BY 3.0 IGO.
9. Kenya National Bureau of Statistics. 2007. *Basic Report on Well-Being in Kenya*. Nairobi.
10. Haughton, Jonathan, and Shahidur Khandker. 2009. *Handbook of Poverty and Inequality*, World Bank, Washington DC.
11. NISR (National Institute of Statistics of Rwanda), 2018. *Rwanda Poverty Profile Report 2016/2017*, Kigali.
12. NISR (National Institute of Statistics of Rwanda), 2018b. *Rwanda Multidimensional Poverty Report 2016/2017*, Kigali.
13. Pogge, Thomas, and Scott Wisor. 2016. Measuring Poverty: A Proposal.
14. UN: Transforming our world: the 2030 Agenda for Sustainable Development. 2015. https://sustainabledevelopment.un.org/post2015/transformingourworld
15. UNICEF. 2018. *Progress for Every Child in the SDG Era*. New York.
16. Wang, Xiaolin, Hexia Feng, Qingjie Xia, and Sabina Alkire. 2016. On the Relationship between Income Poverty and Multidimensional Poverty in China. OPHI Working Paper No. 101, University of Oxford.
17. World Health Organization. 1985. *Energy and Protein requirements,* WHO Technical Report Series 7, WHO, Geneva.

Correlation between salaries and number of employees in manufacturing sector with the presence of outliers

Mohd Zulhairi Omar, Syed Samshiee Syed Abd Kadar, Umi Salmah Zainol
Department of Statistics Malaysia

Abstract

This study is to focus on correlation between salaries and number of employees in manufacturing sector in Malaysia from June 2017 to June 2018. This study used the data from Monthly Manufacturing Report that be published by Department of Statistics Malaysia (DOSM) on June 2018 that covers 155 out of a total of 259 industries in the Manufacturing Sector (based on the Malaysia Standard Industrial Classification, 2008). This study involves on applying the classical and robust r measure between salaries and number of employees. Apart of classical correlation coefficient, robust correlation coefficient can be used to show the relationship because it is less affected by outliers. The result of the study indicates that there is a linear relationship between the salaries and number of workers with the presence of outliers.

Keywords

Outliers, relationship, salary, employees, manufacturing, Monthly Manufacturing

1. Introduction

Manufacturing is one of the main contribution sectors for Malaysia's economy. It contributes 23.6% (RM 70.8 billion) of gross domestic product (GDP) for second quarter of 2018, increase by 0.1% compare to second quarter of 2017 (RM 67.5 billion). This growth is contributes by manufacturing of electric and electronic products, petroleum, chemical, rubber, plastic, transport equipment and other repair products.

Based on Monthly Manufacturing Statistics Malaysia, June 2018, number of employees that engaged in the Manufacturing sector in June 2018 was 1,070,776 persons, an increase of 2.2% or 22,556 persons as compared to 1,048,220 persons in June 2017. Total of salaries that be paid in June 2018 is RM 3.86 billion as compared to RM 3.50 billion, an increase of 10.2% or RM357.2 million. The result shows that as more employees were appointed to the sector, more money will be spent for their salaries.

The relationship between total of salaries and number of employees has been discussed often by many researchers such as Ho and Yap, 2001 and Nailah et. al, 2012. However, it is mainly focus on productivity rather than number of employee and involves all the economic sectors. Hence, this paper aims to produce an empirical evidence for showing the relationship between number of employees and salaries in manufacturing sectors from June 2017 to June 2018 without excluding the outlier value.

2. Methodology

In this study, the measure of correlation coefficient was computed on a real data set by using Pearson's correlation coefficient. Let (x_1,y_1), …, (x_n,y_n) be n observations from a bivariate normal distribution with parameters (μ_x,μ_y,σ^2_x, σ^2_y), where μ_x and σ^2_x are the mean and variance of x, μ_y and σ^2_y are the mean and variance of y.

The Pearson's correlation coefficient r is the define as:

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(X_i - \bar{X})\ \sum_{i=1}^{n}(Y_i - \bar{Y})}$$

Where: $\bar{X}=\frac{\sum_{i=1}^{n} X_i}{n}$ $\bar{Y}=\frac{\sum_{i=1}^{n} Y_i}{n}$

For robust correlation coefficient measure, this study used the robust Median Absolute Deviation (MAD) introduced by Gideon (1987, 2007) to calculate correlation coefficient. This method is very general and provides a robust estimation procedure in basic correlation analysis and in every advanced statistical procedure.

$$\left(\frac{1}{2}\left[med\left(\frac{x_i'}{MAD_x}+\frac{y_i'}{MAD_y}\right) - med\left(\frac{x_i'}{MAD_x}+\frac{y_i'}{MAD_y}\right)\right]\right)$$

Where:

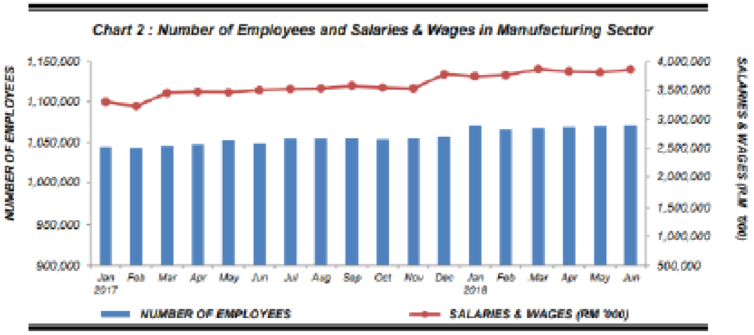
$$x_i' = x_i - med(x)$$
$$y_i' = y_i - med(y)$$
$$MAD_x = med\,(x_i - med(x))$$
$$MAD_y = med\,(y_i - med(y))$$

This study used the data from Monthly Manufacturing Report that be published by Department of Statistics Malaysia (DOSM) on June 2018. This survey covers 155 out of a total of 259 industries in the Manufacturing Sector (based on the Malaysia Standard Industrial Classification, 2008).

3. Result

The variables that be selected for this study is between the salaries and wages and number of employees from June 2017 to June 2018, as can be seen in Figure 1 below:

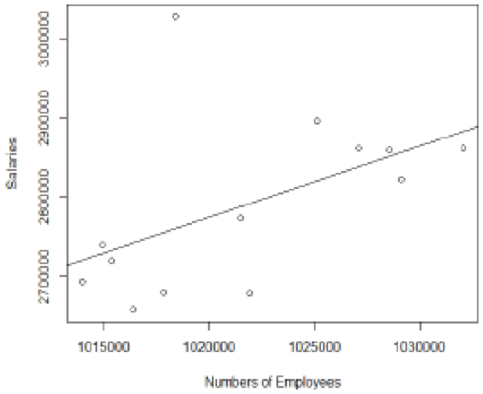
Figure 1: Number of Employees and Salaries & Wages in Manufacturing Sector



Source: Monthly Manufacturing Statistics Malaysia, June 2018

The reports only inform the readers about the trends of number of employees and salaries and wages respectively. It did not discuss about the relationship between them as it can briefly seen that there is likely relationships occur between the variables. However, the value of salaries did not move consistently with the numbers of employees as at December 2017, the total of salaries has a major increase while the number of employees has a slightly change compare to previous month.

Graph 1: Correlation between Salaries and Number of Employees



Based on Graph 1, it shows that there is a point where is defiantly far from others which is outliers. The value of correlation coefficient measure for these data was **0.51**, generated from **R 3.1.2 software**. It shows that there is moderate relationship between salary and wages and number of employees. However, the result of correlation coefficient measure above was affected by the presence of outliers. Hence, by using the robust correlation coefficient, the result obtained was **0.73** which fairly close to 1. It showed that there is a strong linear relationship between these two variables with the presence of outliers.

4. Discussion and Conclusion

The result of this study shows that there is a linear relationship between the data variables, in this case; the number of employees and value of salaries. By using Pearson's correlation coefficient measure, the value is not very significant since there are outliers in the data.

However, if the data is applied with robust correlation coefficient measure, the result shows that there is strong linear relationship between each other with the presence of outliers. The evidence from this study suggests that data is valid with the presence of outliers. The number of workers has a linear relationship between the salaries with the presence of outliers.

References

1. Barnett, V and Lewis, T 1984, *Outliers in Statistical Data*, John Wiley and Sons, New York
2. Department Of Statistics Malaysia 2018, *Monthly Manufacturing Statistics: June 2018*, viewed 2 December 2018, http:// newss.statistics.gov.my/newss-portalx
3. Gideon, R. A. (2007), The Correlation Coefficients, *Journal of Modern Applied Statistical Methods*, 6: 517-529

The impact of tourism industry on consumer price index: the case of Melaka

Syafawati Abdul Refai, Mohamad Hamizan Abdullah
Department of Statistics Malaysia

Abstract

The tourism industry is a dynamic industry where it is one of the fastest growing services industries in Malaysia's economy. The tourism industry is the third major contributor to the Malaysia's Gross Domestic Product (GDP) of which in 2017 it contributed 14.9 per cent to the GDP. Since the launched of Visit Malaysia Year campaign in 1990, the tourism industry has thrive vigorously over the years as a backbone of the services sector. This article presents the relationship between tourist arrival and Consumer Price Index (CPI) to observe the impact of tourism industry on Melaka's CPI. Simple Linear Regression was applied to examine the relationship between both variables. Finding indicates that there exist a positive relationship between tourist arrival and CPI. The tourism industry in Melaka need to undergo several phases of transformation and development through plans and programs developed by the government in order to remain sustainable and robust.

Keywords

tourism industry; tourist arrival; consumer price index

1. Introduction

Over the decade, Malaysia's tourism industry has experienced dynamic growth and rapid development to become one of the fastest growing economy sector in Malaysia. As Malaysia move from industrialized economy to services driven economy, the tourism industry plays an important role as one of the main contributor to the country's economy. In order to sustain growth and face challenges particularly competition among other countries, the Malaysia government has implemented the Malaysia Tourism Transformation Plan 2020 and Tourism Malaysia Integrated Promotion Plan (2018-2020) with the vision to make Malaysia's tourism industry a primary source of national revenue and a prime contributor to the socio-economic development of the nation.

Tourism industry continues to be the prime mover of the services sector growth where its share of Gross Value Added of Tourism Industries (GVATI) to Malaysia's GDP increased to 14.9 per cent in 2017 as against 14.8 per cent in 2016. GVATI value in 2017 recorded an increase of 10.3 per cent to RM201.4 billion as compared to RM182.6 billion in 2016. The Malaysia's tourism industry remain robust where in 2017 the Tourism Direct Gross Domestic Product (TDGDP) contributed a share of 6.1 per cent to Malaysia's GDP. The value of TDGDP steadily increased by 7.8 per cent to registered RM82.6 billion in 2017 as compared to RM76.6 billion in 2016.

The Malaysia's tourism industry remain resilient largely attributed to the active participation of both the public and private sector. In 2017, a total of 25.95 million tourist arrival recorded with a total revenue of RM82.1 billion. The number of tourist arrival is expected to increase with concerted efforts from the government towards realising its target to achieve 36 million tourist arrival and RM168 billion in receipts by year 2020. The diversification in tourism industry is a catalyst to other economic areas such as retail trade, food and beverage services and accommodation. The retail trade industry contributed 44.8 per cent to the total GVATI in 2017 while food and beverage and accommodation both contributed 16.3 per cent and 12.8 per cent respectively to GVATI in the same year. The tourism industry impact on the economic growth is felt mostly by foreign currency exchange, labour and employment as well as wages. In 2017, the tourism industry has employed 3.4 million persons and contributed 23.2 per cent to the Malaysia's total employment.

Melaka's tourism industry has flourished tremendously over the years since being recognized as UNESCO World Heritage Site in 2008. The state government is actively promoting and developing the tourism industry towards enhancing its contribution to the services sector in particular, and the state economy in general. In 2017, a total of 16.8 million tourist arrival recorded in Melaka, an increase of 3.15 per cent as compared to 16.3 million tourist arrival in 2016. Melaka continues to show positive growth in the tourism industry, where the total number of foreign tourist expanded by 3.15 per cent to registered 5.06 million foreign tourist in 2016 as compared to 4.47 million foreign tourist in the previous year. In conjunction with Visit Melaka Year 2019, the state government is focusing more on improving services in the state's tourism industry as well as developing new tourism products in an effort to achieve 20 million tourist in 2019.

As tourism activities involve both the consumption and purchase of goods and services, the impact of those activities would be reflected in the various sectors of the national economy. The main purpose of this study is to develop an understanding of the impact of tourism industry in consumer price index in Melaka. However due to practical constraint, this paper cannot provide comprehensive review of the tourism industry for Malaysia.

2. Methodology

Data

The objective of this study is to determine the impact of tourism industry on CPI for the case of Melaka. The monthly time series data on total number of tourist arrival and CPI for Melaka over the period of 2013 to 2017 are utilised in this study. The data of the total number of tourist arrival was obtained from Melaka Tourism Board (MTB), while the data on CPI was obtained from Department of Statistics Malaysia (DOSM). The variables used in this study are symbolized and describes as follows:

Tourist : Total number of tourist arrival to Melaka
CPI : Consumer price index

Methodology

This section described briefly about the statistical techniques applied to analyse data collected from DOSM and MTB. Two methods were used in this study; Simple Linear Regression Model and Stepwise Regression Analysis.

a) Simple Linear Regression Model (SLR)

Regression analysis is a statistical methodology that attempts to explore and model the relationship between two continuous variables. SLR is a model with single regressor x that has a relationship with a response y that is a straight line. The SLR model can be expressed as:

$$y = \beta 0 + \beta 1 x + \varepsilon$$

where x denotes the independent variable (tourist arrival); y is the dependent variable (CPI) and ε is a random error.

b) Stepwise Regression Model

Stepwise regression is a semi-automated process of building a model by successively adding or removing variables based solely on the t-statistics of their estimated coefficients. In order to use the stepwise regression, simple (pair-wise) correlation coefficient and partial correlation coefficient between y and each of x variables need to be calculated.

The simple correlation coefficient between two variables, y and x, is simply the ratio between their covariance and the product of their respective standard deviation, which is:

$$r_{yx} = \frac{\Sigma yx}{\sqrt{\Sigma y^2}\sqrt{\Sigma x^2}}$$

When the correlation coefficient between y and x is computed by first eliminating the effect of all other variables, it is called partial correlation coefficient. It is computed as follow:

$$r_{yx1.x_2} = \frac{r_{yx1} - r_{yx2}r_{x1x2}}{\sqrt{(1-r^2_{yx_2})(1-r^2_{x_1x_2})}}$$

The statistical test and data analysis were done through SPSS and Microsoft Excel.

3. Result

The relationship between tourism industry and CPI is obtained by applying Simple Linear Regression to monthly data of Melaka's tourist arrival and CPI for the year 2013 – 2017. Table 1 shows that there is a low degree of correlation between the two variables (R = 0.48). Furthermore, only 23 per cent (R2 = 0.230) of the variation in CPI can explained a linear relationship with the number of tourist arrival as indicated in Table 1.

Table 1: Simple Linear Regression Test Model Summary, Tourist Arrival and CPI

Model	R	R2	Adjusted R2	Standard Error of the Estimate
1	0.480[a]	0.230	0.217	3.8957

Table 2 shows the coefficient table and indicates the value of beta (standardized and unstandardized) for the two variables. Results indicate that there is a significant relationship between tourist arrival and CPI where the *p-value* = 0.000 which is less than α = 0.05. From Table 2, it further proved that the relationship between tourist arrival and CPI is at a low degree of correlation where β = 1.014E-5

Table 2: Coefficients[a] of Simple Linear Regression Model, Tourist Arrival and CPI

Model		Unstandardized Coefficien ts		Standardized Coefficients	t	Sig
		B	Std. Error			
1	Constant	99.486	3.211		30.988	0.000
	Tourist	1.014E-5	0.000	0.480	4.165	0.000

Note: 'a' denote dependent variable: CPI

CPI is used to measures the weighted average of prices of a basket of consumer goods and services. Those goods and services are broken into 12 main groups: food and non-alcoholic beverages; alcoholic beverages and tobacco; clothing and footwear; housing, water, electricity, gas and other fuels;

furnishings and household equipment; health; transport; communication; recreation services and culture; education; restaurants and hotels; and miscellaneous goods and services. To understand further the impact of tourism industry on CPI, a simple linear regression analysis is applied to determine the relationship between tourist arrival and 9 selected main groups of CPI. Based on Table 3, all main groups has a significant relationship with tourist arrival except transport. However, the level of correlation between all 8 main groups and tourist arrival is very weak where the value of unstandardized beta is less than 0.01.

Table 3: Simple Linear Regression Test Model Summary, Tourist Arrival and 9 Main Groups of CPI

Main Group	Unstandardized Coefficients		Standardized Coefficients	t	Sig
Food and nonalcoholic beverages	1.546E-5	0.000	0.453	3.866	0.000
Alcoholic beverages and tobacco	5.211E-5	0.000	0.466	4.014	0.000
Clothing and footwear	-6.325E-6	0.000	-0.523	-4.675	0.000
Health	8.861E-6	0.000	0.458	3.925	0.000
Transport	2.731E-6	0.000	0.120	0.924	0.360
Communication	4.508E-6	0.000	0.423	3.552	0.001
Recreation services and culture	5.250E-6	0.000	0.430	3.632	0.001
Restaurants and hotel	9.806E-6	0.000	0.450	3.836	0.000
Miscellaneous goods and services	8.367E-6	0.000	0.422	4.544	0.001

A stepwise regression is used to determine the main group that is most affected by the tourism industry. Based on Table 3, only 8 main groups has a significant relationship with tourist arrival thus were selected for stepwise regression analysis. Table 4 shows that the group Clothing and Footwear is statistically significant and moderately correlated with *p-value* = 0.000 and variance inflation factor (VIF) = 1.0. The other main groups were automatically removed by the stepwise regression due to multicollinearity which will increases the standard errors of the variables.

Table 4: Coefficients[a] of Stepwise Regression, Tourist Arrival and 8 Main Group of CPI

Model	Unstandardized Coefficient		Standardized Coeffic ient	t	Sig.	Correlations			Collinearity Statistics	
	Beta	Std. Error				Zero orde r	Partial	Part	Tolerance	VIF
Constant	5169432.766	827380.33		6.248	0.000					
Clothing and footwear	-43273.737	9255.686	-0.523	4.675	0.000	-0.523	-0.523	-0.523	1.000	1.000

4. Discussion and Conclusion

Discussion

Based on the result obtained in the analysis done with SLR, there exist a positive relationship between tourist arrival and CPI in Melaka. This shows that with the increase in the total number of tourist, there will also be an increase in the CPI. One of the factors that could lead to the positive relationship is with the higher number of tourist arrival, the demand for tourism goods and services will also increase which could lead to an increase in the price of goods and services. However, the impact of tourist arrival on the CPI is very minimal where only 23 per cent of CPI is affected by the tourism industry. This may be due to the CPI's basket of goods does not comprehensively covers all products and services in the market particularly tourism products. The CPI's baskets of goods only covers selected product and services based on the findings from Household Expenditure Survey.

This could further be proven by Table 3 where the correlation value for selected main groups is less than 0.01. This may be attributed by smaller number of foreign tourist arrival where in 2016 only 32 per cent of foreign tourist arrival recorded in Melaka, which is less than half compared to domestic tourist arrival at 68 per cent. Thus there is a minimal effect on price of goods and services due to tourism activity or foreign currency exchange. Furthermore, high demand from domestic tourism does not affect the price of goods and services.

Although the overall trend of the tourist arrival and main group of CPI shows a positive relationship, there also exist a negative relationship between tourist arrival and the Clothing and Footwear. Based on Table 4, the Beta value for both standardized and unstandardized coefficient prove the negative relationship between both variables. This imply that when the tourist arrival increases, the index for Clothing and Footwear group decreases. One of the possible factors is sellers tend to reduce the price of their products in order to attract more buyer.

The tourism is a dynamic industry, where there are many factors that affect the industry such as facilities and product and services offered by both the public and private sector. Even though price for goods and services is expected to increase particularly in a state with high tourism activity, the findings in this study shows otherwise. The impact of tourism industry to the Melaka's CPI are unsubstantial due to the majority demand for goods and services in the tourism industry are from domestic tourist.

Conclusion

This study set out to examine empirically the impact of tourism industry to the CPI in the case of Melaka by employing Simple Linear Regression and Stepwise Regression. This study has found that generally there exist a positive relationship between the two variables.

The result of this study support the idea that tourism industry indirectly increase the CPI level in Melaka. However, it is unfortunate that this study was limited only to Melaka thus makes these findings less generalizable to understand the effect of tourism industry on Malaysia's CPI. Therefore, further study is necessary in order to understand the impact of tourism activity on Malaysia's CPI.

As the state of Melaka is focusing on making the services sector as the main driver of the economic growth, the tourism industry has to remain sustainable and competitive in order to face the dynamic changes and challenges in the industry. As tourism generates high multiplier effects across many sectors, more coordinated efforts must be taken to mobilise and channel resources to upgrade the requisite tourism infrastructure and facilities as well as developing more innovative tourism products and services.

References

1. Department of Statistics Malaysia. (2018). Tourism Satellite Account, December 2017.
2. Malaysia Tourism Promotion Board. (2018). Tourism Malaysia Integrated Promotion Plan, 20182020.
3. MAMPU. (2018). Statistik Ketibaan Pelancong Mengikut Negara Asal dan Benua ke Negeri Melaka Termasuk Peningkatan Peratus Dari Tahun 2014 sehingga 2016. Retrieved 21 December 2018, from http://www.data.gov.my/data/ms_MY/dataset/statistic-pelancong-mengikut-negaraasal-ke-negeri-melaka/resource/380a0340-43b5-476a-b2cf-2dd08cb3a0e3
4. Ministry of Tourism & Culture of Malaysia. (2018). Malaysia Tourism Statistics, 2017. Retrieved 20 December 2018, from https://www.tourism.gov.my/statistics

Underemployment: a review of methodology

Syafawati Abdul Refai, Sharuddin Shafie
Department of Statistics, Malaysia

Abstract

Malaysia's labour market has already reached full employment with annual unemployment rate below 4.0 per cent. However, low unemployment rate is inadequate to reflect as a good labour market in an economy. One of the key indicators that measure a good welfare in a labour market is low underemployment rate which implies that the economy has met the need for employment in the country. A high underemployment rate will lead to a setback on government's vision to achieve a high income nation by 2020. The purpose of this study is to identify the most suitable method for underemployment within graduates in Malaysia's labour market by using time series data from Labour Force Survey and Salary and Wages Survey for the period 2010 to 2017. The methodology used in this study is based on Surveys on Economically Active Population: Employment, Unemployment and Underemployment: An International Labour Organization (ILO) Manual on Concepts and Method (Geneva, 1990). Findings of this study suggest that the best method to measure underemployment within graduates is Income-Related Inadequate Employment and thus can be use by the government for policy formulation and monitoring the national economic performance and social development.

Keywords

Underemployment; inadequate employment; unemployment; labour market

1. Introduction

The structure of the Malaysian economy has undergone rapid changes since independence in 1957. This has been the result of deliberate economic policies developed and implemented by the government to meet the needs and circumstances of each of its development phases. Malaysia's economy continues to expand and is currently the 3rd largest economy in Southeast Asia with an annual Gross Domestic Product (GDP) growth rate of 5.9 per cent in 2017. Due to buoyant economic conditions, the Malaysia's labour market underwent equally significant transformation. In 2017, the labour force comprised 15.0 million persons with the Labour Force Participation Rate (LFPR) at 68.0 per cent. Of these, 14.5 million were employed while the remaining

500,000 persons being officially unemployed with the Unemployment Rate at 3.4 per cent.

Malaysia aspires to achieve a high income nation by 2020, thus the human capital development plays a pivotal role for driving and sustaining Malaysia's economic growth. Labour market plays a vital role in human capital development particularly in determining socio-economic progress and is one of the key factors that indicates the poverty level. An efficient and effective labour market can act as an effective mechanism for contributing to economic growth and makes the economy less susceptible to shocks and retain a high standard of living. Malaysia's labour market has reached full employment[1] since 1995 with an average annual unemployment rate of 3.2 per cent. However, despite growth in employment and low unemployment rate, these indicators cannot comprehensively cater the scenario of Malaysia's labour market. One of the underlying issues on labour forces that are not captured by unemployment rate is labour underutilization or underemployment[2].

According to the 1966 International Conference of Labour Statisticians (ICLS) resolution, underemployment "exists when a person's employment is inadequate in relations to specified norms or alternative employment, account being taken of his occupational skill (training and working experience)". In general, underemployment or inadequate employment is defined as the situation when the worker is employed, but not in the desired capacity, whether in terms of compensation, hours or level of skill and experience and is willing to seek other or additional work. According to Mehran, Bescond, Hussmanns & Benes, 2008, underemployment is a broad concept reflecting underutilization of the productive capacity of the employed population. As defined in Surveys on Economically Active Population, Employment, Unemployment and Underemployment: An ILO Manual on Concepts and Method (ILO, 1990), two principle forms of underemployment are distinguished: visible underemployment, reflecting an insufficiency on the volume of employment; and invisible underemployment, characterised by low income, underutilisation of skill, low productivity and other factors.

Underemployment is not a new phenomenon particularly in developing countries. It has been appreciated since Gunnar Myrdal's critique of employment data in the Against the Stream: Critical Essays on the Economics

[1] The Organisation for Economic Co-operation and Development (OECD) defines full employment as unemployment rate below 4.0 per cent

[2] A measure of employment and labour utilization in the economy that looks at how well the labour force is being utilized in terms of skills, experience and availability to work. Labour that falls under underemployment classification includes those workers that are highly skilled but working in low paying jobs, workers that are highly skilled but work in low skill jobs and part-time workers that would prefer to be full-time. This is different from unemployment in that the individual is working but isn't working at their full capability

in the 1973. However, there is very few studies regarding underemployment in Malaysia's labour market. Due to the difficulty in measuring underemployment by conventional means, this very useful indicator are not usually reflected in national published data in some developing countries including Malaysia. Malaysia relies solely on unemployment rate as the key indicator of the well-being of the labour market and economic growth. As noted by Sengenberger 2011, the unemployment rate tends to either over-estimate or under-estimate the true magnitude of labour force underutilization or underemployment. Thus measuring and distinguishing underemployment from full employment and unemployment is crucial in determining the efficiency and effectiveness of Malaysia's labour market.

The criterion recommended by ILO in measuring underemployment is by using three approaches: a) time spent in gainful activity; b) income earned from the activity; and c) skill are underutilized. The main purpose of this study is to identify the most suitable method to measure underemployment in Malaysia's labour market. This study compares three approaches recommended by ILO based on data availability and practicality of Malaysia's labour market. The method suggested by this study can be used by the government for policy formulation and monitoring the national economic performance and social development.

2. Methodology

The objective of this study is to identify the most suitable method to measure underemployment in Malaysia's labour market. This study explores the data from Labour Force Survey (LFS) and Salaries & Wages Survey between year 2010 and 2017.

According to Resolution Concerning the Measurement of Underemployment and Inadequate Employment Situations, adopted by the Sixteenth ICLS (October 1998) there are three conditions of underemployment or inadequate employment situations that need to be consider:

a) Skill-related Inadequate Employment

Skill-related inadequate employment is characterized by inadequate utilization and mismatch of occupational skills, thus signifying poor utilization of human capital. Persons in this form of inadequate employment may be understood to include all persons in employment who during the reference period wanted or sought to change their current work situation in order to use their current occupational skills more fully, and were available to do so.

Skill-related inadequate employment is measured as the number of employees with formal education of Diploma or higher working in a semi-skilled or low-skilled occupation. Skill levels of occupation were classified based on Malaysia Standard Classification of Occupation (MASCO) 2013 as follow: i. Skilled workers: 1. Managers; 2. Professional; and 3. Technicians and associate professionals; ii. Semi-skilled workers: 4. Clerical support workers; 5. Service and sales workers; 6. Skilled agricultural, forestry, livestock and fishery workers; 7. Craft and related trade worker; and 8. Plant and machine operators and assemblers; and iii. Low-skilled workers: 9. Elementary occupations

b) Income-related Inadequate Employment

Persons in this form of inadequate employment may be understood to include all persons in employment who during the reference period wanted or sought to change their current work situation in order to increase income limited by factors such as those mentioned above, and were available to do so.

Income-related inadequate employment is measured as the total number of workers with formal education of Diploma or higher with total monthly income less than the cut-off point. The cut-off point of income-related inadequate employment for year 2017 was RM2,210 which was based on average salaries for entry level positions published by JobStreet Malaysia's 2017 Salary Report. However, the cut-off point for year 2010 to 2016 was calculated based on average annual growth rate of average salaries for entry level position which was 1.73 per cent.

c) Time-related Underemployment

Time-related underemployment exists when the hours of work of an employed person are insufficient in relation to an alternative employment situation in which the person is willing and available to engage. Person in this form of underemployment must satisfy three criteria; a) willing to work additional hours; b) available to work additional hours; and c) worked less than a threshold relating to working time. Time-related underemployment is measured as the total number of workers working less than 30 hours a week during the reference period[3].

The underemployment rate is calculated as follows:

$$UDR\ (\%) = \frac{Persons\ in\ underemployment}{Persons\ employed}\ x\ 100$$

The statistical test and data analysis were done through SPSS and Microsoft Excel.

[3] As defined in Labour Force Survey, DOSM

3. Result

The underemployment of Malaysia's labour market was measured by using three approaches: a) skill-related inadequate employment; b) income-related inadequate employment; and time-related underemployment to LFS and Salary & Wages Survey data from year 2010 to 2017. Based on Figure 1, the total number of employed persons with second job was gradually decreased since 2013, while the total number of employed persons steadily increased since 2010. The highest number of employed person with second job is 78,216 persons in Quarter 3 of 2013 with growth rate of 30.2 per cent as compared to Quarter 2/2013.

Figure 1: Total Number of Employed Persons and Total Number of Employed Person with Second Job, 2010-2017

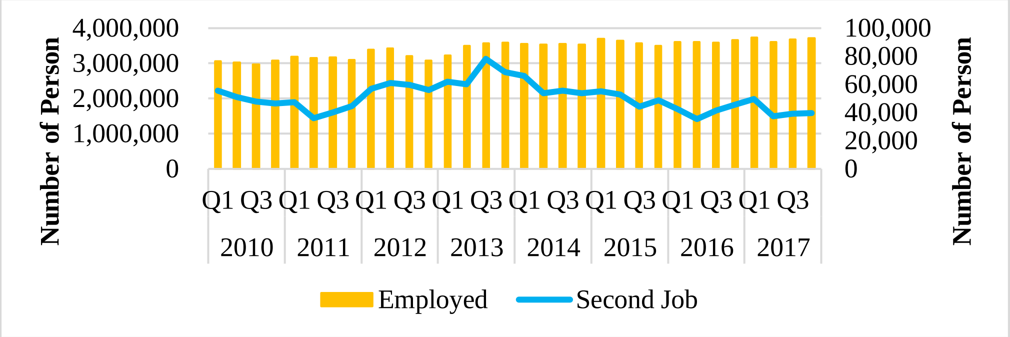
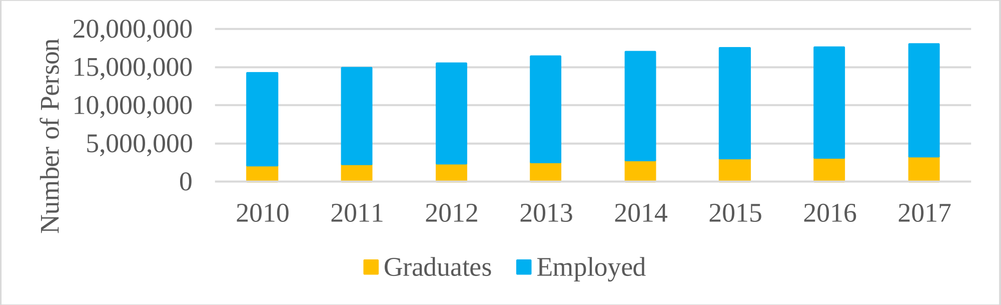


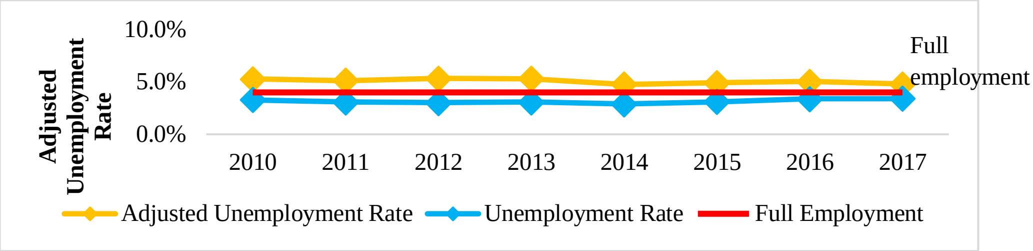
Figure 2 presents the summary statistics for graduates employed in Malaysia's labour market. Based on Figure 2, the total number of graduates actively employed in Malaysia's labour market shows an increasing trend where the highest growth rate of graduates employed was 10.6 per cent in 2014. Malaysia's labour market consists of 19.6 per cent of graduates actively employed in year 2010 to 2017 where the highest percentage of graduates employed was 22.2 per cent in 2017 as compared to 17.5 per cent in 2010.

Figure 2: Total Number of Employed Persons and Total Number of Graduates Employed, 2010-2017



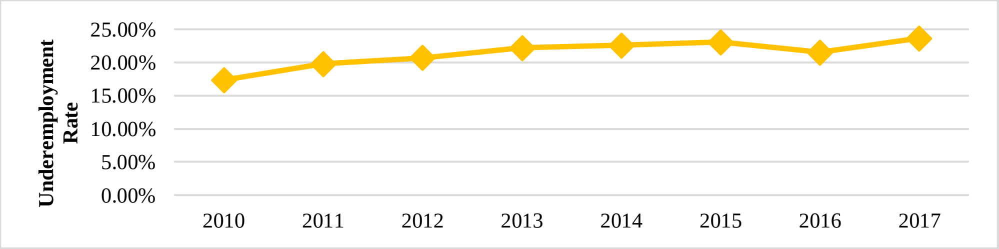
Adjusted unemployment rate was calculated by adding underemployment rate to unemployment rate. Figure 3 shows the adjusted unemployment rate between year 2010 and 2017. Based on Figure 3, the adjusted unemployment rate was between 5.3 per cent and 4.8 per cent. The average annual adjusted underemployment rate was 5.1 per cent as compared to average unemployment rate of 3.2 per cent.

Figure 3: Adjusted Unemployment Rate, 2010-2017



Skill-related inadequate employment approach was applied to determine the underemployment rate of graduates working in semi-skilled and low-skilled occupation. Figure 4 shows that the trend of graduates employed affected by the skill-related inadequate employment within graduates increased yearly with an average annual growth rate of 11.5 per cent. The highest proportion of graduates underemployed was 23.7 per cent in 2017 with a total of 774,906 graduates were underemployed.

Figure 4: Skill-Related Inadequate Employment of Graduates, 2010-2017



However, based on Figure 5 the income-related inadequate employment of graduates shows a decreasing trend since 2012 which recorded the highest underemployment rate of 22.2 per cent. The highest growth rate of income-related inadequate employment of graduates was 42.2 per cent in 2011 with the total of 429,600 graduates were underemployed as compared to 302,111 in 2010. In the year 2017, 22.2 per cent of actively employed person in Malaysia's labour market consists of graduates with a total employed graduates of 3,275,840 persons. Of these, 567,624 graduates or 17.3 per cent were underemployed as shown in Figure 6. The total number of graduates underemployed in 2017 increased by 2.5 per cent as compared to 553,744 graduates underemployed in 2016.

Figure 5: Income-Related Inadequate Employment of Graduates, 2010-2017

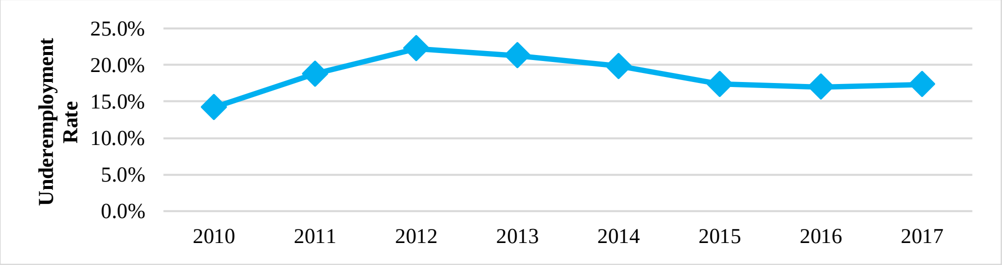


Figure 6: Graduates Employed by Income Group, 2017

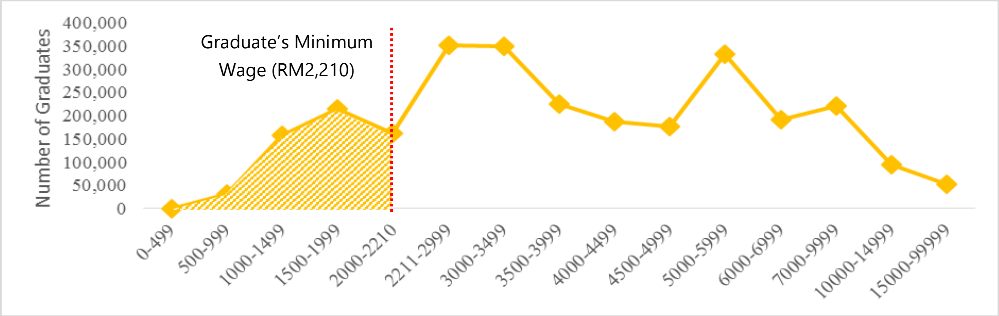
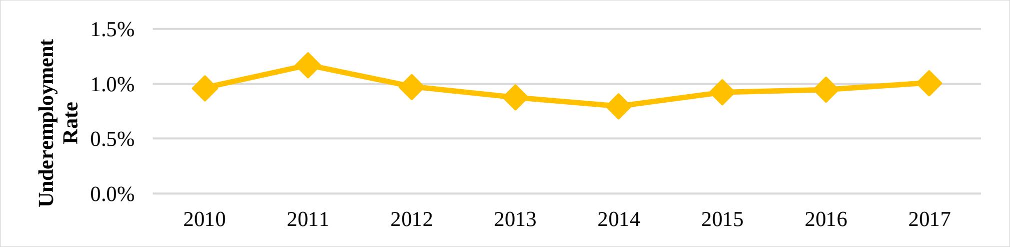


Figure 7 shows the time-related underemployment of graduates between year 2010 and 2017 with an average underemployment rate of 1.0 per cent. As shown in Figure 7, the highest recorded of time-related underemployment of graduates was in 2011 where a total of 26,783 persons or 1.2 per cent of graduates employed were underemployed. In 2017, Malaysia's labour market consists of 14,750,327 employed person and a total of 475,610 persons were working less than 30 hours a week. Of these, 228,580 persons (48.1 per cent) and 33,053 graduates were officially underemployed as defined by ILO.

Figure 7: Time-Related Underemployment of Graduates, 2010-2017



4. Discussion and Conclusion

Similarly to unemployment, underemployment or underutilization is a serious constraint to economic progress at macro and individual level. The standard labour force framework currently used in Malaysia is biased towards unemployment rate as an indicator of a healthy labour market. This systematically undervalues the degree of the unemployment problem, hence it is essential to introduce an underemployment indicator to complement the unemployment indicator in measuring underutilization of Malaysia's labour market. The adjusted unemployment rate is basically calculated by adding the total underemployed person to total unemployed person and can be used to reflect the real issue in the economy particularly in labour market. Malaysia economy has reached full employment with average unemployment rate below 4.0 per cent. However, based on this study indicates that there exists an underlying issue on the labour market where the average adjusted unemployment rate between 2010-2017 was 5.1 per cent.

This study is set out to determine the best approach in measuring underemployment of graduates in Malaysia's labour market. The current recommended approach by ILO is time-related underemployment due to the data availability and practicality in most countries. However, time-related underemployment does not comprehensively explain the underutilization of labour market particularly in a job where the normal duration of work is less than the cut-off point. In line with the government aspirations to achieve a high-income nation by 2020, an income-related inadequate employment should be considered to be as a method to measure underemployment. This is because employed person tend to seek other or additional job when the income received is inadequate to cover the cost of living.

Empirically, this study found that the variance between time-related underemployment and income-related inadequate employment is relatively low, where both approaches show a similar trend in the underemployment rate between year 2010 and 2017. Despite both approaches show a decreasing trend, it is still imperative to have a new indicator to complement the unemployment rate in measuring labour underutilisation and thus can assist in building a better understanding of the true employment situation.

Based on the findings of this study, the skill-related inadequate employment shows an increasing trend between year 2010 and 2017. However, this does not necessarily mean an increase in inadequate underutilisation as skill utilisation is extremely difficult to measure, as it involve an evaluation of the quality of the jobs against the skills of the occupant. Furthermore, there is a limitation in determining the levels of occupation of the incumbent as it relies on the enumerator's knowledge and understanding on the job description provided. Moreover, the determination of skill underutilisation threshold is a complex procedure as it may be necessary to establish different thresholds for different occupational groups or different economic activity.

The result of this study suggest that there are two approaches that can be used in measuring underemployment: a) time-related underemployment; and b) income-related inadequate employment. For the purpose of international requirement, the ILO recommended the use of time-related underemployment approach due to its practicality and data availability. However, the time-related underemployment approach does not comprehensively reflect the full extent of underutilisation of Malaysia's labour market. Though, the time-related underemployment approach is still necessary as a supplementary indicator particularly for international comparison.

Thus, this study suggest that the income-related inadequate employment is the best method in measuring underemployment or labour underutilisation as it reflect more on the Malaysia's labour market scenario and the data necessary for the indicator are available in the existing Labour Force Survey and Salary and Wages Survey. The main reason this study suggest the income-related inadequate employment as the best method to measure underemployment is that because income received by employed person is the main factor for the occupant to stay in one occupation. Furthermore, this approach can be used as a tool to evaluate government's mission in achieving high income nation by 2020. However, further research is required in establishing the suitable cut-off point of income-related inadequate employment in Malaysia's labour market.

References

1. Department of Statistics Malaysia. (2018). Labour Force Survey Report, Malaysia 2007-2017
2. Department of Statistics Malaysia. (2018). National Accounts Gross Domestic Product, 2017
3. Department of Statistics Malaysia. (2018). Salaries & Wages Survey Report, 2007-2017
4. Gunnar Myrdal. (1973). Against the Stream: Critical Essays on the Economics
5. International Labour Organization. (Geneva, 1966). Resolution Concerning Measurement and Analysis of Underemployment and Underutilisation of Manpower, P. 33-36
6. International Labour Organization. (Geneva, 1990). Surveys on Economically Active Population: Employment, Unemployment and Underemployment: An International Labour Organization Manual on Concepts and Method, P. 121-145
7. International Labour Organization. (Geneva, 1998). Resolution Concerning the Measurement of Underemployment and Inadequate Employment Situations, adopted by the Sixteenth International Conference of Labour Statisticians
8. JobStreet Malaysia. (2018). 2017 Salary Report
9. Mehran, F., Bescond, D., Hussmanns, R., & Benes, E. (2008). Beyond Unemployment: Measurement of Other Forms of Labour Underutilization. ILO: Geneva
10. Sengenberger, W. (2011). Beyond the Measurement of Unemployment and Underemployment. The Case for Extending and Amending Labour Market Statistics

Towards measuring the digital economy: initiative and challenges

Fadzilah Aini Mutaffa, Yusrina Mohd Yusoff, Malathi Ponnusamy [1]
Department of Statistics Malaysia

Abstract

The aim of this paper is to study the digital economy and to share Malaysia's experiences in development towards a digital economy. The rapid growth of digital technologies has become increasingly challenging due to the transformative impacts on economic activity. The innovations through digital technology have changed the way products are produced, consumed and traded; jobs and income are generated, and investments are financed. The businesses are taking more innovative ways to maximize and utilise digital technology to their business activities. The new technology such as the Internet of Things (IoT), Cloud Computing, Big Data Analytics (BDA), Artificial Intelligent (AI) which all of these technologies will drive towards to digital economy. The criteria for the digital economy was included how the transaction is made (digitally ordered, enabled or delivered), what is transacted (goods, services or data), and who is involved (consumer, business or government). Therefore, this paper highlighted the definition; scope and coverage; and data sources for measuring the digital economy. The current international statistical initiatives in measuring digital economy for selected countries and also for Malaysia in the context of Department of Statistics, Malaysia (DOSM) will be discussed at the last part of this study and subsequently followed by challenges and conclusions.

Keywords

Digital economy, digital technology, new technology

1. Introduction

The rapid growth of digital technologies has transformed the way operation of businesses and indirectly has affected the pattern of human life. This growth has become increasingly challenging due to the transformative impacts on economic activity. The businesses are taking more innovative ways to maximise and utilise digital technology to their business activities. The new technology such as the Internet of Things (IoT), Cloud Computing, Big Data Analytics (BDA), Artificial Intelligent (AI) was driving the digital economy (Kylasapathy, Hwa, & Mohd Zukki, 2018).

[1] The views expressed in this paper are solely the responsibility of the author and should not be interpreted as reflecting the views of Department of Statistics, Malaysia

According to UNCTAD (2017), digital economy evolving is a result of the development and use of new technologies and innovations over several decades. Characterisations of the digital economy began in the mid-1980s with mass-produced personal computers (PCs). This was followed by advanced on manufacturing computerised in the 1990s and e-commerce and offshoring in the 2000s.

The digital economy has driven economic interests and policymakers to gauge the importance and impact on the economy. Malaysia Digital Economy Corporation (MDEC) is the lead agency in driving the digital economy in Malaysia. In 2011, MDEC launched the Digital Malaysia Initiative with the goal of national Digital Transformation. It is one of the initiatives of the transformation by the Malaysian government. The Prime Minister has mandated to MDEC on 19th October 2011 during the 23rd MSC

Malaysia Implementation Council Meeting (ICM) to set up Malaysia Digital Economy Satellite Account (DESA). The establishment of DESA consists of an ICT Satellite Account (ICTSA) and other indicators to build the complete picture of the impact of digital transformation.

In accordance with this revolution, Department of Statistics, Malaysia (DOSM) took the initiative to study and compile the holistic ICT Statistics through the framework of satellite account. Referring to the Eleventh Malaysia Plan 2016-2020, the share of ICT industry to GDP was targeted at 18.2 per cent or approximately RM324.9 billion by 2020. ICTSA is a statistical framework which provides the detail transaction of supply and use of ICT products. The purpose of compiling this statistic is to present the contribution of ICT industry inclusive of e-commerce to Malaysia’s economy.

2. Malaysia’s government initiative towards the digital economy

Malaysia economy has started to shift from a resource-driven economy towards knowledge based since the 1990s as mention in Eleventh Malaysia Plan, 2016-2020. It will focus on ICT as an imperative enabler for a knowledge economy, especially in the areas of industry, infrastructure, human capital and digital inclusion. Various ICT initiatives have been implemented in the effort to transform the country into an innovative digital economy.

The initiative to develop ICTSA was documented under the DOSM Corporate Plan 2004-2009 and it was continued presented in Strategic Plan 2010-2014. Therefore, DOSM collaborated with other government agencies such as Ministry of Finance (MOF), Economic Planning Unit (EPU), Ministry of Communication & Multimedia Malaysia (KKMM) and MDEC; to set up an InterAgency Technical Working Group. All input requirement such as definitions, methodology and data sources are discussed in this committee.

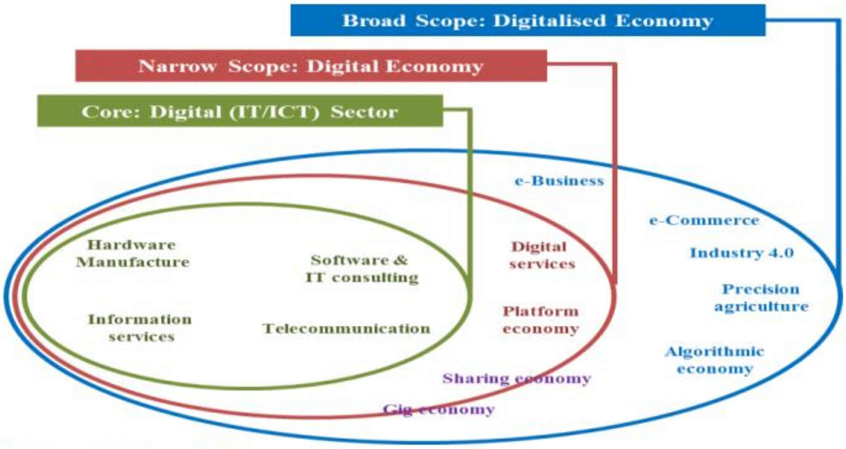
The first ICTSA report was published with limited circulation among committee members and main stakeholders in December 2012. The ICTSA publication was officially released to the public in November 2014. The compilation of ICTSA is by annual basis and the latest publication was ICTSA 2017.

3. Results

Even though the global digital economy is evolving at a rapid pace, there is a significant disparity among the development of the digital economy in different countries around the world. The digital economy now permeates countless aspects of the world economy, impacting sectors as varied as banking, retail, energy, transportation, education, publishing, media or health. Information and Communication Technologies (ICTs) are transforming the ways social interactions and personal relationships are conducted, with fixed, mobile and broadcast networks converging, and devices and objects increasingly connected to form the Internet of Things (OECD, 2015).

According to Bukht & Heeks (2017), there are three elements relating to the conceptualisation of the digital economy comprises of digital sector or ICT sector, broad scope and narrow scope. The digital sector was define using OECD definition covers International Standard Industrial Classification Revision 4 (ISIC Rev. 4). Broad scope covers e-business (ICT enabled business transactions) and its subset, e-commerce (ICT enabled external business transactions), algorithmic decision making in business, use of digitally automated technologies in manufacturing and agriculture including Industry 4.0 and precision agriculture, etc. Meanwhile, narrow scope was based on the notion of intensive and extensive applications of ICTs. Through this approach, the digital economy would represent all extensive applications of digital technologies plus the production of those digital technologies covers the digital sector, digital services, and emergent phenomena such as platform economy, gig economy and sharing economy (**Exhibit 1**).

Exhibit 1: Scoping the digital economy



Source: Bukht and Heeks (2017)

Cited in the paper Ahmad & Ribarsky (2018), the criteria for distinguishing digital transactions include how the transaction is made (digitally ordered, platform enabled or digitally delivered), what is transacted (goods, services or data), and who is involved (consumer, business or government). Digitally ordered was the transaction in goods and services that reflect e-commerce. The examples for platform enable were the sharing economy, gig economy and collaborative economies, such as Airbnb, Uber and e-Bay.

a. Mexico

Mexico measures the digital economy by calculating the value added of e-commerce. E-commerce has extended over growing sectors of the Mexican economy, which now use the internet for conducting business as a matter of course (Palacios, 2003). As a first approach, an estimate was made to quantify the gross added of e-commerce. Examples of digital economy are automatic vehicles, social networks, e-commerce, open courses online and personalized medicine. E-commerce is a process of purchase, sale or exchange of goods, services and information conducted over computer networks. For the sales of goods and services, the buyer places an order and both the price, and the terms of the transaction are negotiated through internet, email or web. The payment thru e-commerce may or may not be done online and the estimations do not include cross border transactions. The measurement of the gross value added of e-commerce was made under a supply approach related to the wholesale, retail and other services commercialisation. Total use is implicit since the SUT are balanced. The approach that has been used by Mexico to measure e-commerce is similar to Malaysia.

b. Thailand

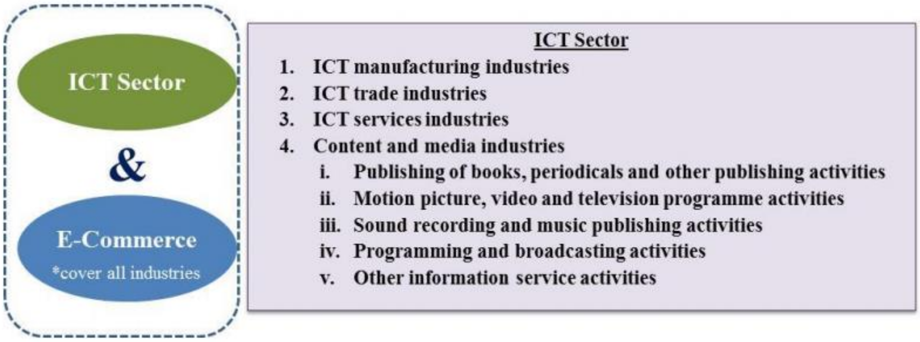
Thailand is working to improve economic growth by shifting its economy from an industry driven country to one that is high-tech driven. Focusing on this goal, the government launched the Thailand 4.0 initiative and also the Digital Thailand plan in 2016. There are six strategies of the digital economy plan in Thailand, that is building country-wide high-capacity digital infrastructure; boost the economy with digital technology; create a knowledge-driven digital society; transform into digital government; develop workforce for the digital era; and build trust and confidence in the use of digital technology (Bukht & Heeks, 2018).

Thailand measures the digital economy using ICT statistics, e-transaction statistics and ICT infrastructure statistics. ICT statistics comprising the findings from household and establishment survey on the use of ICT. E-transaction statistics encompasses e-payment, e-trading and services, e-certificate, e-health, e-filing and e-reporting, and e-tax invoice. ICT infrastructure statistics contains the services provider infrastructure for telecommunication (fixed telephone and mobile phone) and the internet.

c. Malaysia

DOSM has taken the initiative to measure towards digital economy using satellite accounts approach through ICTSA publication. At present, the coverage under ICTSA is consisting of the ICT sector and e commerce (Exhibit 2). OECD Guide to Measuring Information Society 2011 is used as a reference for the definition and classification of ICT Sector (OECD, 2011). The measurement of e-commerce value added is based on the recommendations by OECD Internet Economy Outlook 2012 (OECD, 2012).

Exhibit 2: Coverage for ICTSA



ICT Sector

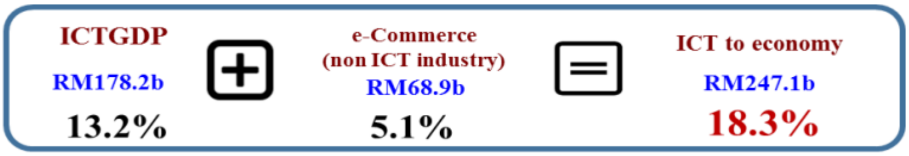
1. ICT manufacturing industries
2. ICT trade industries
3. ICT services industries
4. Content and media industries
   - i. Publishing of books, periodicals and other publishing activities
   - ii. Motion picture, video and television programme activities
   - iii. Sound recording and music publishing activities
   - iv. Programming and broadcasting activities
   - v. Other information service activities

There are two recommended approaches to measure e-commerce, which are narrow and broad. The narrow approach takes into account the value added of wholesale and retail sectors, while the broad approach includes all industries across the economy. For Malaysia's case, the broad approach was applied to measuring the value added of e-commerce. It is assumed the share of revenue from e-commerce in total revenue for each industry sector is proportional to the share of value added from e-commerce in total value added for the same industry.

4. Main Findings of ICTSA 2017

DOSM have released the ICTSA annually and the latest publication was ICTSA 2017. The contribution of ICT to the national economy recorded value of RM247.1 billion in 2017 registering a growth of 10.3 per cent (2016: 8.7%). ICT contributed 18.3 per cent to GDP comprising of Information and Communication Technology Gross Domestic Product (ICTGDP), 13.2% (RM178.2 billion) and e-commerce for non-ICT industries, 5.1% (RM68.9 billion) as shown in Exhibit 3.

Exhibit 3: Contribution of ICT to the economy



ICTGDP
RM178.2b
13.2%

➕

e-Commerce
(non ICT industry)
RM68.9b
5.1%

=

ICT to economy
RM247.1b
18.3%

5. Challenges

Even though at the moment, DOSM is measuring the digital economy through ICTSA, the initiative to compile DESA is under study, by exploring its scope, coverage and methodology. DOSM is facing various challenges in measuring the digital economy as per stated below:

a. The definition has not been internationally agreed

There are many definitions is being used to define the digital economy worldwide. Nevertheless, none of the definition has been finalized and commonly understood. Most of the countries are trying to measure it based on the availability of the data that they can obtain in the ICT field and the components of each differ by country. Therefore, we need to have a standard definition for digital economy to measure the satellite account which can be internationally comparable.

b. Classification of Digital Economy

The latest ISIC was ISIC Rev 4. The list of industries in ISIC Rev. 4 may not evolve as fast as digital technology. It is difficult to identify the major characteristics of the products under the digital economy and it has a mix of digital and non–digital components. Thus, the existent statistical data (ICT/e-Transaction/ICT Infrastructure) does not seem to reflect the digitalized economy split of e-commerce from traditional sales. DOSM is using Malaysia Standard Industrial classification (MSIC) to classify the ICT Industries and Malaysia Classification of Products by Activity (MCPA) to classify the ICT Products which is in concordance with ISIC and Central Product Classification (CPC). Therefore, Malaysia is looking forward for a comprehensive classification of industries and products in the digital economy.

c. Conceptual Framework

Since there was no standard definition for the digital economy, it is very hard to construct a framework to measure DESA. DOSM is in the process of studying the framework of DESA based on the activities undertaken in the Malaysia Economy.

d. Importance of GDP and Digital Supply-Use Table (SUT)

Digital SUT describes how the new tables expand on the current standard supply and use tables. It also identifies the various industries and product classification that have been developed for use in the account (OECD, 2018). It is understood that the GDP digital economy and SUT digital economy play a vital role to compile a comprehensive Digital Economy satellite account. In order to have the GDP on Digital Economy, the economic census questionnaires need to be improvised; which can accommodate questions on the digital economy. However, there is a problem to determine what data to collect due to lack of sufficient guidelines in the digital economy.

6. Conclusion

The momentum of ICT technology development is rapidly expanding, but the scope of the digital economy is still in the grey area and there is no standard manual, methodology or specific guidance on how to measure the digital economy. Therefore, international comparisons are difficult to measure between countries. To overcome the above-mentioned challenges, DOSM is preparing references to acquire training / courses and workshops related to the digital economy, especially in compiling satellite accounts of expertise. DOSM is in high demand to engage in specialized compilations with adequate training on satellite accounts, digital economics and ecommerce statistics to meet the needs of data from multiple users.

References

1. Ahmad, N., & Ribarsky, J. (2018). Towards a Framework for Measuring the Digital Economy. *16th Conference of the International Association of Official Statisticians (IAOS)*. Paris.
2. Bukht, R., & Heeks, R. (2017). *Defining, Conceptualising and Measuring the Digital Economy*. United Kingdom: Centre for Development Informatics Global Development Institute, SEED.
3. Bukht, R., & Heeks, R. (2018). *Digital Economy Policy: The Case Example of Thailand*. United Kingdom: Centre for Development Informatics Global Development Institute, SEED.
4. Eleventh Malaysia Plan. (2016-2020). *Driving ICT in the Knowledge Economy*. Economic Planning Unit, Prime Ministers Department.
5. Kylasapathy, P., Hwa, T., & Mohd Zukki, A. (2018). *Unlocking Malaysia's Digital Future: Opportunities, Challenges and Policy Responses. Malaysia*. Central Bank of Malaysia.
6. OECD. (2011). *OECD Guide to Measuring the Information Society*. OECD Publishing. http://dx.doi.org/10.1787/10.1787/9789264113541-en.
7. OECD. (2012). *OECD Internet Economy Outlook 2012*. OECD Publishing. http://dx.doi/org/10.1787/9789264086463-en.
8. OECD. (2015). *OECD Digital Economy Outlook*. Paris: OECD Publishing.
9. OECD. (2018). *A Proposed framework for Digital Supply-Use Tables: Meeting of the Informal Advisory Group on Measuring GDP in a Digitalised Economy.*
10. Palacios, J. L. (2003). The Development of E-Commerce in Mexico: A Business-Led Passing Boom or a Step Toward the Emergence of a Digital Economy?
11. UNCTAD. (2017). *Information Economy Report 2017: Digitalization, Trade and Development*. New York: United Nations Publication.

The Impact of Institutional Quality and Development of Human Ressources on Exports in the West African Economic and Monetary Union (WAEMU) countries

Zakaya Ramde
Professional Data Scientists Association of Burkina Faso

Abstract

This study examined the relationship between export, institutional quality and human resources development in WAEMU countries. Panel data specification were used over the period 1990-2005 and allowed the use of error correction modelling (VECM). The results reveal a long-term relationship between export, institutional quality and human resources development. In a short term perspective, they also show that human resources and institutional quality have positive significant impact for exports development in WAEMU area. Therefore, these results show the need of policies to improve human resources development and strengthen institutional quality to impact exports flows of WAEMU countries, especially the rule of law of countries, governance stability and developed existing anti-corruption strategies.

Keywords

Institutional quality; Human resources; Panel VECM; WAEMU

1. Introduction

The majority of African countries have economies that are outward-looking. For most of them, they ae sufficient of raw materials, which they export to earn foreign exchange for development in certain domains of their economies. In this way, exports are an important tool of economic policy.

In West Africa, the first export products are raw materials with low human contribution of quality (gold in Burkina Faso, cocoa in Ivory Coast, etc). In most cases, foreign direct investment underpins production in sectors that are low in quality of human resources. The Human Development Index, which aggregate three (3) index as education level of the population aged 15+, remains relatively low in these countries. This is characteristic of a less qualified workforce for the production of goods and services so a weakness in exports flows.

Costinot (2009) shows that countries with skilled human resources and quality of institutions have comparative advantages that can positively impact their export performance.

Moreover, the quality of the institutions appears as a necessary condition to a better production and better returns of their exports. Credibility of institutions are guaranties of foreign investment in the economy that can lead to economic prosperity North D (1990). On the other hand, advanced exports can deteriorate institutions in countries considered as economically weak Levchenko A (2009). Several indicators measure the quality of institutions. However, stability in governance, clear rules of law over property rights, and minimal corruption can increase the productivity of economies and increase exports. If we consider that the quality of human resources as well as stable institutional frameworks can impact the level of production of economies, and hence indirectly boost export, then we can look for questions of link that can exist between them.

2. Methodology

2.1 Data

Data use in our model are exports, human development index and institutional quality. Exports represents exports of good and services in % of GDP. Exports data from World Development Indicators. Human ressources development comes from human development index database. Institutional quality data comes from ICRG international country risk guide presented by the PRS political risk service group. It is the mediane of three (3) variables which are corruption, rule of law and governance stability.

2.2 Panel data specification

We use panel data specification for West African conomic and Monetary Union (Waemu) gives by this specification**:**

$$\begin{aligned} Exportsi{,}t &= \alpha 0 + \beta Human\ Ressources\ Developmenti{,}t \\ &\quad + \delta Institutional\ Qualityi{,}t + \eta i + \lambda t + \epsilon i{,}t \end{aligned}$$

Where η_i shows country specific effect and, temporal specific effect and $\epsilon_{i,t}$ terms of error.

The i and t represents respectively the country and time.

The econometric model process follows three (3) steps. Firstly, the homogeneity test of HSIAO is made to prove evidence of nature to perform a panel treatment approach. Looking for the results of the test, P-value of H10 is 4.818e-20 which is low than 0.05. This shows that alpha and beta are not identitiques. And then, P-value of H20 is 0.11065515 which is higher than 0.05. So we cannot reject the beta coefficients of explicatives variables are equal for different countries. The P-value of H30 is very low than 0.05. This shows existence of specific effects.

Table 1: Hsiao test

Hypothesis	F-stat	P-value
H^10	13.434427	4.818e-20**
H^20	1.4558271	0.11065515
H^30	55.35322	9.440e-30**

Note: significance at 5%.

2.3 Unit root test

In other parts, we will discuss on stationarity of variables. The analysis of stationarity of exports, human resources development and institutional quality is necessary before econometric application, indeed, to avoid wrong regressions. The tests proposed by Levin-Lin-Chu (2002) and Im-Pesaran-Shin (2003) will be use to find the order of our series in panel.

The two (2) tests shows that we cannot reject the null hypothesis of a unit root for Exports (X), Human Resources Development (HDI) and Institutional Quality (IQ) at 5% level of significance. The P-values obtained are less than 5%. So these variables are not stationary at level.

Table 2: Unit root test (in level)

	Test Im-Pesaran-Shin		Test de Levin-Lin-Chu		
	Intercept	Intercept+trend	Intercept	Intercept+trend	None
X	0.0006*	0.0260*	0.0000*	0.0000*	0.3080
HDI	1.0000	0.4326	1.0000	0.9921	1.0000
IQ	0.0069*	0.9981	0.0009*	0.7702	0.3756

*Note : * means statistically significant at 5% level of significance. Source : Our construction on Eviews.*

But, after differencing them at one time, the null hypothesis of existence of a unit root was rejected at 5% level of significance for both tests, and then we conclude at the stationarity of these variables at their first difference.

Table 3: Unit root test (in difference)

	Test Im-Pesaran-Shin		Test de Levin-Lin-Chu		
	Intercept	Intercept+trend	Intercept	Intercept+trend	None
X	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*
HDI	0.0080*	0.1577	0.0000*	0.0000*	0.0202*

IQ	0.0005*	0.0000*	0.0000*	0.0000*	0.0000*

*Note : * means statistically significant at 5% level of significance. Source : Our construction on Eviews*

And then, they are integrated at order one I(1). And then, the variables are eligible to test the long run cointegration analysis between them.

a. Trace, max-eigen and Pedroni tests of cointegration

- **Trace and max-eigen tests**

  According to Trace test, one (1) cointegrating equation at the level of 5% were accepted. The P-values from trace test and max-eigen statistic are significant at 5% level. This shows existence of a long-run relationship between X, HDI and IQ.

Table 4: Trace Test for cointegration

	Trace test	P value	Max-eigen test	P value
None	55.14	0.0000	62.48	0.0000
At most 1	9.906	0.7690	10.76	0.7049
At most 2	10.67	0.7120	10.67	0.7120

- **Predroni test**

  Pedroni (2004) is used to test cointegration. And then, as results shows, four (4) of seven (7) statistics gives by Pedroni test concluded on existence of cointegration relation between exports, human resources development and institutional quality. And then, according to Granger representation theorem, an vector error correction model (ECM) can be perform.

Table 5: Pedroni test for cointegration

Alternative hypothesis: common AR coefs. (within-dimension)				
	Statistic	P value	Weighted Statistic	P value
Panel v-Statistic	-1.079967	0.8599	-1.079967	0.8599
Panel rho-Statistic	-1.350867	0.0884	-1.350867	0.0884
Panel PP-Statistic	-2.428329	0.0076	-2.428329	0.0076
Panel ADF-Statistic	-1.908328	0.0282	-1.908328	0.0282
Alternative hypothesis: individual AR coefs. (between-dimension)				
	Statistic	P value		
Group rho-Statistic	-0.242182	0.4043		
Group PP-Statistic	-2.679863	0.0037		
Group ADF-Statistic	-1.873918	0.0305		

3. Results

Cointegration equation

Cointegration equation is given by: $\boldsymbol{X}= -\mathbf{39{,}073}\ \boldsymbol{HDI}+ \mathbf{2{,}011}\ \boldsymbol{IQ}$

On the long term, HDI have negative impact to X, exports and IQ have positive impact on X.

Table 6: Estimation of VECM model

	D(X)	D(HDI)	D(IQ)
CointEq1	-1.573054	-4.39E-05	-0.053147
	(0.09191)	(0.00029)	(0.02070)
	[-17.1157]	[-0.14945]	[-2.56806]
D(X(-1))	0.269332	0.001165	-0.004773
	(0.06964)	(0.00022)	(0.01568)
	[3.86726]	[5.24060]	[-0.30436]
D(HDI(-1))	203.9609	0.187286	6.162952
	(32.5102)	(0.10381)	(7.32048)
	[6.27375]	[1.80403]	[0.84188]
D(IQ(-1))	5.053402	-0.006198	0.430450
	(0.47897)	(0.00153)	(0.10785)
	[10.5505]	[-4.05204]	[3.99110]
C	-1.120724	0.004708	-0.014155
	(0.19647)	(0.00063)	(0.04424)
	[-5.70437]	[7.50359]	[-0.31996]
R-squared	0.811807	0.344799	0.189686
Adj. R-squared	0.803713	0.316618	0.154834
Sum sq. resids	88.84066	0.000906	4.504555
S.E. equation	0.977382	0.003121	0.220082
F-statistic	100.2933	12.23530	5.442598
Log likelihood	-134.2479	428.9286	11.85806
Akaike AIC	2.841795	-8.651604	-0.139960
Schwarz SC	2.973681	-8.519718	-0.008074
Mean dependent	-0.050254	0.005714	0.017857
S.D. dependent	2.206064	0.003775	0.239394
Determinant resid covariance (dof adj.)		2.36E-07	
Determinant resid covariance		2.01E-07	
Log likelihood		338.3399	
Akaike information criterion		-6.537548	
Schwarz criterion		-6.062758	

3.1 The long run causality

The OLS (ordinary least squares) method were used to discuss about the significance of the model. We found that the error correction term C(1) which is -1.57 is negative and significant at the level of 5% because of -17.11 statistic which is low 0.09. So there is a long-run relationship between the three (3) variables. Then, it exists an impact in long-term between exports flows, human ressources and institutional quality.

3.2 The short run causality

The wald test were conducted to determine the significance of coefficients.

- Human resources development

According to the short term causality results, the p value statistic associated to C(3) is lower than 0.05. So the coefficient is significant at the 5% level. We can consider human resources devlopment explains exports flows in short term.

- Institutional quality

As we can see, the the p value statistic associated to C(4) is lower than 0.05. So the coefficient is significant at the 5% level. We consider that institutional quality explains exports flows in short term in WAEMU countries.

Table 7: Estimation of VECM model

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-1.573054	0.091907	-17.11567	0.0000
C(2)	0.269332	0.069644	3.867262	0.0002
C(3)	203.9609	32.51020	6.273751	0.0000
C(4)	5.053402	0.478971	10.55053	0.0000
C(5)	-1.120724	0.196468	-5.704366	0.0000
R-squared	0.811807	Mean dependent var		-0.050254
Adjusted R-squared	0.803713	S.D. dependent var		2.206064
S.E. of regression	0.977382	Akaike info criterion		2.841795
Sum squared resid	88.84066	Schwarz criterion		2.973681
Log likelihood	-134.2479	Hannan-Quinn criter.		2.895140
F-statistic	100.2933	Durbin-Watson stat		2.278503
Prob(F-statistic)	0.000000			

3.3 Granger causality

According to Granger causality test, we rejected that IQ does not Granger cause HDI. We also reject HDI does not Granger cause X and also X does not Granger cause HDI were rejected. So there is a relation of causality of IQ on HDI, HDI on X and X on HDI.

Table 9: Granger causality

Null Hypothesis:	Obs	F-Statistic	Prob.
X does not Granger Cause IQ	105	0.00144	0.9698
IQ does not Granger Cause X		0.01048	0.9187
HDI does not Granger Cause IQ	105	0.78679	0.3772
IQ does not Granger Cause HDI		4.71742	0.0322
HDI does not Granger Cause X	105	5.31028	0.0232
X does not Granger Cause HDI		6.71723	0.0109

4. Discussion and Conclusion

The objective of this study was to analyze the impact of institutional quality and human resources development on exports flows in WAEMU countries. Based on data from Human Development Index, World Development Indicators and PRS Group over the period 1990-2005, an vector error-correction were build and gives two (2) results. Firstly, there is a long-term relationship between exports flows, human resources development and institutional quality. The institutional quality and human resources have positive impact on short term on exports. Secondly, institutional quality implies a better human resources development as exports too, and a qualified human resources implies exports in WAEMU areas.

References

1. North D. (1990), Institutions, Institutional Change and Economic Performance, *Cambridge University Press and Robert Donnelly Review Essay*
2. Muhammad TM and Eatzaz A (2006). Determinants of Exports in Developing Countries. Pakistan Institute of Development Economics; Institute of Development Economics (Pakistan)
3. Levchenko A (2009). Trade, Inequality and the Political Economy of Institutions.

Improving middle school students' expectations of variability in a two-dimensional context

Dan Canada
Department of Mathematics, Eastern Washington University, Cheney WA 99004 (USA)

Abstract

The purpose of this paper is to report on the emerging results of a project that used an instructional intervention designed to improve middle school students' informal expectations of variability in a two-dimensional context. Specifically, one aim of the project was to compare how students reasoned about variability to make informal inferences both before and after modelling a task physically and then via computer simulation. A simultaneous goal was to have students pursue their own additional questions, beyond the initial prompts given, that were prompted by an analysis of the data they had gathered.

Keywords

Statistics; Education; Probability; Variation; Teaching

1. Introduction

The underlying task in this project, based on work by others (e.g. Engel & Sedlmeier, 2005; Green, 1982; Piaget & Inhelder, 1975), posits raindrops just beginning to fall across a patio of sixteen square tiles in a 4 x 4 array: Where might the first sixteen drops land? In Engel and Sedlmeier's work, using falling snowflakes as a context, their "objective was to find out how children decide between random variation and a global uniform distribution of flakes" (2005, p. 169). Using a framework that considered the degree to which student responses reflected a perspective of randomness versus determinism, those researchers found evidence across a range of tasks and grade levels that students' ability to coordinate randomness and variability seems to deteriorate with age.

Of particular interest was the call by the researchers for instructional interventions that would leverage technology (such as computer simulations) to bolster gathering experimental data in a quest to develop "students' intuitions about chance variation" (Engel & Sedlmeier, 2005, p. 176). In fact, as detailed in the next section covering methodology, the intervention in the current project reflects the first four aspects of Engel's (2002) five-step procedure: Making initial conjectures or observations of a given phenomenon, developing a model for the purposes of simulation, gathering data, and comparing subsequent results to initial predictions. The fifth step, involving formal mathematical analysis, was beyond the purview of the middle school students.

2. Methodology

The main reason "falling raindrops" was used as the context for the task instead of "falling snowflakes" is because the project initially took place with 12 students in a city Tanzania, and again at a later date with 21 students in a city in Vietnam (both places where snow was generally unfamiliar). The students had some basic skills in probability and statistics: For example, they could make simple graphs of data, and talk about distributions of data in terms of centers and ranges. Also, they could discuss likelihoods and compute probabilities for simple one-stage events.

Initially, when presented with the question of "Where might the first sixteen drops land?" as described in the previous section, students made marks on a 4 x 4 grid and also wrote down why they held that view. Whole-group discussion ensued, with student opinions ranging from a more deterministic approach (i.e. expressing that each of the sixteen tiles should contain a raindrop in the center of each tile) to more of random approach (i.e. the raindrops should look like less of a discernible pattern). The nature of the discussion had similar types of thinking as reported in similar results from other researchers (Engel & Sedlmeier, 2005; Green, 1982). Some students wondered how it was possible to make any prediction since "anything can happen" or "rain can fall anywhere", while others mused about how factors like wind might influence the results.

As we transitioned to the question of "How could we model this idea?", students were very creative. Among the ideas were finding a way to "splatter" water over a grid, or other (more viscous) liquids that were easier to record a single drop. Eventually students turned to other methods like tossing coins, blocks, and even "confetti" they made from shredded newspaper (the latter actually gave a strong impression of falling snow). Some students went up to a 2nd - floor balcony to distribute their "raindrops" (many of which missed the grid entirely), tossing things out into an alley or hall, and others used the height of a desk, chair, or simply standing up in a room over a grid. We allowed for all kinds of different materials and different sizes of grids (as long as they comprised sixteen squares in a 4 x 4 array), with the only requirement being that students felt their modelling technique was "as unpredictable as rain". All of the physical experimentation was photographed and videotaped for further reflection.

Once sixteen token "raindrops" had landed somewhere on the 4 x 4 grid of their choice, we did provide uniform pages of identical 4 x 4 grids on paper where they could record their results, carefully marking on the recording paper what their physical model showed. We then hung the recording papers all around the classroom: Each paper recorded one "trial" of their successful toss of sixteen "raindrops". After having at least thirty trials recorded and up around the room (all on identically-sized recording paper grids), we then entered in a period of reflection: In particular, students were asked what they noticed, and what they wondered about.

In this phase of generating new questions to pursue, the first thing many students noticed was that none of the experimental results looked like the typical "one raindrop per tile, perfectly centered in each square" which so many had suggested beforehand. In fact, soon the observation arose that most if not all of the grids up on display were *without* a "one raindrop per tile" result (let alone the idea of being perfectly centered). This led to the obvious connection: If there *wasn't* "one raindrop per tile" on a grid, then by necessity there must be some empty squares on that grid. Students began to wonder how many empty squares were among their displayed experimental results: What was the most and least number of empty squares? What was the most number of raindrops in any given square?

As students tabulated different aspects they were interested in, based on the questions about the results they raised, the notion of likelihood came up by wondering what would happen if we repeated the whole experiment on another day? The language of a "batch" of results was used to describe how many trials were on display: For example, if there were thirty grids of experimental results, we just called it a batch of thirty "trials". If, at another time, we generated a new batch of thirty results, how would students think the new batch would compare to the initial batch? As an example of a specific observation, students saw in their initial batch a grid with five empty squares, which seemed surprising to them: Would we expect to see such a grid in another batch of thirty results?

During the next part of the intervention, occurring on a different day, instead of generating more data using physical experimentation, the dynamic software "Fathom" was used (Finzer, 2000). A simulation was created in Fathom that randomly placed sixteen dots on a 4 x 4 grid, and by toggling the animation feature, a single "trial" would unfold so the dots could slowly appear. In showing students the animation of a single trial, it was vital for students to question the veracity of the displayed result: How could they be sure the computer was doing it correctly? More salient was the question: Did the Fathom results look reasonable when compared to what the students had just done physically? Figure 1 has an example of the end result for a single trial.

After some discussion that led to the class accepting Fathom as being just as unpredictable as their physical models, we then were able to use Fathom to look at many trials, very quickly. In fact, whereas we had previously displayed on paper a batch of at least thirty trials, we know could use Fathom to see a batch of thirty trials within seconds.

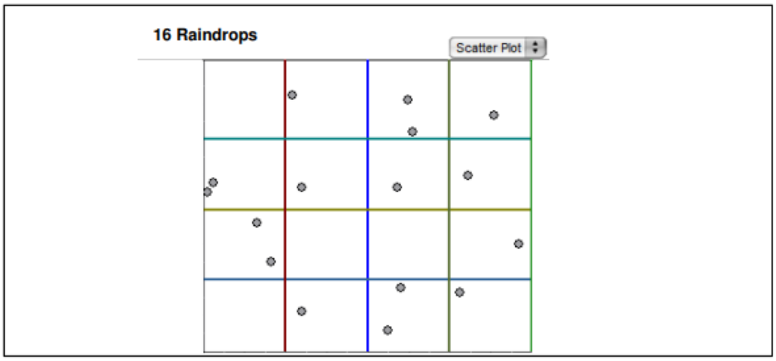
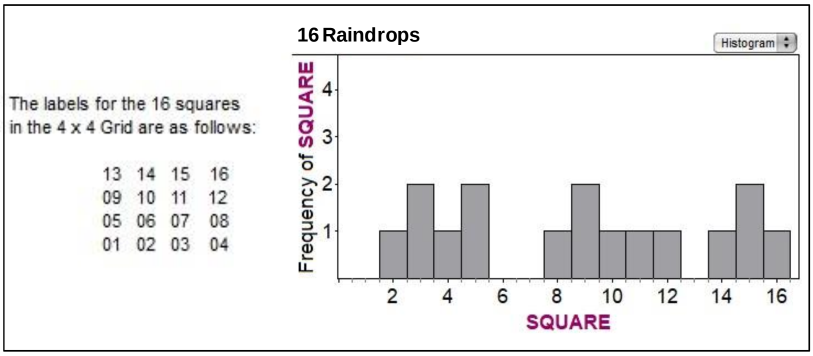


Figure 1: A single trial (via Fathom)

The point in generating more data was to investigate some of the question's students had raised earlier, and here is where Fathom played a key role. For example, Fathom could easily record how many "raindrops" were in each tile (which were numbered 1 – 16). Figure 2 below shows the Fathom tabulation of frequencies in squares for the result corresponding to Figure 1, along with a legend showing the square labelling convention for the grid.


The labels for the 16 squares
in the 4 x 4 Grid are as follows:

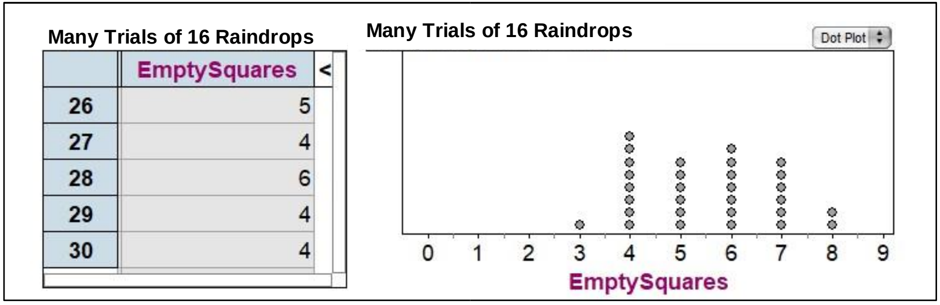
13	14	15	16
09	10	11	12
05	06	07	08
01	02	03	04

Figure 2: How many raindrops in each of 16 squares from Figure 1

At this point, the comments and questions made by students about the Fathom results paralleled those made regarding the students own experimental results. For instance, in Figure 2 we see a trial having at most 2 raindrops in any given square. How likely is such a result from any given trial? Were there any paper grids from students' own collective batch that matched that Fathom result or came close?

Again, the power of Fathom in quickly generating results came to bear as students noticed in Figure 2 that there were exactly four empty squares (they could look back at the actual grid in Figure 1 to verify that squares 1, 6, 7, and 13 were indeed empty, as cross-referenced with the labelling legend). But Fathom can record this result of "four empty squares" and then do another trial (recording how many empty squares), and so on. By using the animation feature of Fathom, we were able to slow things down in generating a batch of thirty trials, each time recording how many empty squares were in a trial.

It was important to run the initial "batch of 30 trials" on Fathom as slowly as possible, so that students could see that everything Fathom was doing mirrored the same ideas they had explored with their own paper recording grids. For example, Figure 3 shows results of such a batch of thirty trials, with frequencies for how many empty squares were in each trial. The last (30th) trial had exactly four empty squares such as was seen in Figures 1 and 2. And so a tally mark (a dot in this case) was added to that column. Students could see that of the thirty trials, eight trials had happened to have exactly four empty squares. And if needed, they could go back through the other displays and match a tally mark with the grid result it came from to verify that tally mark.



Many Trials of 16 Raindrops

	EmptySquares
26	5
27	4
28	6
29	4
30	4

Figure 3: Counting the empty squares in each of thirty trials

By generating more data, whether in increasing the number of trials (beyond 30, for instance), or in simply replicating many batches of the same number of trials, students were able to pursue deeper questions about what was expected. They also used their insights into what was likely to make inferences about purported results. At the end of the intervention, a series of "results" of physical experimentation was given to students, which were claimed to come from a single trial, or a batch of trials, depending on what was being asked. Students were then asked to imagine that "some other class from a school across town" had submitted these "results", but we weren't sure if the other class just made up the results or if they actually came from the other class doing the physical trials. Particular attention was given to way students based their inferences of "real or fake?" on the variability inherent in the Fathom data they had just been exploring.

3. Results

Among the questions in seeing repeated "batches of thirty trials" (which we sped up once the idea of what was going on was understood and accepted) was about what was reasonable to expect in terms of how many empty squares might be in any given trial. In Figure 3, representing a single batch of thirty trials, we see a minimum of three and a maximum of eight empty squares. So, what would be typical for the number of empty squares? If zero empty squares was considered very unlikely (corresponding to one raindrop per square), then wouldn't one or two empty squares be fairly likely?

In fact, students realized that Figure 3 was of poor use in ascertaining what was typical, since nothing too definitive emerges regarding the center of that distribution. After examining repeated batches of thirty trials, students wanted to aggregate the batches and we ended up doing 100 or more trials per batch. The time it would take Fathom to generate such results varied according to the relevant computer power, but usually something like 1000 trials only took about one minute or less. Figure 4 (below) shows the same idea of Figure 3 but a much stronger sense of distribution emerges.

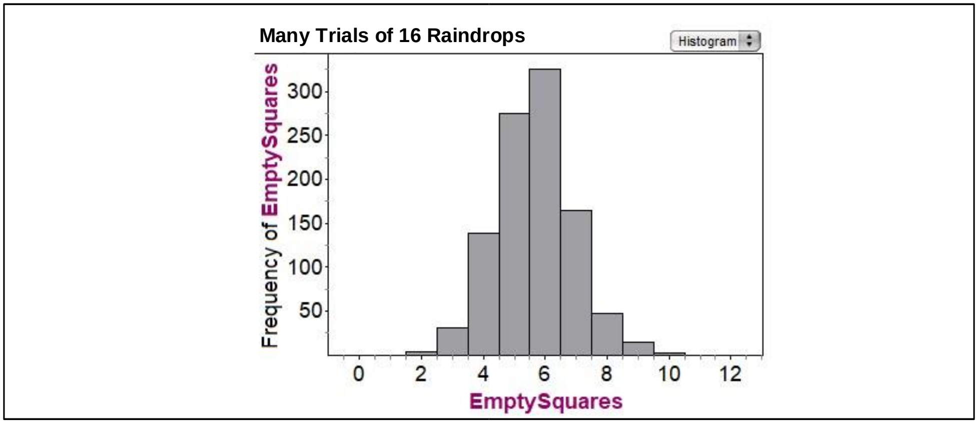


Figure 4: Counting the empty squares in each of 1000 trials

When shown some "real or fake?" data, for instance, students moved away from their claims such as "Who can tell for sure?" or "You never know, that [result] might have happened". Such claims show an over-appreciation of variability, especially when looking at the tails of distributions like that in Figure 4. Instead, as students reasoned about what was likely, they showed more sophistication in reconciling expected values with the variability they had witnessed in analysing many results.

Especially encouraging was the way students developed new questions to help decide on what real or fake data might be, such as "How likely is it that any given trial has 6 or more raindrops on a square?" Another question that we were able to gain data on from the Fathom simulations was "How many squares in any given trial are likely to contain exactly 2 raindrops?" This manner of generating questions, along with the student comments that picked up on the variability inherent in the supporting data, was a highlight for the results of the project so far.

At this stage in the analysis of students' written responses, a rubric is being developed to better affix a quantitative measure to the degree of improvement in students' use of variability in their reasoning. For example, Engel & Sedlmeier (2005) ascribed the "Novice" label to students whose predictions included between one and three empty squares, and the label of "Expert" to those between four and eight empty squares. They then structured numeric scores based on those labels (and also the lower levels of "Deterministic" and "Moderately Deterministic"). So far, an initial qualitative examination of responses shows a general increase in student confidence in making predictions, with a markedly stronger emphasis on trying to balance variability with expected values.

4. Discussion and Conclusion

Among the surprising results of the project so far has been the new avenues for questions that came from looking at the data generated by Fathom. A good example was when students asked the waittime question "How many trials would we expect before we hit exactly 6 empty squares?" This question seemed natural enough, given that one student after another might do a trial and not have that particular result. Or it might happen on the first try.

Some students did have bit of prior knowledge about an expected value for wait time as the reciprocal of the underlying probability, although it wasn't phrased that way. For instance, they might expect to roll a die six times to hit a "4". But again, there is variability to consider. In the context of the "falling raindrops" task, students could see that six empty squares had a high likelihood, say 0.342 for example. They then wonder if in fact 1/0.342 ≈ 2.92 might mean that "three trials" ought to be reasonable to hit exactly six empty squares. We then turned to Fathom to see if that in fact "three" was a reasonable answer for the above question on wait-time.

Perhaps the most intriguing question had to do with the probability of a square having a particular nonzero number of raindrops. They surmised a correlation between "number of empty squares" and "maximum number of raindrops in a square", but it turned out to be a challenging question to determine a specific probability for a given nonzero number of raindrops. For instance, "What's the likelihood any given trial will have 6 raindrops as a maximum on a square?" was a question that arose. Certainly, we could look at our original experimental data – the paper grids up around the room – and compute that experimental probability. But getting Fathom to "keep track" of how many trials had exactly six raindrops on a square (and no more than six) was complicated for us.

Instead, it was very easy to have Fathom run trials until the number of raindrops was six or greater. So, we changed the question to "What's the likelihood any given trial will have 6 raindrops or more on a square?" To gain insight into that question, we ran 100 experiments on Fathom, where an experiment was defined as "Count how many trials are needed to be run until a trial hits 6 raindrops or more on any given square".

Before running 100 experiments, we discussed what the results might look like. Students knew an experiment could end with "one trial" because we could get 6 or more drops on a square with the first trial. Some students thought an experiment could go on for "thousands of trials" since maybe it would take a while to get the desired result. We also noted that none of our initial experimental data had that result. After discussion of initial expectations, we ran Fathom for 100 experiments as defined above, and the results are in Figure 5.

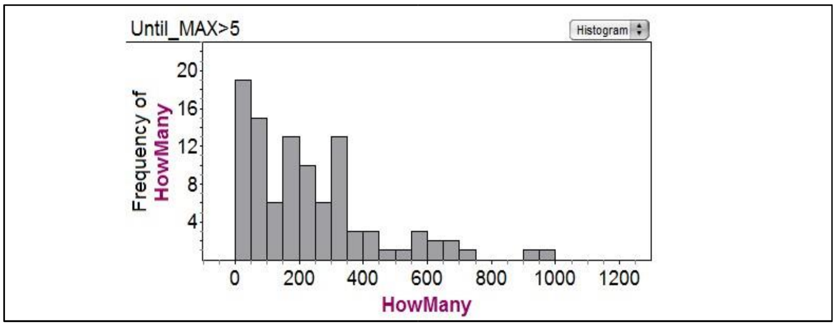


Figure 5: 100 Experiments of "How many trials to get a square with 6 or more drops?"

Using the mean result from 100 experiments, which was about 230 trials for Figure 5, students conjectured that the question of "What's the likelihood any given trial will have 6 raindrops or more on a square?" might be the reciprocal of the wait-time: 1/230 ≈ 0.0043. However, this low probability did not satisfy those who thought any given trial must surely have be fairly likely to have the desired result. Again, precise mathematical computations were not the aim of the project, but students did raise very interesting probabilistic and statistical questions. They were left musing about the correlation between "maximum number of drops" and "number of empty squares", so in that sense their curiosity had not been fully slaked.

Overall, by the end of the intervention students seemed to demonstrate three features useful for developing a more robust engagement in a world beset by variability. First, students markedly changed their predictions of where sixteen raindrops might land, as they were exposed to ever-increasing amounts of experimental data. Second, students were better able to integrate a reasoning about variability in making inferences about hypothetical results. Third, and perhaps most intriguing, students generated further questions that were based on what they noticed, and what they wondered about, in the face of large amounts of simulated data.

The latter questions are what really made this project and paper unique, in the way that students furthered their investigation of a well-known task. The next step will be to employ a conceptual framework to assess student responses in order to describe in more detail the ways in which their appreciation and use of variability improved by the end of the instructional intervention.

References

1. Engel, J. (2002). Activity-based statistics, computer simulation, and formal mathematics. In B. Phillips (Ed.), Proceedings of the Sixth International Conference on Teaching Statistics. CD ROM.
2. Engel, J. & Sedlmeier, P. (2005). On Middle-School Students' Comprehension of Randomness and Chance Variability in Data. Zentralblatt füur Didaktik der Mathematik (2005) 37: 16
3. Finzer, W. (2001). Fathom! (Version 1.16) [Computer Software]. Emeryville, CA: Key Curriculum Press.
4. Green, D. R. (1982). A Survey of Probability Concepts in 3000 Students aged 11–16 Years. In D. R. Grey (ed.), Proceedings of the First International Conference on Teaching Statistics, Teaching Statistics Trust, University of Sheffield, 766–783.
5. Piaget J., & Inhelder, B. (1975). The Origin of the Idea of Chance in Children. London: Routledge & Kegan Paul

If Retail Trade Sales Falls, Will GDP Follow? A Case Study of Malaysia

Norzarita Samsudin, Zainuddin Ahmad
Department of Statistics Malaysia

Abstract

Sales value of retail trade for Malaysia grew quite significant lately. In 2016, it registered 8.7 per cent growth, while in 2017, it posted 11.5 per cent. In the meantime, Malaysia's Gross Domestic Product (GDP) registered 4.2 per cent growth in 2016, and 5.9 per cent in 2017, respectively. The objective of this study is to examine relationship between sales value of retail trade and GDP of Malaysia where quarterly sales value of retail trade and GDP of Malaysia is used. This study uses econometric analysis i.e. Vector Autoregressive (VAR). Granger causality test is applied to study causality between the studied variables. This paper also examines future reaction of the studied variables where Impulse Response Functions (IRF) and Variance Decomposition are applied. The results of Johansen Cointegration Test show that there is no cointegration in the long run between the studied variables. In the short run, sales value of retail trade lag one and Malaysia real GDP lag one have positive relationship with Malaysia real GDP. Granger causality test shows real GDP Granger cause retail sales, while retail trade sales doesn't Granger cause GDP. If there is a positive shock to Malaysia real GDP, sales value of retail trade will react positively both in the short and long run. Conversely, a shock to retail sales gives a marginal negative impact to GDP in the short and long run. In Variance Decomposition of real GDP study, it is found sales value of retail trade's contribution is very small in short and long run. Meanwhile, Variance Decomposition of sales value of retail trade reveals that contribution of real GDP is increasing more significant in long run.

Keywords

Sales value of retail trade; growth; GDP; sales; Malaysia

1. Introduction

Retail trade is one of important activities in any economy. Sales value of retail trade is also a very essential component of GDP, whereby rising retail trade means a growth in consumption and a fall in unemployment. Sales value of retail trade consists of selling merchandise in the state that it is purchased (or after minor transformations), generally to a customer base of private individuals, regardless of the quantities sold. In addition to sales, retail trade activity may also cover delivery and installation at the customer's home (of furniture or household appliances for example). Gross domestic product (GDP) is a monetary measure of the market value of all the final goods and services produced in a period of time, often annually or quarterly. Real GDP estimates are commonly used to determine the economic performance of a whole country or region, and to make international comparisons.

Sales value of retail trade for Malaysia grew quite significant lately. In 2016, it registered 8.7 per cent growth, while in 2017, it posted 11.5 per cent. Meanwhile, in quarter two 2018, sales value of Malaysia retail trade grew 9.9 per cent, while 13.0 per cent was recorded in quarter three 2018. In the meantime, Malaysia's GDP registered 4.2 per cent growth in 2016, and 5.9 per cent in 2017, respectively. In quarter two 2018, Malaysia GDP grew 4.5 per cent, while 4.4 per cent was recorded in quarter three 2018 (Department of Statistics Malaysia, 2018).

The objective of this paper is to study relationship between real GDP and sales value of retail trade of Malaysia.

2. Literature Review

According to Deloitte Insight Article (2018), retail trade is a key avenue of consumer spending and makes a major contribution to the labour market. The retail sector however, is undergoing some changes, which have impacted total sales, revenues of different store types, and employment. Retail trade is also a major contributor to employment. According to the Bureau of Labour Statistics' establishment survey, retail trade made up 10.8 per cent of total non-farm payrolls in the economy in the year until November 2017. In this paper, it is also mentioned that although retail sales have gone up, they have fallen slightly when compared to GDP.

Shehnaz Tehseen (2016), in his study mentions that retail industry is given much more importance because of its direct link with the final consumers. Retailing as a commercial transaction is very important due to the consumption of goods and services by the buyers through family, personal or household use. In 2017 Malaysia has been ranked third in the 2017 Global Retail Development Index (GRDI) for the second consecutive year. Management consulting firm A.T. Kearney attributed the high ranking to the influx of tourists, higher disposable income and government investments in infrastructure had boosted the retail industry. Malaysia's retail market continued to grow despite a slight dip in overall GDP growth and short-term pressures of currency fluctuation and inflation.

In other study, Raja Nurul Aini and Amalina (2017) investigate the relationship between GDP growth and the factors such as Inflation, Foreign Direct Investment (FDI) and Female Labour Force Participation in Malaysia. Least Square Method (OLS) and Augmented Dickey Fuller (ADF) are used for the analysis. The results identify that among the factors of FDI and Female

Labour Forces have positive impact on GDP growth. However, FDI is the only variable that contributes significantly to GDP growth in Malaysia.

3. Methodology

This study used econometric analysis i.e. Vector Autoregressive (VAR) to examine the relationship between Malaysia sales value of retail trade and GDP at constant price. Secondary data is used that is quarterly data from quarter one 2010 to quarter three 2018. Because of VAR models represent the correlations among a set of variables, they are often used to analyze certain aspects of the relationships between the variables of interest (Rossi, 2018). In this study, after testing the variables involved, “Unrestricted VAR” model is suitable to be used to examine short run relationship between Malaysia real GDP and sales value of retail trade. In this study it was found that Malaysia Real GDP has seasonality. Hence the seasonality is removed before conducting relationship study. The general model of “Unrestricted VAR is as follow:

$$Y_t = b_{10} - b_{12}X_t + \gamma_{11}Y_{t-1} + \gamma_{12}X_{t-1} + \varepsilon_{yt}$$

$$X_t = b_{20} - b_{21}Y_t + \gamma_{21}Y_{t-1} + \gamma_{22}X_{t-1} + \varepsilon_{xt}$$

Where:

Y = real GDP of Malaysia
X = Malaysia sales
value of retail trade
b and γ = constant
term t = time
trend ε = error
term.

Stationarity test is conducted in order to select appropriate VAR model where Augmented Dickey– Fuller test (ADF) is applied in this study. It is found that the studied variables are not stationary at level. Real GDP and sales value of retail trade are stationary at 1st difference. Hence, Johansen Cointegration Test is performed to examine long run relationship. Granger causality test is used to investigate the causality between real GDP and sales value of retail trade. This is to determine whether Malaysia real GDP can influence the sales value of retail trade, or sales value of retail trade can cause the Malaysia GDP in the short run.

In this study also, Impulse Response Functions (IRF) and Variance Decomposition are used to examine future reaction of the studied variables. IRF identify the responsiveness of the dependent variable (endogenous variable) in the VAR when a shock is applied to the error term. Variance Decomposition technique separates the variation in the endogenous variable into the component shocks to the VAR. Thus, the variance decomposition provides information about the relative importance of each random shock in affecting the variables in VAR (Ogungbenle, Olawumi and Obasuyi, 2013).

4. Discussion and Conclusion

Correlation analysis is conducted between Malaysia real GDP and sales value of retail trade. The studied variables has very strong positive relationship (r = +0.96). Unit root test is conducted to ascertain level of integration of the studied variables. It is found that both the real GDP and sales value of retail trade are stationary at first difference. The summary is in Table 1.

Table 1: Augmented Dickey-Fuller Test Statistic

Variable	Stationarity	t-Statistic	Prob.
Gross Domestic Product	First difference	7.15	0.0000
Sales value of retail trade	First difference	-6.34	0.0000

Source: Author's computation

Akaike Information Criterion (AIC) is used to identify optimal lag. The optimal lag is Lag one. The summary result is shown in Table 2 below.

Table 2: Optimal Lag: Akaike Information Criterion (AIC)

43.59738	Lag 0
36.74791*	*Lag 1
36.91179	Lag 2

Source: Author's computation

The next step is to find out whether the real GDP and sales value of retail trade are integrated in the long run.

4.1 Johansen Cointegration Test

Johansen Cointegration Test is used to determine the long run cointegration of the studied varibles. Trace test and Max-eigenvalue test indicated that there are no cointegration between real GDP and sales value of retail trade at the 0.05 level. The result of Johansen co-integration test is presented in Table 3.

Table 3: Johanson Co-integration Test
Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Trace Statistic	0.05 Critical Value	Prob.
None	10.40900	18.39771	0.4419
At most 1	0.167248	3.841466	0.6826

Trace test indicates no co-integration at the 0.05 level

Source: Author's computation

Therefore, "Unresricted VAR model" is used to study relationship of real GDP and sales value of retail trade of Malaysia.

a. Unrestricted VAR Model

Table 4 presents result of the Unrestricted VAR model. It is found that sales value of retail trade lag one (p-value < 1 per cent) and real GDP lag one (p-value < 5 per cent) are significant to explain sales value of retail trade. One per cent increase in sales value of retail trade lag one will result in 0.75 per cent increase in sales value of retail trade. Meanwhile, one per cent increase in real GDP lag one will result in 0.18 per cent increase in sales value of retail trade.

It is also found that only Malaysia real GDP lag one is significant to explain Malaysia real GDP. Coeficient of Malaysia real GDP lag one is 1.014 (p-value < 1%).

Table 4: Unrestricted VAR Model

Variables	Coefficient	Std. Error	t-Statistic	Prob.
Dependent variable				
(Sales value of retail trade)	0.747950	0.124225	6.020949	0.0000
Sales value of retail trade 1				
Real GDP lag 1	0.177940	0.077533	2.295022	0.0251
Dependent variable (GDP)				
GDP lag 1	1.014394	0.024219	41.88504	0.0000

Source: Author's computation

4.3 Granger Causality Test

Granger Causal Test was performed to determine whether independent variable can cause dependent variable. Table 5 present the result.

Table 5: VAR Granger Causality Tests

Null Hypothesis:	Obs	Chi-sq	Prob.
Real GDP does not Granger Cause sales value	34	5.267124	0.0217
Sales value does not Granger Cause real GDP		0.030952	0.8603

Source: Author's computation

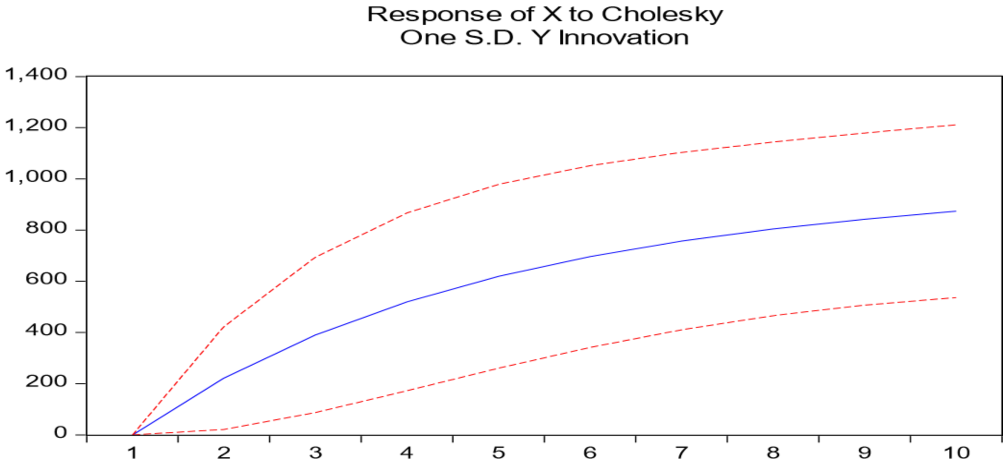
It is found that real GDP Granger cause sales value of retail trade, while sales value of retail trade doesn't Granger cause real GDP.

4.4 Impulse Response Function (IRF)

IRF is to study reaction (in future) between Malaysia real GDP and sales value of retail trade in a period of time if there is a shock in the VAR system (model). In this study, ten periods or quarters are selected.

Response on sales value of retail trade (Graph 1). A one standard deviation shock to GDP has noticeable impact on retail sales. The response significantly increases until period 10th where it maintains in positive region.

Graph 1: Impulse Response Function: Response on Sales Value of Retail Trade

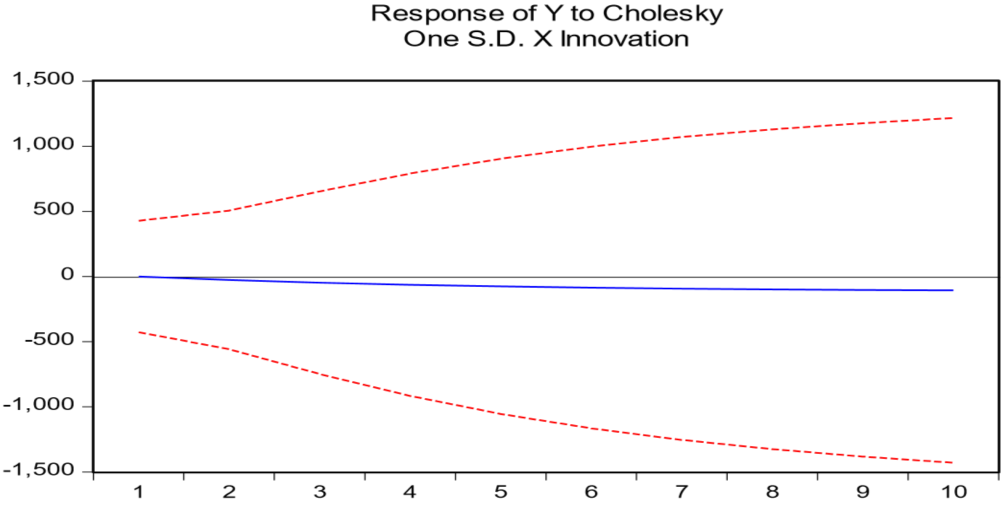


Source: Author's computation

Note. X is sales value of retail trade; Y is real GDP.

Response on Real GDP (Graph 2). A one standard deviation shock to sales value of retail sales initially has no reaction to real GDP in the short term. But there appears negative reaction gradually and steadily until to period ten where it remains in the negative region. We conclude a shock to sales value of retail sales give marginal negative impact to GDP in the short and long run.

Graph 2: Impulse Response Function: Response on Real GDP



Source: Author's computation

Note. X is sales value of retail trade; Y is real GDP.

4.5 Variance Decomposition

In Variance Decomposition study, ten quarters period were selected. This study enabled us to forecast ten quarter ahead on the variance decomposition of real GDP and sales value of retail trade of Malaysia.

Variance Decomposition of real GDP (Table 6). In short run (period 1), almost 100 per cent of forecast error variance in real GDP is explained by real GDP itself. The contribution of retail sales is very small. This implies very weak influence of this variable in the future. In the long run, the influence of GDP on itself is still very strong, while contribution of retail sales almost no change.

Table 6: Variance Decomposition of Real GDP

Period	S.E.	X	Y
1	1245.553	1.10E-05	99.99999
5	2863.124	0.155249	99.84475
10	4172.925	0.346001	99.65400

Source: Author's computation

Note. X is sales value of retail trade; Y is real GDP.

Variance Decomposition of retail trade sales (Table 7). In short run (period 1), almost 100% of forecast error variance in retail sales is explained by retail sales itself. In the long run, the influence of sales value of retail trade on itself is gradually decreasing, while contribution of GDP increases to approximately ten per cent in period ten. This implies real GDP will influence retail trade sales in the future.

Table 7: Variance Decomposition of Retail Sales

Period	S.E.	X	Y
1	3987.510	100.0000	0.000000
5	5904.051	97.54495	2.455054
10	6300.303	89.83461	10.16539

Source: Author's computation
Note. X is sales value of retail trade; Y is real GDP.

4.6 Diagnostic Test

The Unrestricted VAR model has a significant R-square and F-statistics. VAR Residual Serial Correlation LM Tests shows there is no serial correlation (prob. 0.5642). VAR Residual Heteroskedasticity Tests demonstrates there is no heteroskedasticity (prob. 0.4041). However, VAR Residual Normality Tests Jarque-Bera illustrates the residual is not normal (prob. 0.01),

5. Conclusions

The objective of this study is to examine relationship between Malaysia real GDP and retail trade sales where Vector Autoregressive model, Impulse Response Function and Variance Decomposition statistical technique were used. Result showed that there was no co-integration in the long run between the studied variables. In the short run, retail trade sales lag one and Malaysia real GDP lag one have positive relationship with Malaysia real GDP.

If there is a positive shock to Malaysia real GDP, retail trade sales will reacts positively both in the short and long run. Conversely, if there is a positive shock to retail trade sales, the Malaysia real GDP will react insignificantly in the short and long run. Variance decomposition of Malaysia real GDP study showed that in the short and long run, retail trade sales has no influence to Malaysia real GDP. Meanwhile, in the short run, the retail trade sales influence on itself is very strong, while in the long run, the contribution of Malaysia real GDP will increase to almost ten per cent in the variance decomposition of retail trade study. In this study, there are only two variables are examined. For other research, other variables that can be included are investment, consumption or credit.

References

1. Deloitte Insight Article (2018). Retail sales trends in revenue and employment. Retrieved February 4, 2019 from https://www2.deloitte.com/insights/us/en/economy
2. Department of Statistics Malaysia, (2018), Malaysia Economic Performance. Retrieved January 25, 2019 from https://www.dosm.gov.my
3. Ogungbenle, S., Olawumi, O.R., and Obasuyi, F.O.T. (2013). Life Expectancy, Public Health Spending and Economic Growth in Nigeria: A Vector Autoregressive (VAR) Model. European Scientific Journal, Vol. 9, No. 19, 210 – 235.
4. Raja Nurul Aini Raja Aziz and Amalina Azmi (2017). Factor affecting Gross Domestic Product (GDP) in Malaysia. International Journal of Real Estate Studies. Vol. 11, No.4, 61-67.
5. Shehnaz Tehseen, (2016). Malaysian Service Sector: An Overview of Wholesale and Retail Industry. Retrieved February 4, 2019 from https://www.researchgate.net/publication

Poverty incidence in Kelantan State "Finding of Household Income Survey (HIS) vs. E-Kasih"

Syamaizar Razali, Wan Aziam Wan Awang
Department of Statistics Malaysia

Abstract

The phenomenon of poverty is a condition that will exist in any place or country. This due to differences in the income levels and purchasing power of consumers in human needs such as food, clothing and shelter. The environmental factors also influence of a household to be functioning in the communities is unsuccessful and not equal.

Hence, various plans have been made by the Federal Government in helping to eradicate these poverty issues. On this issues, the Department of Statistics Malaysia (DOSM) have conducted the Household Income Survey (HIS), in the meantime the ICU JPM also conducted the Poor Household Census (BIRM) for the purpose of registering in the e-Kasih system to monitor the impact of the government's incentives that has been channelled to these low-income groups.

The relevance of both studies will shows that HIS can provide a good vision of areas with high rates of poverty, while the data of e-Kasih will provide the information/ profile of low income groups to provide to all government agencies under their ministers that providing assistance such as Ministry of Finance Malaysia (MOF), Ministry of Rural and Regional Development (KKLW), Ministry of Housing and Local Government (KPKT), Ministry of Woman, Family and Community Development (KPWKM), and others.

The finding shows that HIS can provide a real vision of the incidence of poverty, while e-Kasih plays a role in monitoring or co-ordering programs of poverty alleviation. The right target audience will definitely have the maximum impact in eradicating poverty issues in a country.

Keywords

Poverty; HIS; e-Kasih

1. Introduction

Poverty is a condition which is the person lack of the living things and this will associated with money shortage problem. Poverty phenomenon will happen when household (IR) will not be able to funtioning in community surrounding in fulfilling the basic requirements such as food, clothes and residence. Poverty also can define the disability of person to generate income to advancing their family institution. At the national level, poverty issue can be a measure of a progress in a country. For example, there are three categories of poverty which is poor hardcore, poor and B40 that three of this is under the national poverty line income (PLI).

Therefore, the Department Of Statistics Malaysia (DOSM) take the initiative to do the study about poverty issue in research, titled Household Income Survey (HIS). The indicators that will be measured under HIS such as mean household income, median household income, sources of income, Gini coefficient, incidence of poverty and others.

The Implementation & Coordination Unit (ICU) under Prime Minister Office (PMO) also carry out poverty the studies but rather to get the list of the lower – income groups (Below 40 as known as B40) to register in the e-Kasihsystem. e-Kasihsystem is a not a database that show whole household income under the Poverty Line Income (PLI) because the listing into the e-Kasihdatabase system is based from some limited resources such as reports from mass-media, reported by the Government or NGO's Department or other Agencies and their own nominees who present at the ICU officefor the purpose of this registration.

Hence the main purpose of this study is to look at the importance and use of the data reported in the **Household Income Survey (HIS)** and the **e-KasihSystem** in explaining the **Poverty Incidence.**

2. Methodology

This study uses descriptive analysis to see the situation and poverty rate in Kelantan. The referenced data are based on the HIS publications in 2009, 2012, 2014 and 2016 and also the data recorded in the e-Kasih system.

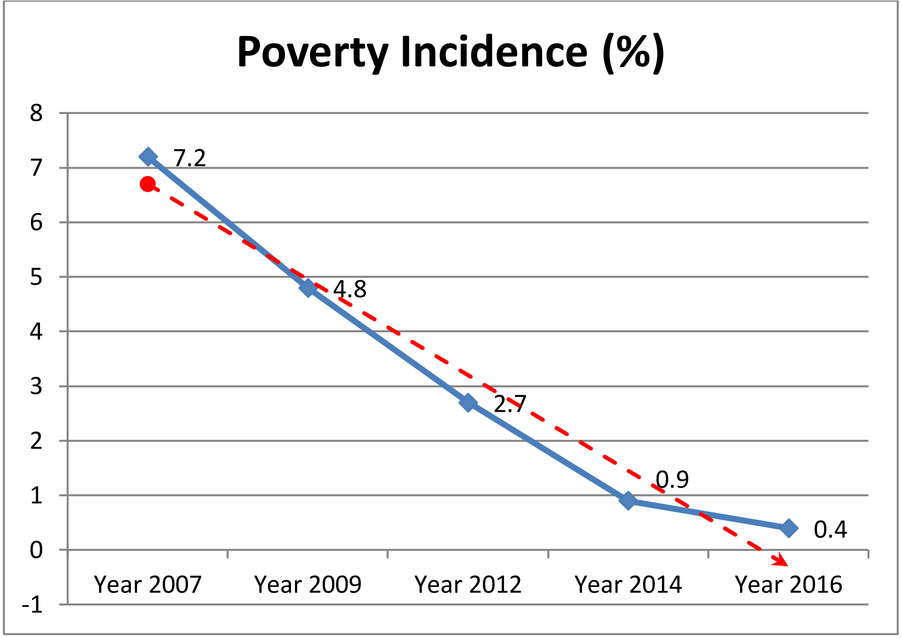
HIS wasfirstimplemented in 1974 and conducted in every two years. The list of residenceselected randomly by using the probabilistic sampling method that covers all states and administrative districts in Malaysia. This survey covers all households living in private residences and citizens only.

Cencus of Poor Households (CPH) under ICU was first implemented in June 2008 and conducted continuouslybased on the current number of registrations for the groups under PLI for the purpose of registering in the e-Kasih system. The e-Kasih system is also dynamic in which the number of registered participants and their poverty status will be change from time to time based on updated information or profile additions in this system.

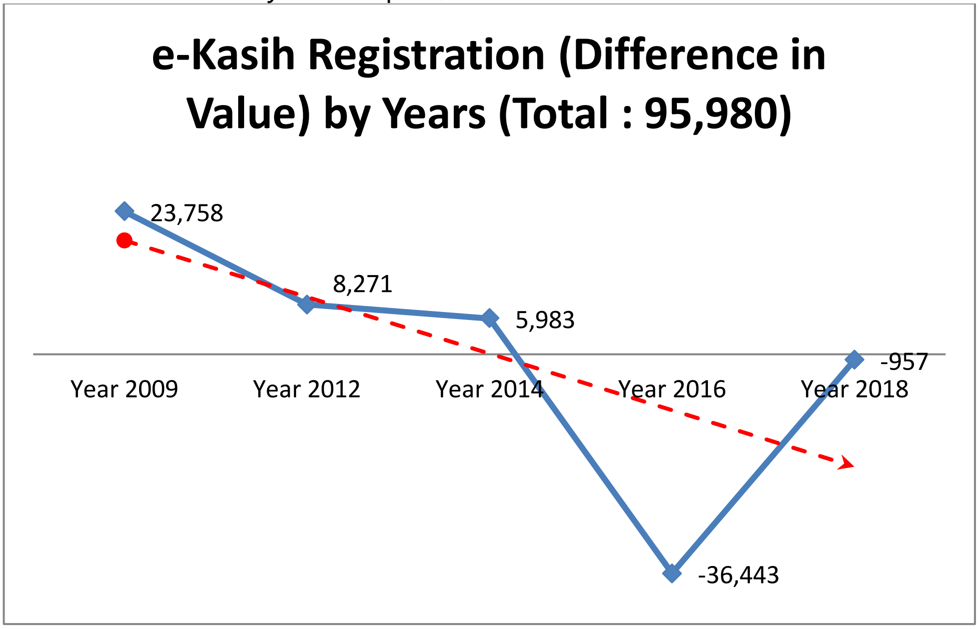
The main purpose of the selection of these two studies is to see how far the use of indicators in each of these studies is used in overcoming the poverty of the state, especially in Kelantan.

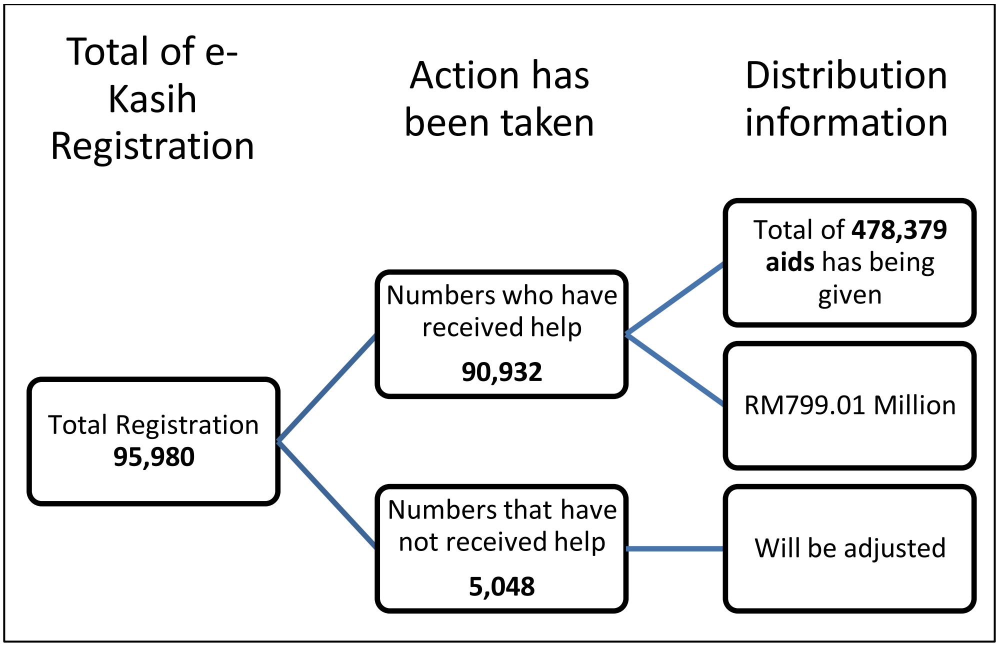
3. Results

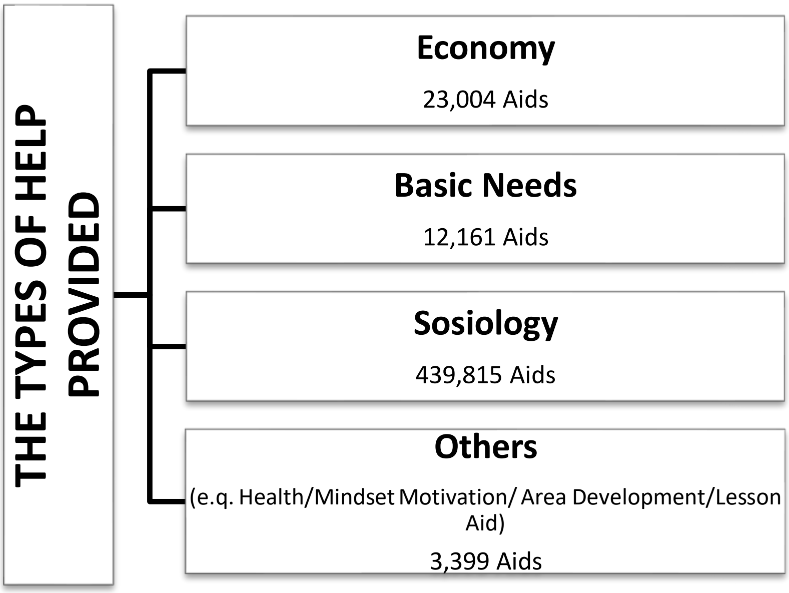
a. HISReport, Kelantan



b. e-Kasih System Report, Kelantan




Total of e-Kasih Registration
Action has been taken
Distribution information
Total Registration **95,980**
Numbers who have received help **90,932**
Numbers that have not received help **5,048**
Total of **478,379 aids** has being given
RM799.01 Million
Will be adjusted


THE TYPES OF HELP PROVIDED
Economy 23,004 Aids
Basic Needs 12,161 Aids
Sosiology 439,815 Aids
Others (e.q. Health/Mindset Motivation/ Area Development/Lesson Aid) 3,399 Aids

4. Discussion and Conclusion

Overall, the results of this study explain that Kelantan's poverty rate by HIS report has show the decreased from 7.2% in 2007 to 0.4% in 2016. The decline was due to the implementation of the Poverty Eradication Programme from 2009 to 2019 involving an allocation of RM799.09 million with channeling 478,379 assistance to groups under this PLI.

References

1. “LAPORAN PENYIASATAN PENDAPATAN ISI RUMAH DAN KEMUDAHAN ASAS 2016”by the Department of Statistics, Malaysia
2. The study by Arius Jonaidi entitled“ANALISIS PERTUMBUHAN EKONOMI DAN KEMISKINAN DI INDONESIA”
3. The study by Katty Hsiao Jia Wong, Ti Ching Yan & Janice Lay Hui Nga entitled“KEMISKINAN DAN PENDIDIKAN: KESAN DAN CABARAN DALAM KALANGAN BELIA MALAYSIA”

Factors affecting the gross output value of real estate subsector in Malaysia

Siti Salwani Ismail[1], Wan Rahifah Wan Ramli[1], Mohamad Helmi Hidthiir [2].

[1] Department of Statistics, Malaysia

[2] Universiti Utara Malaysia

Abstract

The thriving of high-value projects would continue to catalyse property business activities in Malaysia and bring about new growth drivers and opportunities. There are expectations that the real estate subsector to boom at the end of 2018 as a spill over from this robust economic expansion. This development would be desirable but still with uncertainties as there are many factors contributing to the growth in the real estate subsector. Therefore, the aim of the paper is to study the factors that affecting the gross output value of real estate subsector in Malaysia from year 2010 to 2015. This study only focused on three factors namely value of loan applied, base lending rate and average house price as independent variables. Linear regression is used to determine the relationship between the variables. Granger causality test is employed to investigate the causality between economic variables and output of real estate subsector. Based on the findings of the study, loan approved and base lending rate has negative effect to dependent variable. While, average price house was positive effect with dependent variable of this study. The output of real estate subsector was granger cause to base lending rate.

Keywords

real estate subsector, gross output and regression analysis

1. Introduction

The real estate subsector is one of the main drivers of Malaysia economic growth. This subsector is expected to remain its steady growth driven by the implementation of various projects under the Economic Transformation Program (ETP), the 2013 Budget and the Tenth Malaysia Plan (RMK-10). In 2015, the real estate subsector contributed 1.4 per cent to the total gross domestic product (GDP) and 2.7 per cent to the total of services sectors. Meanwhile, value of gross output generated for these services amounted to RM28.1 billion. Within the period of 2010 and 2015, there was an increase of RM9.4 billion registering a CAGR of 8.5 per cent.

The business main enabler in the real estate subsector are the real estate developers. Generally, these real estate developers or property developer, involve in the activities encompassing from renovation and re-lease of existing buildings to purchase of raw land and the sale of developed land or parcels to others. Thus, the performance of the property developers in this study will be evaluated based on the gross output generated by their establishments.

2. Objectives

The objectives of this study are:

i) To investigate the economic variables that affects the value of gross output of real estate subsector in Malaysia.

ii) To identify the relationship between output of real estate subsector and economic variables.

3. Literature Review

Gross output by industry provides important insights into an industry's contribution to the overall economy. Gross output is principally a measure of sales or revenue from production for most industries (SNA, 2008). The value of gross output for the real estate subsector is defined to include commissions and brokerage received on sales from land, residential, non-residential and other properties; commissions and brokerage received on rental/lease transaction from land, residential, non-residential and other properties; rental income received from land, residential, non-residential and other properties; sale income received from land, residential, non-residential and other properties; income received from valuers / appraiser of real estate; income received from property management; income received from management services and other income (DOSM, 2016).

According to Wilkinson, S. J. (2008), property development is an exciting and occasionally frustrating, increasingly complex activity involving the use of scarce resources. It is a high-risk activity that often involves large sums of money tied up in the production process, providing a product that is relatively inseparable and illiquid.

The sales of Malaysia's property are driven by various factors. Chia, J (2016) identified five variables which were found to be significant and have positive relationships with house purchase intention. Their findings showed that financing, distance, superstition numbers, environment and house features were important attributes to house buyers when they purchase a house. While according to Sean, S. L., & Hong, T. T. (2014). location, financial and structural factors prove to be significant at 5% with p-value of 0.026, 0.011 and 0.024, respectively as the factors taken into the consideration in acquiring residential properties.

From January to May 2017, commercial banks approved over RM25.7 billion and disbursed more than RM24.6 billion in loans for the purchase of residential property. Housing loans formed the single largest component of commercial banks' total loan portfolio, representing 34.4% of the total outstanding loans as at end May 2017 (The Star, 2017). This study will use the value of loan approved in five years period as one of the independent variables to be analyse.

Customers of banks and other financial institutions borrow for a number of reasons and their ability to pay back their loans can be attributed to a number of factors which include lending rates at which they borrowed the loans. High lending rates will impact on borrower's ability to pay which also hinder the consumer to purchase the property in the first place (Evans, 2014). The second independent variable which will be observed in this study is the base lending rate.

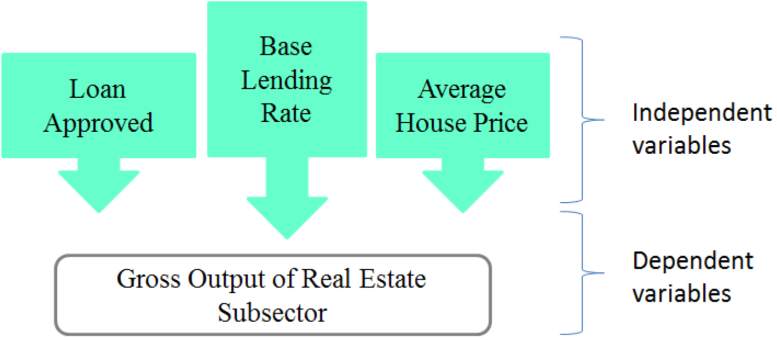
In year 2012, the Malaysia's housing market was ranked 9th in the list of "The World's Hottest Real Estate Markets" by a real estate consultancy Knight Frank with 5-year price growth of 28.5 per cent (Lim Sze Yoong et al.2016). This is achieved by the rapid increase in overall property prices, openness to foreign investment and future growth potentials of the housing market. In relation to it, according to MacLennan (2002), economists have characterised housing as a bundle of attributes. Some of these attributes are derived from the internal characteristics of the house unit itself such as the rooms available, whilst examples of external are location, accessibility to utilities, services and facilities. These attributes has influence in determining the house price. According to the Valuation and Property Services Department's The Residential Prices Quarterly Update Q2 2018, the average of all houses' pricing in Malaysia has continued to increase from 2010 till date. Thus, the average house price was selected as another independent variable in this study.

3. Methodology

The purpose of this section to apply the conceptual discussion and quantitative data econometric test of variables that effect to output of real estate subsector. This research aims to determine the relationships that exist between macroeconomic variables (loan approved, base lending rate and average house price).

4. Theoretical Framework

The following theoretical frame work shows how the independent variables which were represented by value of loan approved, base lending rate and average house price will influence the gross output value of residential developer.



5. Data

The data for this study was drawn from three main sources: Economic Census / Survey of Real Estate Services, Central Bank of Malaysia Statistical Bulletin and statistical compilation by National Property Information Centre which is available in its website. For the purpose of analysis, a time series of 5 years data from year 2010 to year 2015 was used. The data extracted were gross output of real estate subsector, base lending rate, loan applied for residential property and average price house in Malaysia. In order to reduce the volatility, all variables were converting to log linear form (for example using ln output of real estate subsector instead of actual value of output of real estate subsector).

6. Model

In order to develop the regression model for factors affecting the gross output value of real estate subsector in Malaysia, multiple regression function was used.

The regression function was shown in Equation 1.

$$\boldsymbol{Y_i = \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{3_i} + e_i}$$
(1)

Where Y is the dependent variable, X_2 and X_3 the explanatory variables (or regressors), e the residual term and i the ith observation; in case there are time series, the subscript t will denote the t^{th} observation.

For this study, the function of gross output value for real estate subsector was given in Equation (2).

O = f (LA, BLR, APH) (2)

Following variables are used throughout the model:

O = Gross output of real estate subsector

LA = Loan approved

BLR = Base lending rate
APH = Average house price

Independent variable this model was loan approved. The loan approved will affect the increase or decrease of gross output value for real estate subsector. This is because approval of loan is formal authorization to get a loan usually from a bank.

Residential property is the larger subsector that is contributed to the real estate subsector in Malaysia. Therefore average house one of the proxy to study the factor of effect to value gross for real estate subsector. Average house price may the factor of demand of house in Malaysia. Real house prices are directly determined by the willingness of households to pay for (and willingness of builders to supply) a constant-quality house. Changes in the quantity of housing demanded will affect real prices only to the extent that the long-run housing supply schedule is positively sloped.

Hypothesis:

H_1: There is a significant relationship between gross output value and its independent variables.

7. Estimation Results

7.1 Descriptive Statistics

The next section of the analysis was concerned with the summary statistics for all the variables. As Table 1 showed, the value of gross output had a mean of RM22,838 million with standard deviation RM3,612 million. The loan approved had a mean of RM103,270 million with standard deviation of RM15,715 million. The mean for base lending rate is 6.50 with standard deviation 0.26. The mean value for average house price recorded RM285,581 and a standard deviation of RM53,466.

Table 1: Summary Statistics

Variables	Output Real Estate (RM Million)	Loan Approved (RM Million)	Base Lending Rate	Average Price House
Mean	22,838	103,270	6.50	285,582
Median	22,320	99,355	6.53	286,674
Std. Dev.	3,612	15,715	0.26	53,466

7.2 Correlation Analysis

Based on analysis below showed that that loan applied had strong positive relationship between gross output of real estate subsector with loan approved, base lending rate (BLR) and average price house with 0.673, 0.865 and 0.989 respectively. Meanwhile, loan approved also stated strong relationship with base lending rate and average price house. Base lending rate also has strong relationship with average price house with 0.899.

Table 2: Correlation Analysis

Variables	Output Real Estate	Loan Approved	Base Lending Rate	Average Price House
Output Real Estate	1.000	0.673	0.865	0.989
Loan Approved	0.673	1.000	0.631	0.724
Base Lending Rate	0.865	0.631	1.000	0.899
Average Price House	0.989	0.724	0.899	1.000

7.3 Unit Root Test

The unit root test stationarity in Table 3 was based on the Phillips-perron (PP) unit root test. The result reported that value output of real estate subsector is stationary at second differencing, while other economy variables stationary at first differencing.

Table 3: Unit Root Test

variables	levels		First differences		Second differences		Conclusion
	t-statistics	p-value	t-statistics	p-value	t-statistics	p-value	
Output of real estate subsector	-1.5998	0.8838	-1.6105	0.1337	-1.6562	0.0261	I(2)
Loan Approved	-1.5973	0.8613	-1.5998	0.0783	-1.6105	0.0559	I(1)
Base Lending Rate	-1.5973	0.9904	-1.5998	0.0106	-1.610	0.0053	I(1)
Average Price House	-1.5998	0.5944	-1.6105	0.0415	-1.6562	0.0077	I(1)

**Represent 10% significance*

7.4 Estimation Equation

This section will further discuss on the result of estimation equation. This equation measure the relationship between three independent variables that was loan approved, base lending rate and average house price with gross output value of real estate subsector.

Table 4: Model Summary

Variables	Coefficient	t-Statistic	p-value
Loan approved	-0.1137	-0.8901	0.4389
Base lending rate	-0.0557	-0.6789	0.5459
Average price house	0.9324	6.5141	0.0073

R-squared	0.9805
Durbin-Watson stat	1.8755

The R^2 shows that the loan approved, base lending rate and average house accounted for over 98 per cent of the variation in gross output of real estate subsector in Malaysia over the research period. Meanwhile, from Durbin-

Watson test shows that there are positive autocorrelation between gross output value and its independent variables.

In the final part of the analysis, the function was developed as follows:

Gross output value of real estate subsector = -0. 1137 Loan approved- 0.0557 Base lending rate +0.9324 Average price house + ei

As these result showed, the coefficient for all independent variables was significant at 10 per cent level with the expected sign. When loan approved and based lending rate increased by one per cent, gross output will decrease by 0.11 per cent and 0.06 per cent respectively. While average price house increase by one per cent, gross output for real estate industry will increase 0.93 per cent.

7.5 Granger Causality Test

The granger causality test reject the null hypothesis if the p-value is more than 5 and 10 per cent otherwise fail to reject the null hypothesis if the p-value is greater than that per cent. The result in Table 5 indicates that base lending rate granger cause to average house price. Meanwhile output of real estate subsector was granger cause to base lending rate.

Table 5: Granger Causality Test Result

Null Hypothesis:	Obs	F-Statistic	p-value
Base lending rate does not Granger Cause **Average house price**	5	14.548	**0.062
Average house price does not Granger Cause **Base lending rate**		5.308	0.148
Loan approved does not Granger Cause **Average house price**	5	0.019	0.904
Average house price does not Granger Cause **Loan approved**		0.847	0.455
Output of real estate does not Granger Cause **Average house price**	5	1.308	0.371
Average house price does not Granger Cause **Output of real estate**		7.525	0.111
Loan approved does not Granger Cause **Base lending rate**	5	6.530	0.125
Base lending rate does not Granger Cause **Loan approved**		0.182	0.712
Output of real estate does not Granger Cause **Base lending rate**	5	19.581	*0.048
Base lending rate does not Granger Cause **Output of real estate**		0.154	0.733
Output of real estate does not Granger Cause **Loan approved**	5	0.107	0.774
Loan approved does not Granger Cause **Output of real estate**		0.411	0.587

**Represent 5% significance*

***Represent 10% significance*

7.6 Diagnostic Checking

To make sure that the stated result is free from spurious inference, the competency of the model specified is further verified through diagnostic tests. The result in Table 5 pointed out that the null hypothesis of no serial correlation, hoterocedasticity and as well as normality of the distribution of the residuals were fail to reject. Therefore, we can conclude that the model has passed for the diagnostic test.

Table 6: Diagnostic Checking Result

Test statistics	**F-statistics**	**Probability**
Autocorrelation	0.000	0.9819
Normality	1.5799*	0.4539
Heterocedasticity	0.3835**	0.7793

*Jacque-bera
** Breush-Pagan-Godfrey

8. Discussion and Conclusion

Malaysia's real estate market is a significant contributor to the country's gross domestic product. This study aims to identify the factor affecting real estate subsector performance using a proxy of real estate gross output value. From the findings, it revealed that loan approved and base lending rate has negative effect to the gross output value, while the changes of average price house have positive effect to the output. In terms of causality relationships, the results found that there are have relationship between base lending rate and average price house and output of real estate with base lending rate.

References

1. Chia, J., Harun, A., Wahid, A., Kassim, M., Martin, D., & Kepal, N. (2016). Understanding Factors That Influence House Purchase Intention Among Consumers In Kota Kinabalu : An Application Of Buyer Behavior Model Theory, *03*(02).
2. European Communities, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and World Bank*. System of National Accounts 2008*. (2009).
3. Evans, Oteng. (School of Business, Valley View University-Techiman Campus). (2014). The Impact of High Lending Rates on Borrowers' Ability to pay Back Loans in the Tamale Metropolis
4. JPPH. (Valuation and Property Services Department), (2018). Harga Kediaman Sukuan Terkini The Residential Prices Quarterly Update, (7).
5. Maclennan, D. (2002). The Review of Scotland's Cities. Edinburgh: Her Majesty's Stationery Office
6. Malaysia, DOSM (2016). *Economic Census 2016 - Real Estate Services*.
7. Sean, S. L., & Hong, T. T. (2014). Factors Affecting the Purchase Decision of Investors in the Residential Property Market in Malaysia, *5*(2), 1–13.
8. The Star., https://www.thestar.com.my/business/business-news/2017/07/21/banks-high-debts-insufficient-income-main-factors-for-housing-loan-rejections/#tiE8Dgc2kSj7mzC8.99
9. Wilkinson, S. J. (2008). Property Development, (May 2014). Available from: https://www.researchgate.net/publication/255787306_Property_Development
10. Yoong, L. S. (Universiti T. A. R., Ling, S. S., Yun, T. C., Fang, T. S., & Cong, T. W. (2016). Determinants of Housing Price In, (April).

Spatial Demographic Analysis at the Sub-National Levels

Tey Nai Peng[1], Datin Rozita Talha[2], Ezatul Nisha Abdul Rahman[2], Muhamad Fadzil Ismail[2]

[1] University of Malaya

[2] Department of Statistics Malaysia

Abstract

While there is a rather sizable literature on the demographic dynamics in Malaysia, spatial demographic analysis is lacking, although space is a crucial element for demographic studies. A better knowledge of the demographic dynamics at the sub-national levels is essential for planning in terms of resource allocation and provision of infrastructure and services. This paper provides an illustrative spatial demographic analysis using data from Malaysian population censuses, and vital statistics. The paper deals with population growth and distribution/concentration, fertility, mortality, pupil-teacher ratio, and population ageing. It is hoped that this paper will provide inputs for the 2020 population census, stimulate interest in spatial demography, which will result in more e effective utilization of demographic data for development planning.

Keywords

Demographic Dynamics; Sub-National Levels; Population Censuses; spatial

Introduction

There is a rather sizable literature on Malaysia's demographic dynamics (Chandran et al. 1977; Sidhu and Jones, 1981; Arshat et al., 1988; Saw, 2007; Leete, 1996; Tey et al., 2015). Space is a crucial element in demographic studies, and a good knowledge of the spatial demography is needed for planning (Marcia Caldas de Castro, undated). However, demographic analysis at the sub-national levels is lacking. The few spatial demographic analyses reveal wide differentials in the demographic dynamics and the health conditions across the districts (Tey, Tan and Arshat, 1985; Nuzlinda and Syerrina, 2012; Ling, et al., 2014; Azreena et al., 2016).

In keeping with the increasing demand for Small Area Statistics (SAS), DOSM has been publishing SAS such as the decennial population censuses, annual vital statistics, state/district data bank, state/district social statistics, wholesale and retail trade and other statistics by district, and these are available in the form of interactive database. It is hoped that the availability of SAS will encourage and enhance spatial demographic analysis.

This paper is an exploratory and illustrative study with the aim to stimulate spatial demographic analysis and more effective use of demographic data for policy making. It deals with district level analysis on population density, population growth and distribution/ concentration, fertility and mortality, pupil-teacher ratio in the secondary school, and population ageing.

The main sources of data for this paper came from the published reports of the population censuses, vital statistics reports, and state/district social statistics report. Simple tabulations, scatterplots and maps were used to present the findings.

Population Density and Rate of Population Growth

Malaysia has a population density of about 100 persons per square kilometer in 2010, and this ranged from 19 persons per square kilometer in Sarawak to 6,891 in the Federal Territory of Kuala Lumpur. In Peninsular Malaysia, besides Kuala Lumpur, Timur Laut and Petaling are the two most densely populated districts (with a population density of 4.330 and 3,012 respectively). On the other hand, Gua Musang, Jerantut, Lipis and Ulu Perak have the lowest population density of around 13- 17 persons per square kilometer.

Between 2000 and 2015, the rate of population growth was estimated to range from 1.2 percent per annum in Perak to 2.6 percent in Selangor. The variation in the rate of population growth is even more striking across the districts, ranging from -1 percent in Jempol to a high of 6 percent in Sepang during the inter-censal period 2000-2010. Figure 1 shows that most of the densely populated districts have high rate of population growth, and this will result in further concentration of population in areas of rapid growth and aggravating regional inequality.

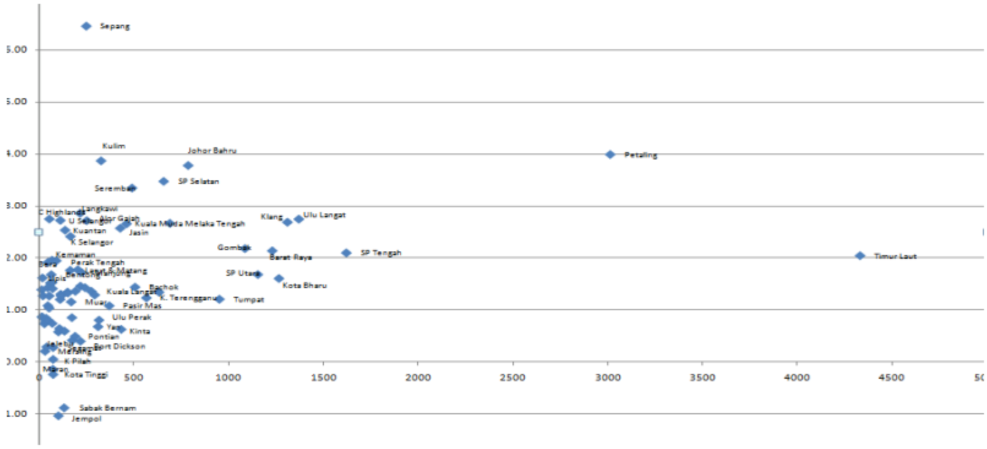


Figure 1: Population density and rate of population growth, by district
Sources: DOSM, Basic Population Characteristics by Administrative Districts, 2010; and Population Censuses, 2000, 2010

The population of Petaling district increased almost five folds from 360 thousand in 1980 to 1.77 million in 2010, at a rate of 5.3% per annum, and it has taken over from Kuala Lumpur as the most populous district. The population of Johor Bahru increased more than three-fold between 1980 and 2010, making it the third most populous district. Ulu Langat, the ninth most populous district in 1980 witnessed the most rapid population growth at 6.2% per annum over this period, to leapfrog into the fourth place in 2010. With a population growth rate of 3.7% per annum over the three decades, Klang remained the fifth most populous district in 2010. Although Kinta remained the sixth most populous district in 2010, its population has been growing much slower than many other districts.

More than a third (or 34.4%) of the national population now lives in the ten most populous districts: Petaling, Kuala Lumpur, Johore Bahru, Ulu Langat, Klang, Kinta, Gombak, Kuching, Seremban and Timur Laut. In the National Physical Plan, the Federal Department of Town and Country Planning (2010) projected a population of 10.37 million, 2.42 million, 2.40 million and 1.38 million for Kuala Lumpur, Georgetown1 (Pulau Pinang), Johor Bahru and Kuantan respectively in 2020. The combined total population of 16.57 million in these four conurbations will make up about 70 percent of the urban population or 60.4 percent of the total population in Peninsular Malaysia.

What are the reasons for the phenomenal population growth in these districts? Petaling and the other four districts in the Klang Valley have attracted migrants from all over the country to take up jobs in administration, commercial, financial, industrial, educational and services sectors. The various economic policies and programmes require the relocation to the cities, and these have led to the dramatic population redistribution over the last few decades.

Eight districts have experienced depopulation between 1970 and 2010- Sabak Bernam, Temerloh, Julau, Betong, Dalat, Sri Aman and Hilir Perak. Sabak Bernam and Temerloh registered the heaviest decrease of more than 50,000 persons each over the 30 years between 1980 and 2010, at a rate of -2.7%, and -1.0% per annum respectively, and the decrease was highest between 1991 and 2000, at a rate of -6,6%, and -3.6% per annum respectively. However, there was a reversal in the trend in Temerloh since then, as the population grew at 1.7% per annum between 2000 and 2010. Julat and Betong, both in Sarawak had a contrasting demographic trend - the population of Julat decreased by half between 2000 and 2010 but that of Betong increased during since 1991 after registering a sharp decline between 1980 and 1991.

Births and Deaths

The fertility rate and mortality rate in Malaysia have fallen to a very low level. However, wide variations in the fertility and mortality rates persist across regions and sub-groups of the population. Information on the number of births in small geographical areas is needed for short-term and medium terms planning in the provision of healthcare and educational services and facilities.

The 2017 Vital Statistics Report shows that the CBR ranged from 4.5 per thousand population in Kinabatangan to 26.4 in Kuala Terengganu, while the CDR ranged from 1.0 per thousand population in Kinabatangan to 9.1 in Kanowit. Interestingly, four of the five districts with the highest CBR are in Terengganu, and Kinabatangan has registered the lowest CBR and CDR in the country. The high CBR districts have a relatively young age structure, while three of the five districts with the lowest CBR have a relatively older age structure. In the case of Timur Laut, the demographic dominance of the Chinese also contributed to the low CBR as the community in general has attained ultra-low fertility. High CDR districts have a high proportion of older population aged 60 and over, while the reverse is true for the low CDR districts. This clearly shows that the CBR and CDR are affected by the age structure of the population. The extremely low level of CBR and CDR in some of these districts must be interpreted cautiously, as these extreme values could be due to under-registration or mis-reporting, especially in the remote areas in Sabah and Sarawak. There is a need for an evaluation of the extent of under-reporting in these remote areas.

As recently as 2015, the district-level CBR and CDE were not presented in the vital statistics reports. However, the child-woman ratio by district can be estimated using data from the population censes to provide an indicator of the fertility level for each district. A comparison of the child-woman ratio with the period measures of CBR and total fertility rate by district may throw some light on the reliability and validity of the district-level fertility rates based on the vital registration system.

Table 1: Districts with the highest and lowest crude birth rate and crude death rate, 2017

		Highest		Lowest
Crude birth rate	Kuala Terengganu (8.3, 24)	26.4	Kinabatangan (1.3, 23)	4.5
	Marang (8.2, 24)	26.0	Putatan (3.9, 24)	5.6
	Julau (12.5, 26)	24.7	Pakan (12.6, 28)	6.4
	Besut (7.9, 21)	24.5	Hilir Perak (12.6, 28)	8.3
	Setiu (8.3, 20)	24.0	Timur Laut (13.0, 33)	8.4
Crude death rate	Kanowit (15.6, 30)	9.1	Kinabatangan (1.3, 23)	1.0
	Sabak Bernam (11.9, 26)	8.8	Samarahan (5.0, 22)	2.1
	Kuala Pilah (14.5, 30)	8.7	Kunak (2.9, 24)	2.1
	Pendang (12.6, 27)	8.7	Belaga (6.3, 24)	2.4
	Kuala Kangsar (14.2, 28)	8.5	Tongod (3.3, 18)	2.4

Sources: DOSM, Vital Statistics Report, 2018

Note: Figures in parenthesis indicate the proportion of population aged 60 and over and the median age of the population)

While the rates are commonly used in spatial and temporal analyses, the actual numbers may be more relevant for planning purposes. The most recent vital statistics report show that a large number of babies were added some districts, as shown in Table 2. Such information is needed by the educational planners to prepare for the human resources and school facilities to cater for the new school entrants, as new-born reach the school-going age in the near future.

Table 2: Districts with the highest number of births, 2016-2017

	2017		2016	
	Number	CBR	Number	CBR
Malaysia	508685	15.9	508203	16.1
Petaling	30044	14.2	32085	15.4
Johor Bahru	25750	16.5	25003	16.3
Kuala Lumpur	21732	13.8	25739	14.4
Ulu Langat	21684	16.1	21953	16.6

Source: DOSM, 2018. Vital Statistics Report, Malaysia

Pupil-teacher Ratio

The pupil-teacher ratio is commonly used as an indicator of education quality. This section uses the pupil-teacher ratio in secondary schools for illustrative purposes. In 2013, the pupil-teacher ratio in secondary school ranged from 9.7 in Putrajaya to 15.1 in Selangor. The spatial differential in pupil-teacher ratio was even wider across districts, ranging from 7.4 in Maran and around 10 in Putrajaya, Beaufort, Port Dickson and Kuala Pilah to around 16 in Klang and Gombak, Ulu Langat and Bau (Table 3). Districts with high population density and rapid population growth tended to have higher pupil-teacher ratio. Hence, more teachers are required in states/districts with high pupil-teacher ratio in order to achieve the standard of 10:1 in the developed countries.

Table 3: States/districts with the highest and lowest pupil-teacher ratio in secondary schools, 2013

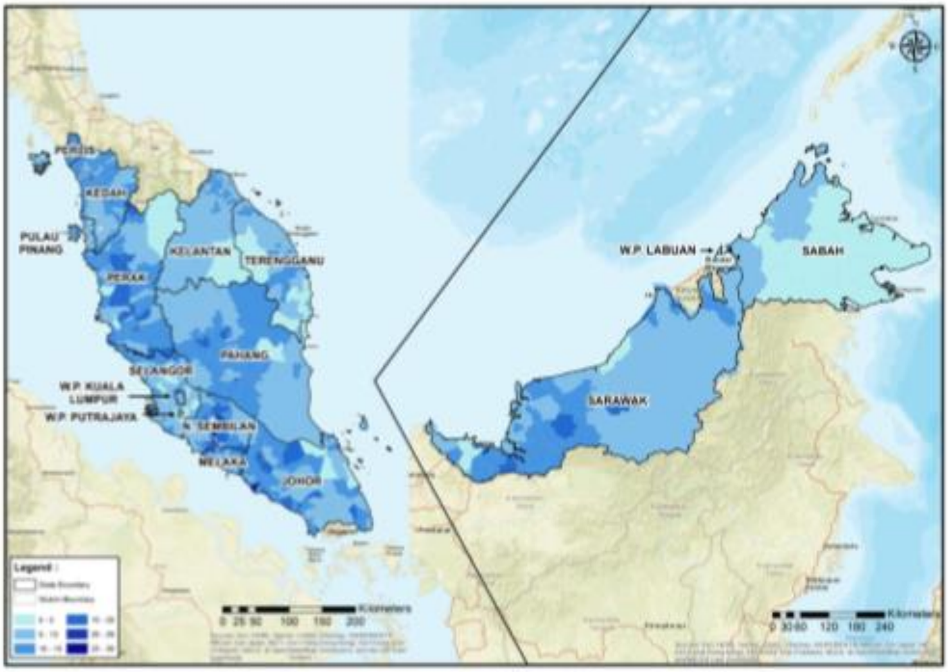
		Highest		Lowest
State	Selangor	15.1	Putrajaya	9.7
	Sabah	13.7	Labuan	10.5
	Pulau Pinang	13.6	Perlis	11
	Sarawak	13.6	Pahang	11.2
	Kedah	13.3	Terengganu	11.8
District	Klang	16.1	Maran	7.4
	Gombak	16.0	Putrajaya	9.7
	Ulu Langat	16.0	Beaufort	9.9
	Bau	15.6	Port Dickson	10.1
	Patatan	15.3	Kuala Pilah	10.1

Source: DOSM: State/District Social Statistics, Malaysia, 2013

Population Ageing

Consequent upon the continuing fertility decline and gain in life expectancy, the Malaysian population is ageing rapidly. Malaysia will become an ageing nation in 2030 when 15% of the population will be aged 60 and over. In 2010, two districts and 98 mukim were already classified as having an aging population. and this number has probably increased to about 12 and more than 200 respectively in 2019. Out-migration of the youths has exacerbated population ageing in the less developed areas. Geographic information on the distribution of the older people and their profiles is crucial for the provision of goods and services as well as public amenities to those in need. There is also a need to provide older people the opportunities for them to continue their active engagement in the society.

Map on population ageing



Sources: Population and Housing Census 2010, DOSM

Discussion and conclusion

Population affects development and is in turn affected by development. Hence, a good knowledge in spatial distribution of the population should be an integral component of EIA and SIA of all the mega projects (e.g. the East Coast Rail Link). Population mobility and redistribution generally result in a more efficient utilization of human resources by moving surplus labour from one region to another region where there is a labour shortage. But these processes also aggravate regional inequality. In its efforts to bring about a more balanced regional development, the Malaysian government has developed the five development corridors, and implemented other strategies. However, as the population continues to gravitate towards the central region, the effectiveness of these development corridors in population redistribution needs to be evaluated.

While the fertility level in Malaysia has fallen below replacement level, high fertility still persists in certain localities where family planning practice is at a low level and unmet need for contraception is high. In these localities, family planning efforts need to be stepped up to provide couples the necessary information and services for them to exercise their reproductive rights, to have the optimum number of children. The reasons for the higher mortality rate in some groups and localities need to be examined and measures taken to reduce the high death rate for these groups.

An in-depth analysis of the spatial distribution of the target groups is required for the allocation of resources to meet the needs of specific target groups such school-going children, the poor and the elderly. Multivariate analyses are needed to determine the covariates and the confounding factors.

There is a need for updating spatial demographic and social data, and the local leaders should be involved in the data collection, analysis and utilization. The 2020 population census provides an opportunity for updating the relevant spatial information for policy making and planning.

References

1. Arshat, H., Tan, B. A., Tey, N. P., & Subbiah, M. (1988). Marriage and Family Formation in Peninsular Malaysia: Analytic Report on the 1984/85 Malaysian Population and Family Survey. Kuala Lumpur:
2. National Population and Family Development Board.78). A Framework for Analyzing the Proximate
3. Chander, R., Palan, V. T., Aziz, N. L., & Tan, B. A. (1977). Malaysian Fertility and Family Survey: First Country Report. Kuala Lumpur: Department of Statistics and National Family Planning Board.
4. Federal Department of Town and Country Planning. (2010). National Physical Plan-2. Putrajaya.
5. Leete, R. (1996). Malaysia's Demographic Transition: Rapid Development: Culture, and Politics. Kuala Lumpur: Oxford University Press.
6. Ling CY, Gruebner O, Kramer A, and Lakes T. (2014). Spatio-temporal patterns of dengue in Malaysia: combining address and sub-district level.Geospat Health. 2014 Nov;9(1):131-40.
7. Marcia Caldas de Castro (undated). Spatial Demography: an opportunity to improve policy making at diverse decision levels, Department of Geography, University of South Carolina
8. Nuzlinda Abdul Rahman, and Syerrina Zakaria (2012). The Household-Based Socio-Economic Index for Every District in Peninsular Malaysia International Scholarly and Scientific Research & Innovation 6(10) 2012
9. Saw, S. H. (2007). The Population of Malaysia. Singapore: Institute of Southeast Asian Studies (ISEAS) Publishing.
10. Sidhu, M. S. and Jones, G. W. (1981), Population Dynamics in a Plural Society: Peninsular Malaysia, Kuala Lumpur: University of Malaya Press.
11. Tey Nai Peng, Tan Boon Ann and Hamid Arshat (1985) Multivariate Areal Analyses of Neo-natal Mortality in Peninsular Malaysia, Malaysian Journal of Reproductive Health, June, Vol. 3 No. 1, pp. 46-58.
12. Tey, N. P., Lai, S. L., & Lee, M. (2019). Population Redistribution and Urbanization in Malaysia, 1970-2010. In W. Y. Lau (Ed.), Statistics in Practice. Kuala Lumpur: University of Malaya Press.

Outlier Detection in Official Statistics

Nadiah Mohamed[1,2], Adzhar Rambli [3], Ibrahim Mohamed[1]

[1] Institute of Mathematical Science, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Negeri Sembilan, 72000 Kuala Pilah Negeri Sembilan, Malaysia

[3] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA 40450 Shah Alam, Selangor, Malaysia

Abstract

Distance-based outlier detection is one of outlier detection techniques for deterministic data. Sliding window method is one of the various streaming techniques to keep track of the most recent data where all mining tasks are performed based on what is “visible” through the window. So far, the use of outlier detection in official data has not been fully utilized yet, though its potential in improving estimation and forecasting are highly useful. Besides, the procedure is also able to identify abnormal pattern or trend in such large data. In this study, we intend to develop new outlier detection procedure which can perform both purposes above on streaming case of official data, in particular, the multidimensional and complex Malaysian economic data.

Keywords

Outlier detection; Official statistic; Sliding window; Water quality data

1. Introduction

Outlier detection for temporal data can be divided into five types that are time series data, data streams, distributed data, spatiotemporal data and network data (Gupta, Gao, & Aggarwal, 2013). This research focuses on water stream data that classifies under data stream. There are some techniques that can be used in order for us to detect outlier in data stream such as distance-based outlier detection, density-based outlier detection, clustering based outlier detection, statistical based outlier detection, frequent pattern mining based outlier detection, classification-based outlier detection and angle-based outlier detection (Souiden, Brahmi & Toumi, 2016).

Streaming data does not have a fixed length compared to static data. Streams can be either multidimensional or time-series. Yamanishi, K., & Takeuchi, J. I. (2002) introduced SmartSifter as a program to compute online-unsupervised outlier detection. Online discounting learning algorithm is used in order to learn about the probabilistic mixture model. However, there is no adjustment made for incremental updates and temporal decay for conventional multidimensional data. Angiulli & Fassetti (2007) used Indexed Stream Buffer (ISB) that support a range query. This exact algorithm of new data structure can compute distance outliers efficiently. Besides that, they also introduced approximate algorithm where they maintain a fraction of safe inliers in ISB and include the fraction of preceding neighbours which is also identified as safe inliers to the total number of safe inliers. Yang, Rundensteiner, & Ward, (2009) propose an efficient algorithm that uses predicted views to calculate distance-based outliers. This "predicted views" can skip the step of maintaining all neighbour relationships across time and maintaining cluster of abstracted neighbour relationships that are expensive. Yang et al., (2009) used dynamic cluster maintenance to the problem of distance-based outlier detection for stream data.

Areas such as electronic commerce, credit card fraud, and even the analysis of performance statistics of professional athletes can lead us to the discovery of unexpected knowledge when dealing with finding outliers (exceptions) in large, multidimensional datasets. The notion of DB- (Distance-Based) outliers and development of cell-based algorithms for computing such outliers by Knorr & Ng (1998) is the best for k ≤ 4, where k is value of dimensional datasets. Efficient Nested-Loop version (ENL) and its Parallel Nested-Loop version (PENL) introduced by Hung & Cheung (1999) shows great results that it is a very good choice to mine outliers in a cluster of workstations with a low-cost interconnected by a commodity communication network. Besides that, Ramaswamy et al. (2000) also developed a highly efficient partition-based algorithm in order to determine very quickly significant number of the input points that cannot be outlier. Micro-cluster-based local outlier mining algorithm introduced by Jin, W., Tung, A. K., & Han, J. (2001) compresses the data and used cut-plane solution for overlapping data.

2. Methodology

Distance-based outlier detection is one of outlier detection techniques on deterministic data. Distance-based outlier detection considers a point as an outlier of a dataset if the number of points within a certain distance from it is below a given threshold (Wang, Yang, Wang, & Yu, 2010). A new definition of distance-based outlier on uncertain data stream given by Wang, Yang, Wang, & Yu, (2010) maintains the basic idea of the traditional definition and employ probability. A dynamic programming algorithm (DPA) is proposed where it can process each single element in linear time, avoiding expensively unfolding the possible worlds of its neighbourhoods. A pruning-based approach (PBA) is also proposed by Wang, Yang, Wang, & Yu, (2010) to effectively and efficiently reduce the processing elements in the sliding window and save detection cost. Outlier detection in big data set up is a data-mining task focusing on the discovery of objects, called outliers that do not seem to have the characteristics of the general population (Kontaki et al., 2016). One of the most widely used definitions of outlier is the one based on distance: an object x is considered as an outlier, if there are less than k objects in a distance at most R from x, excluding x itself. Otherwise, x is characterized as an inlier.

Kontaki et al., (2016) stated that the fundamental characteristic of the majority of the proposed algorithms are operating in a static fashion. The algorithm must be executed from scratch if there are changes in the underlying data objects, leading to performance degradation when updates are frequent. Kontaki et al., (2016) focuses on sliding window method that is one of the various streaming techniques. Since the stream is continuously updated with fresh data, it is impossible to maintain all of them in main memory. Therefore, a window is used where it keeps track of the most recent data and all mining tasks are performed based on what is "visible" through the window. As reported in Gupta et al., (2013), most window-based models are currently offline. The most relevant research works are Angiulli & Fassetti (2007) and Yang, Rundensteiner, & Ward, (2009) where both considered the problem of continuous outlier detection in window-based data streams, without limiting their techniques to multi-dimensional data. However, both methods still have some serious limitations.

3. Result

In this research, we use water quality data that provides information of Dissolved Oxygen (DO) and Biochemical Oxygen Demand (BOD). Figure 1 shows the steps that are used to identify the outlier or inlier of DO and BOD. We use Euclidean distance formula to find the distance for each point of the data by using R Software to identify the outlier and inlier and the result may vary depending on the value of members within the window (W), radius (R) and number of neighbour (k). The value of k=3 and R=4 are used based on (Kontaki et al., 2016) and the value W is set to be 10. Figure 2 shows an example of 1-sliding window on a probabilistic data stream for window 1 and 2. Table 1 shows the result for each window. In window 1, point 4 is not a safe inlier because it became an outlier in window 2. However, all inlier in window 1 is a safe inlier because it still remains inlier in window 2. For window 4, point 4 in Figure 3 is an outlier because it has three neighbours. However, in window 5 in Figure 4, point 4 is an inlier because it has five neighbours.

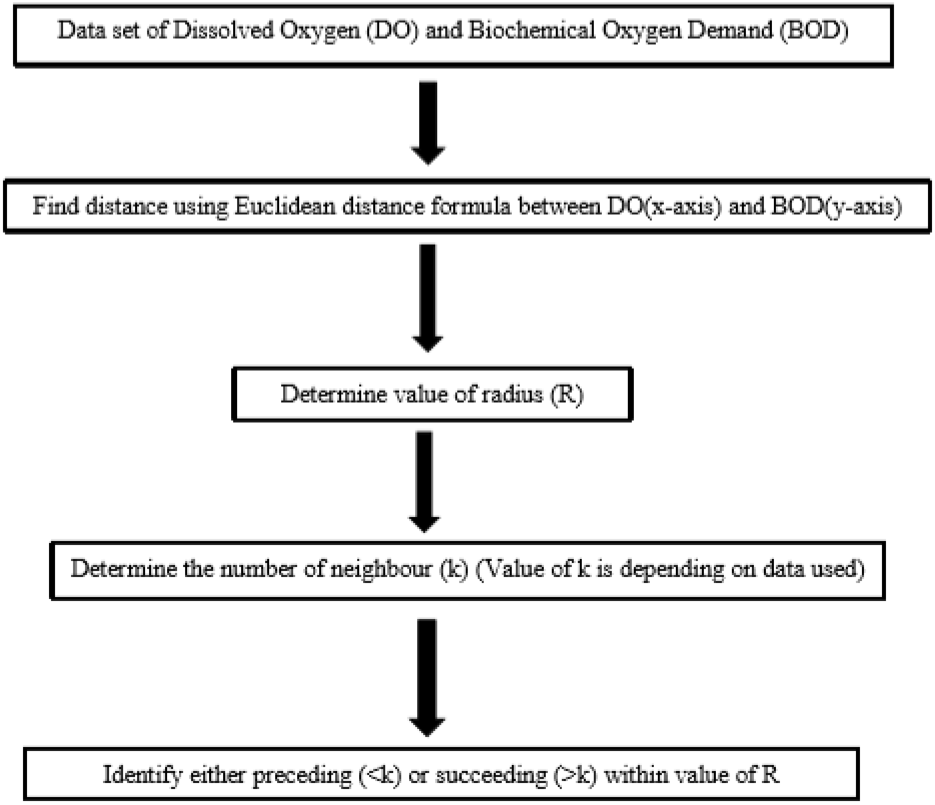
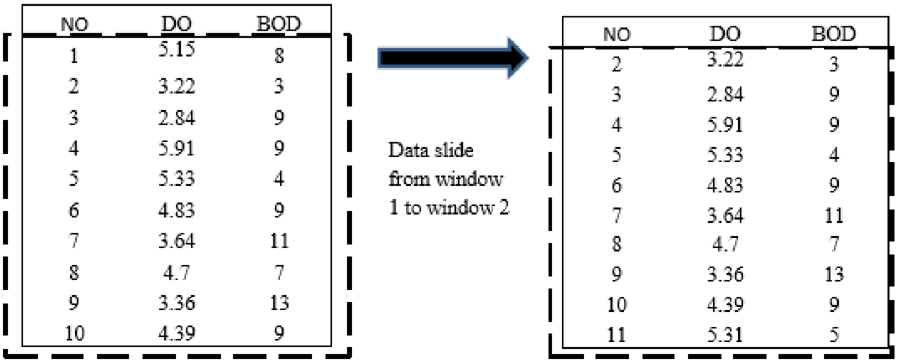

Data set of Dissolved Oxygen (DO) and Biochemical Oxygen Demand (BOD)
Find distance using Euclidean distance formula between DO(x-axis) and BOD(y-axis)
Determine value of radius (R)
Determine the number of neighbour (k) (Value of k is depending on data used)
Identify either preceding (<k) or succeeding (>k) within value of R

Figure 1: Step to identify outlier or inlier in data set



NO	DO	BOD
1	5.15	8
2	3.22	3
3	2.84	9
4	5.91	9
5	5.33	4
6	4.83	9
7	3.64	11
8	4.7	7
9	3.36	13
10	4.39	9

NO	DO	BOD
2	3.22	3
3	2.84	9
4	5.91	9
5	5.33	4
6	4.83	9
7	3.64	11
8	4.7	7
9	3.36	13
10	4.39	9
11	5.31	5

Figure 2: An example of a 1-sliding window on a probabilistic data stream

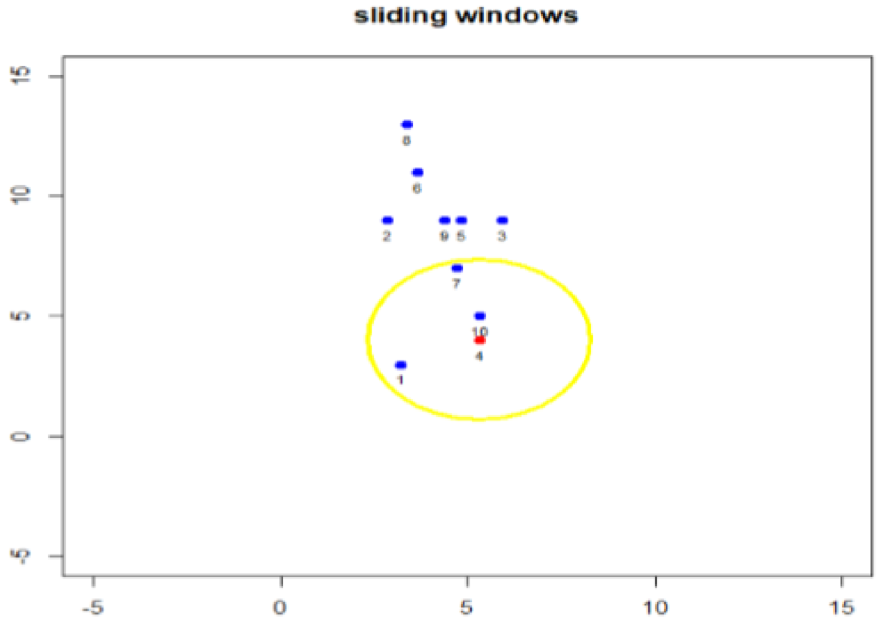


Figure 3: Plots for window 4

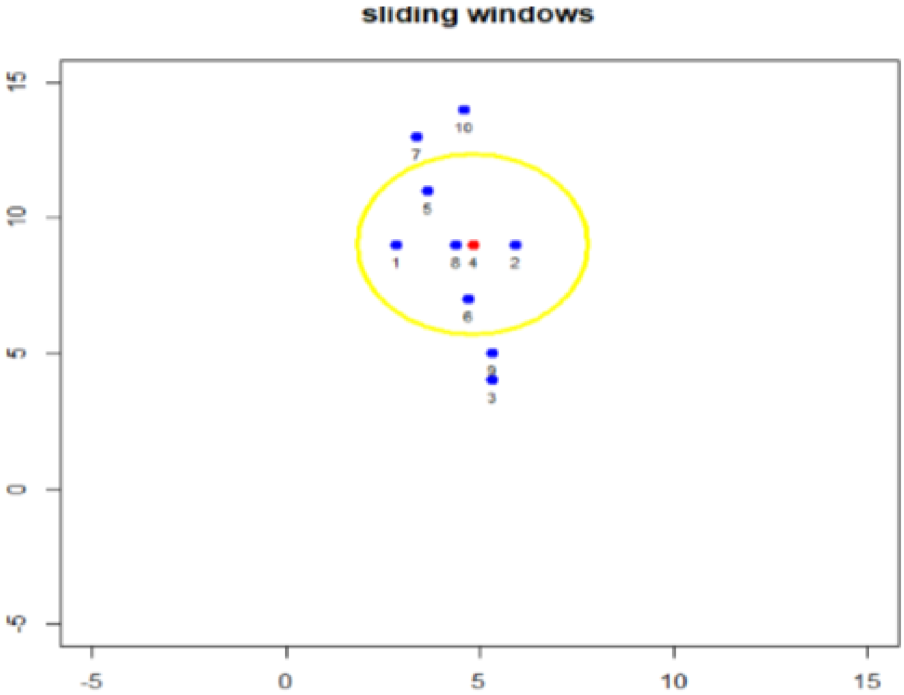
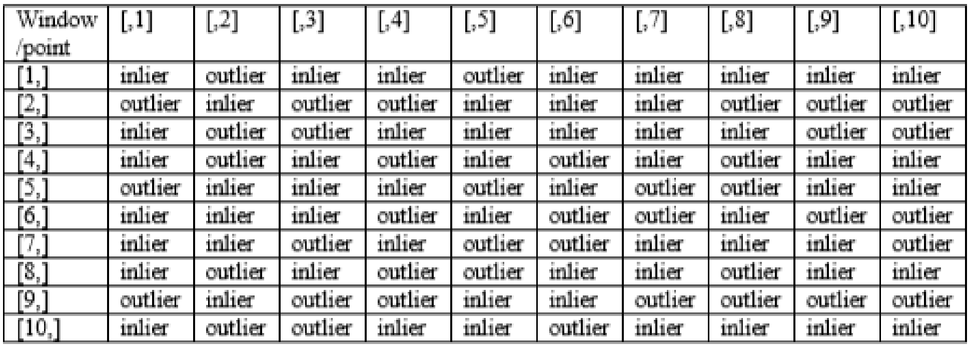


Figure 4: Plots for window 5

Table 1: Result of sliding windows



Window /point	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	inlier	outlier	inlier	inlier	outlier	inlier	inlier	inlier	inlier	inlier
[2,]	outlier	inlier	outlier	outlier	inlier	inlier	inlier	outlier	outlier	outlier
[3,]	inlier	outlier	outlier	inlier	inlier	inlier	inlier	inlier	outlier	outlier
[4,]	inlier	outlier	inlier	outlier	inlier	outlier	inlier	outlier	inlier	inlier
[5,]	outlier	inlier	inlier	inlier	outlier	inlier	outlier	outlier	inlier	inlier
[6,]	inlier	inlier	inlier	outlier	inlier	outlier	outlier	inlier	outlier	outlier
[7,]	inlier	inlier	outlier	inlier	outlier	outlier	inlier	inlier	inlier	outlier
[8,]	inlier	outlier	inlier	outlier	outlier	inlier	inlier	outlier	inlier	inlier
[9,]	outlier	inlier	outlier	outlier	inlier	inlier	outlier	outlier	outlier	outlier
[10,]	inlier	outlier	outlier	inlier	inlier	outlier	inlier	inlier	inlier	inlier

4. Discussion and Conclusion

In this work, we use sliding windows to study the problem of continuous outlier detection over data streams. As shown in the performance evaluation results, we can identify the outlier and inlier in the data depend on the value W, R and k that we choose. However, there are improvement needs to be done in order to get the suitable value of W, R and k to get the optimum result. There are several directions for future research. It is interesting to design outlier detections algorithms over uncertain data streams. A second direction is producer a better framework in processing data in official statistics.

References

1. Angiulli, F., & Fassetti, F. (2007). Detecting distance-based outliers in streams of data. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 811-820). ACM.
2. Angiulli, F., & Fassetti, F. (2007). Very efficient mining of distance-based outliers. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 791-800). ACM.
3. Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. Proceedings of the 2000 Acm Sigmod International Conference on Management of Data, 1–12.
4. Gupta, M., Gao, J., & Aggarwal, C. C. (2013). Outlier Detection for Temporal Data: A Survey. Ieee Transactions on Knowledge and Data Engineering, 25(1), 1–20.
5. Hung, E., & Cheung, D. W. (1999). Parallel algorithm for mining outliers in large database. In Proc. 9th International Database Conference (IDC'99), Hong Kong.
6. Jin, W., Tung, A. K., & Han, J. (2001). Mining top-n local outliers in large databases. In Proceedings of the seventh ACM SIGKDD international

conference on Knowledge discovery and data mining (pp. 293-298). ACM.

7. Knorr, E. M., & Ng, R. T. (1998). Algorithms for Mining Distance-Based Outliers in Large Datasets. Proceedings of the 24th VLDB Conference, 98, 392–403.
8. Kontaki, M., Gounaris, A., Papadopoulos, A. N., Tsichlas, K., & Manolopoulos, Y. (2016). Efficient and flexible algorithms for monitoring distance-based outliers over data streams. Information Systems, 55, 37–53.
9. Paper, C., Notes, L., Processing, B. I., & Souiden, I. (2017). Digital Economy. Emerging Technologies and Business Innovation, 290.
10. Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Record, 427–438.
11. Wang, B., Yang, X. C., Wang, G. R., & Yu, G. (2010). Outlier detection over sliding windows for probabilistic data streams. Journal of Computer Science and Technology, 25(3), 389-400
12. Yamanishi, K., & Takeuchi, J. I. (2002). A unifying framework for detecting outliers and change points from non-stationary time series data. In Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 676-681). ACM.
13. Yang, D., Rundensteiner, E., & Ward, M. O. (2009). Neighbor-based pattern detection for windows over streaming data. Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology - EDBT '09, Newyork, 529–540.

The ultimate comprehensive statistical frame: busting the myth

Hidzir Hamzah
Department of Statistics Malaysia

Abstract

In order to produce good statistics, a high quality and comprehensive frame is need so the data collected from samples really represent the whole population. Household (HH) frame is from the Population and Housing Census that conducted every 10 years. Whereas business establishment (EE) frame, records from business registry is being used as the main source. Both household and establishment frame is independent since both derived from a totally different source. This paper will provide a brief review on how both frames can complement each other and deriving an ultimate comprehensive statistical frame for each other.

Keywords

Household Frame; Business Establishment; Business Registry; Population; housing census

1. Introduction

Household (HH) frame is a comprehensive list of all the living quarters (LQ) and the demographic characteristics in it. Population and Housing census conducted every 10 years so that household frame is always representing the current situation. Enumeration block (EB) maps (GIS) are complementing the HH frame to visualize the listing for all the living quarters and its locality. Combining the HH Frame and GIS will produce a complete frame for Household. Business establishment (EE) frame on the other hand is using purely administrative data from the business registry. The entire establishment registered in the business register are included in the EE Frame. The operation status for the establishment registered whether it's still in operation and also the company performance report (for Register of companies) can easily be known through the business register website and database. Hence, business register is the most comprehensive and suitable source for the foundation of EE frame.

2. Limitations and barrier

a. HH frame

Complementing each other, HH frame and GIS supposedly will produce a near perfect list for the living quarters (LQ) in a specific location. Unfortunately, there are plenty of "hidden" LQ that missed out from the HH frame and can lead to under coverage. Due to skyrocketing house price and suitability of the location, business premises that old and already lost its commercial value usually will be turned into LQ. Since it's not practical to open up business in this kind of premises, owners decided to rent it as a house or warehouse rather than leaving it empty. Most of this premises usually rent by immigrants or employer providing hostel for their workers. This type of premises that are "hidden" and usually provide uncertainty since it can change from business premise to living quarters or to non-living quarters at anytime. Every year, there will always an increase since buildings getting older every single day and the price of houses always increase. On paper, this premise still registered and classified as business premises and is not included in the household frame and also during census.

b. EE Frame and Business Registers.

   i. Operation status of an establishment.

   The operation status of an establishment can simply be known through business register website. But this is only reliable for Register of Company (ROC) but not Register of Business (ROB). For ROB, an establishment can just register once and not renewing their registration but continue operating business as usual. Vice versa, it can also be in active status but the business is dormant. As we know ROB is usually defined as small business unit. But if the contribution of this establishment can reach to significant percentage of the total value, then this small business unit is really vital to the statistics

   ii. Change in address

   When an establishment decided to change their location of business, it is not compulsory for them to change the address in the business registry. Hence, leading to challenges determining the operation status and also the new business address.

   iii. Company Secretary

   For ROC, establishment using company secretary address for both business and registered address, it is a challenge to determine the location of the business since most company secretaries having problem revealing this information to the agency. Since establishment approach is used in EE frame, the location of the business is important information in deriving statistics.

   iv. Coverage

   Using business registry as the main source, the question on how thorough and complete is the coverage can never be answer. Having flaws in the registration system and not having a complete locality listing, we can never be sure the percentage of the coverage for business establishment is.

v. Small Area Statistics

Using business registry as the main source has lead to problem in assigning locality for establishments. Since Enumeration Block (EB) continuously change and there's no mechanism in checking the accuracy of the locality code assigned. Usually locality code for establishments is assigned with a dummy code. This will definitely make deriving small area statistics is merely impossible.

vi. Census operation planning

Without the locality codes in Para v., it is a huge challenge to plan for census operation since we won't be able to assign human resource according to area and locality. Even with the locality codes available, miscalculation of human resource will likely to happen due to the conflict number of cases between the database and reality on the field.

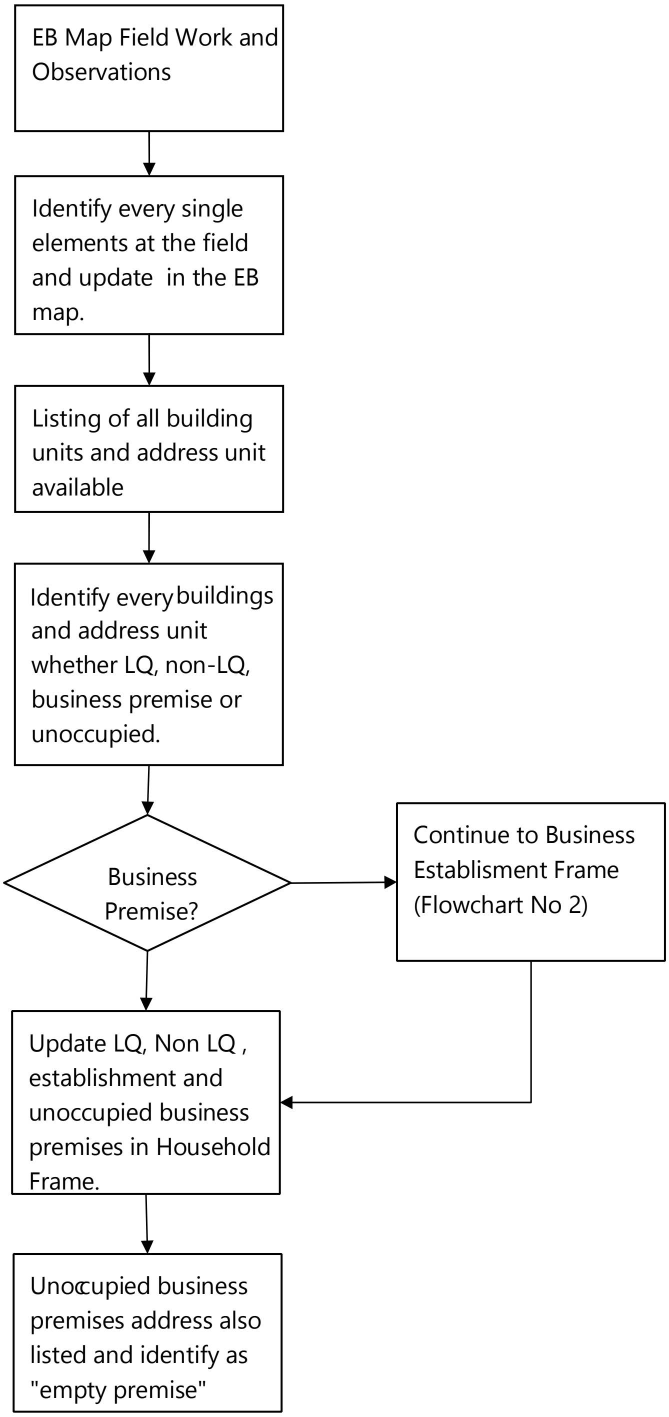
vii. Address for "unofficial" buildings/premises

Plenty of establishment operates in an unofficial premises that does not have a proper address (eg : permanent roadside stalls and container office . This will lead to missing coverage since there is no business address for this kind of establishment.

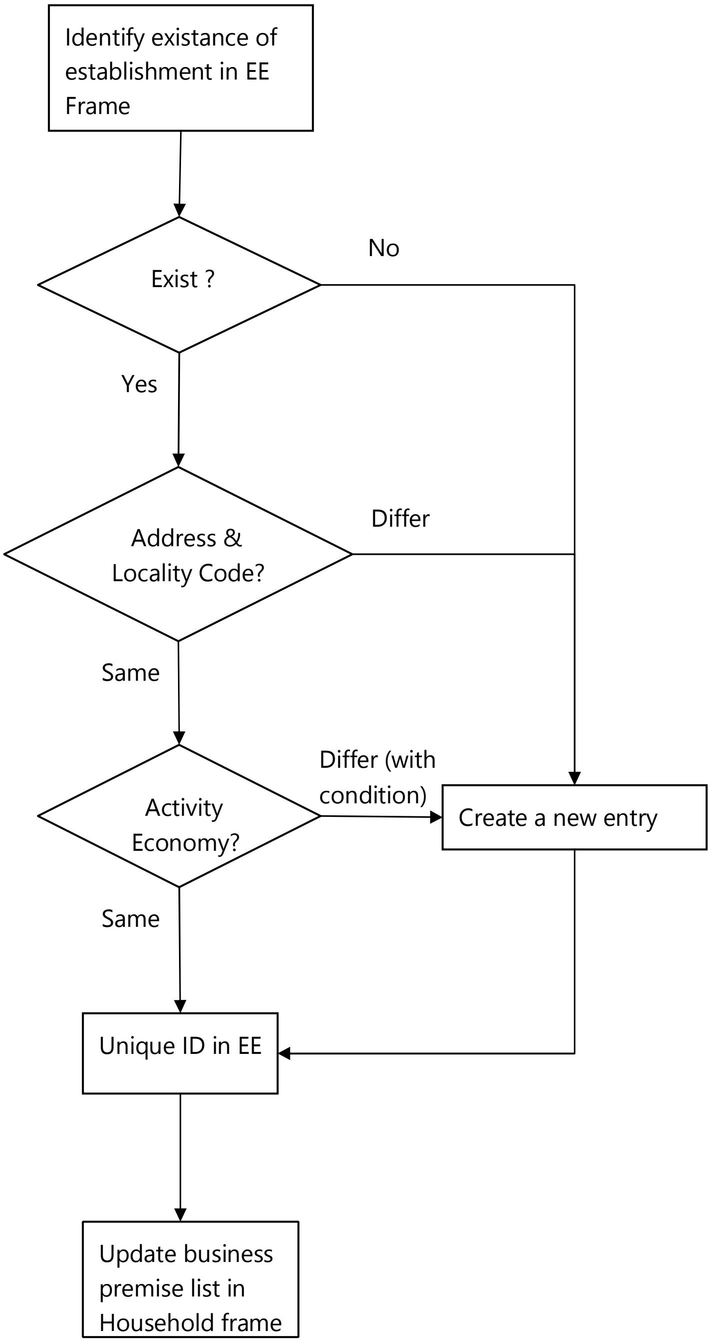
3. Methodology

In order to have an ultimate comprehensive frame for HH and EE, both frame cannot stand on their own. Both frames must complement each other to have what we call an ideal frame. For every EB map, steps that need to be done are visualize in Flowchart 1 and 2 below.

Flowchart 1 : Field Work, GIS and Household Frame


EB Map Field Work and Observations
Identify every single elements at the field and update in the EB map.
Listing of all building units and address unit available
Identify every buildings and address unit whether LQ, non-LQ, business premise or unoccupied.
Business Premise?
Continue to Business Establisment Frame (Flowchart No 2)
Update LQ, Non LQ , establishment and unoccupied business premises in Household Frame.
Unoccupied business premises address also listed and identify as "empty premise"

Flowchart 2 : Business Establishment Frame



i. Identify every elements in the map (buildings roads etc) so the map are updated and represent the reality.
ii. Identify every unit address and every unit address whether it is a LQ, non-LQ or business premises. Some units maybe consist both LQ and business premises. (eg; barbershop operate as business premise at front and as a LQ at the back).
iii. Unoccupied LQ and business premises also listed for provision in updating process in the future. This is also useful in planning for census operation (economic) to determine the actual workload since every single business premise has already been covered. This also helps in Population and Housing Census to determine the potential of LQ in business premises.
iv. When every elements and address unit has been listed and identify whether LQ, nonLQ or business premise, it can be said that 100% of coverage of LQ and establishment (with physical business premise) has already achieved.
v. For EE frame, 100% of establishment with physical premise has already been covered. This is where business registry and other administrative records will complement for the establishment without physical business premise. (eg: online retail from home, cleaning service, catering service). There will be 2 sets of database for EE frame, one from the business registry and the other one is from the business premises where these two datasets are connected with a unique ID.
vi. With this complete statistical frame foundation, changes and update in the EB map in the future will require changes made to both HH and EE frame. Hence, killing two birds with one stone

4. Impact and Advantages

By using this approach most of the limitations in HH (2a) and EE (2b) frame can be solved.

For HH frame, it can be said that 100% of the living quarters are in the coverage. Projection of the population and target during census can be easily determined whether possible to achieve with justification.

For EE frame, all the issues raised in Para 2b can be resolved as follows:

i. Operation status of an establishment can be easily determined by referencing to the address in business premise database.
ii. It can be said that 100% coverage for business establishment that operates under a physical premise.
iii. Changes in establishment address can be crosscheck with the business premise database.
iv. Exact location of an establishment with only company secretary address can be determined.

v. Small Area Statistics can be derived from the business premise listing even at EB unit level.

vi. A proper census operation planning. From the business premise listing, it is known the total of business premise and cases by locality. Allocation of human resource for census operation is fully utilized.

vii. GIS elements can be used for unofficial buildings address.

5. Challenges

A few challenges need to be addressed :

i. Time consuming Updating EB maps and make a complete list for all unit address available is really time consuming.

ii. Once only The assignment to update and list everything must complete in by one individual only. High turnover of employee will waste a lot of time and resource

iii. 3-in-1 In order to really make it a success, one must have broad knowledge the three elements in statistical frame which is Household, Establishment and GIS since there will be plenty of issues, treatment and solutions needed along the way.

6. Conclusion

An ultimate and comprehensive frame is plausible with the integration of the three elements, HH frame, EE frame and GIS. Even though on the surface, HH and EE frame is independent and mutually exclusive, but both frames can complement each other to provide a foundation in planning for census operation and an ultimate comprehensive statistical frame.

Food security/ insecurity and policy responses in Sri Lanka: A theoretical review

Anura Kumara
Department of Census & Statistics, Battaramulla, Sri Lanka.

Abstract

The research is based to scrutinize the policy responses for the food insecurity. It is being considered for fulfilling food for an active and healthy life for all people in a territory (FAO, 1996). If it gets in to the problem: "Why does inadequate food insecurity continue to be a policy challenge, despite the government responses the food insecurity?" it would be an issue for Sri Lanka. The welfare programmes for the Sri Lanka society are very high with a larger government involvement.

By the year 2030, also Sri Lanka has to achieve and full fill the sustainable development goals (SDGs). The problem of reaching targets in relation to nutritional status of children and women are shown in Sri Lanka.

According to the existing literature, institutional levels of population are needed to be examined for food insecurity. The research study is based on the food practices data of households to calculate into various numerical indicators which are commonly derived in different ways. For the calculation, it can be used existing data or timely collected data. It is very useful to achieve goals through given targets for indicators at different time points. Here it follows in food security sector studies by both research instruments such as quantitative and Qualitative methods parallel context.

Keywords

The Sustainable Development Goals (SDG's); Food Security Indices; The Global Environmental Change and Food System (GECAFS).

1. Introduction

"Food security exists when all people, at all times, have physical, social and economic access to sufficient, safe and nutritious food which meets their dietary needs and food preferences for an active and healthy life." (FAO, 1996)

That means the international attention has set for the food security of the world from the last century. With the country context, there is a problem of "Why does inadequate food security continue to be a policy challenge, despite the government responses the food insecurity?". And it would be an issue for Sri Lankan society.

Objective of the Study: It is to identify the new sound policy interventions for the Food Insecurity at Household levels of the Country strategically. And the broad objectives of the Study

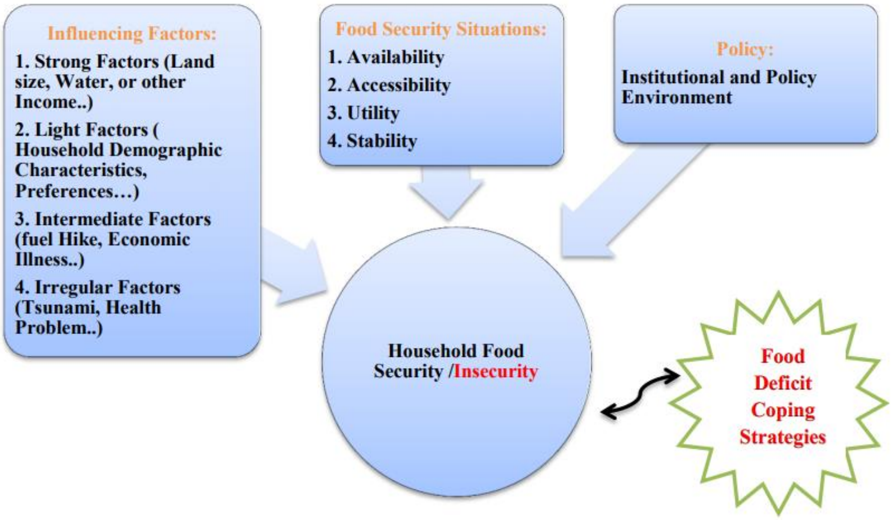
- To increase the quantity and quality of food available, accessible and affordable to all people at all times.
- To achieve good nutrition for optimum health of all people.
- To protect vulnerable populations using innovative and cost-effective safety nets linked to long-term development for all.

The global warming leads to the environment change and then it allows to the biodiversity. In backward, human activities for unsustainable returns, lead to the environment change and the global warming. (Ingram & Brklacich, 2002) The food is considered as the basic need for the people, it is very important matter to study further to achieve common goals as globally. By the year 2030, also Sri Lanka has to achieve and full fill the sustainable development goals (SDGs) in the context of United Nations common Agenda in 2015. Among those, SDG 2 is explained that "end hunger, achieve food security and improved nutrition and promote sustainable agriculture". (UNO, 2016) In the country context, there is a problem of reaching a lower phase and set targets for some indicators in health sector in Sri Lanka for some years. Particularly in relation to nutritional status of children and women are not shown a better picture. According to that wasting has increased 11.7 in 2009 to 19.6 in 2012. Based on this figure, Sri Lanka was ranked as having 3rd highest prevalence in the world only behind Djibouti and South Sudan from selected countries. (MRI, 2012) And presently also it is not shown bigger different on these indexes.

Even though the country has many positive indicators in many sectors, the food is a remarkable sub sector to be policy intervention/or developed in the context of healthy and secured food for all at all time. According to the Medical Research Institute of Sri Lanka, the recommendation of the minimum target for the food security is 2030 kcal and 53 grams per person per day in calorie and protein levels respectively. (MRI, 2012)

After I followed many scholarly articles in the sector in higher level forums, it has been designed to conceptualize the framework for this study as follows.

Figure 1 : Conceptual Frame work


Influencing Factors:
1. Strong Factors (Land size, Water, or other Income..)
2. Light Factors (Household Demographic Characteristics, Preferences…)
3. Intermediate Factors (fuel Hike, Economic Illness..)
4. Irregular Factors (Tsunami, Health Problem..)
Food Security Situations:
1. Availability
2. Accessibility
3. Utility
4. Stability
Policy:
Institutional and Policy Environment
Household Food Security /Insecurity
Food Deficit Coping Strategies

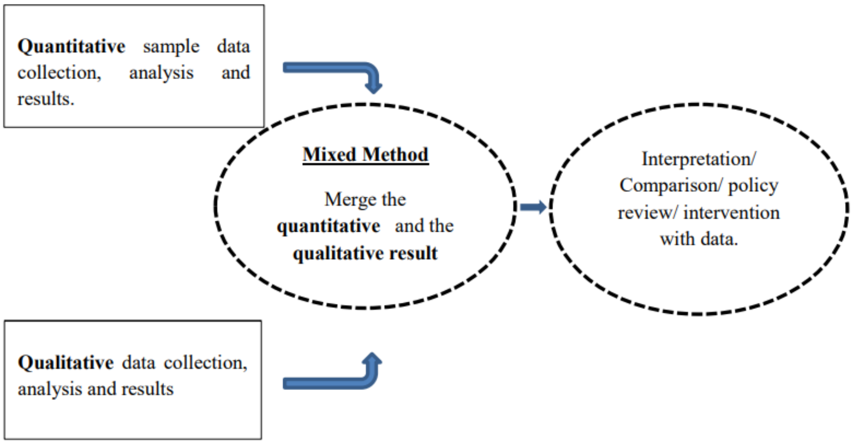
2. Methodology

According to the existing literature in the context of the food security, there are three levels of population to be examined in a country as following. (Deitchler, Ballard, Swindale, & Coates, Introducing a Simple Measure of Household hunger for Cross-Cultural Use, 2011)

1. National level,
2. Household level,
3. Individual level.

The study for food insecurity is measured using a quantitative questionnaire and other qualitative parallel ways. These strategies for methodologies of the PhD studies are recommended by various scholars. (Mukherjee, Hoare, & Hoare, 2002) According to general practice to measure the food security, it has managed to calculate numerical indicators which are commonly derived in many countries. In addition to that, the relevant group discussion and other qualitative methods are used to clarify the food insecurity in the country context by this paper.

Figure 2: The Mixed Method which is used for the Research study.


Quantitative sample data collection, analysis and results.
Mixed Method
Merge the quantitative and the qualitative result
Interpretation/ Comparison/ policy review/ intervention with data.
Qualitative data collection, analysis and results

A Survey Technique was developing to introduce the food insecurity level of the country with food poverty and the policy suggestions. (Ifeoma & Agwu, 2014). That discusses the survey module to go for the food security measuring for a territory. The method will be useful if it generalize to a country context. Since the survey application is to be generalized at all level of a country it is acceptable to apply to a certain country. Otherwise it will be vague effort.

As analytical tools the indices are playing very important role. For the calculation, it can be used existing data or timely collected data. Those are used to set goals through given targets through indicators. To analyze for a country food security, it is very important to identify suitable and handy indicators relevant to the country. Therefore, as researchers; we have own responsibility to select the most important and relevant as same as practically achievable indicators for the study. As global food sector indicators are explained by the Aspen Institute Food Security Strategy Group, it will cover a bigger study area. It starts from food and people then it ensures to a line through various socioeconomic activities using suitable policies. (Group, 2015) It is very common tool, the indices relevant to food security sector study. (Vhurumuku, 2014)

In addition to these quantitative accesses, the regular practices at household level should be studied in qualitative manner also according to my research knowledge. Therefore, it is included in the research methodology to collect information through the research instrument called case studies at the household levels.

Ethical consideration for the respondents

Food security study will be running through field work process since it is a measuring and calibrating a social and demographic practice in food sector consumption. That is bounded to many disciplinary practices. There are also three ethical concerns as it, a lot of information is collected from participants;

1. Protecting participant identity (Privacy)
2. Treating participants with respect (Sensitivity)
3. Protecting participants from both physical and psychological harm (Injury or pressure).

All the statistical sector workers are known and commonly practiced these activities in the worldwide during their day to day activities and for respecting the policies which have been set to develop the statistics field.

This should be carefully address by correct training for the enumerators and questions and information are designed to collect carefully to avoid such situation. Pre-informing letter is posted for the selected household and there concern is taken in written forms which have already designed.

Identified food policies in the country

The present food policies can be evaluated and some of them are direct and others may be indirect or intermediate involvements. Some of the major relevant programs are operationalized presently as below.

1. "Samurdhi" Programs are for the Low-Income Household. 2. "Thriposha" Programs are for the Child baring woman and early childhood (needed). 3. Mid-day Meal promoting program is for the Schooling cohort from lower grades. 4. Since 2009, World food program is for the supporting postconflict recovery interventions. All above categories cater commonly in post conflict areas. 5. Other indirect Programs (eg.: Concessions of fertilizer to farmers/ or giving technical knowledge/ encouraging farmers, and various concession packages are there for small entrepreneurs etc., also Free Health & Education helps and reserves to cover the cost for food indirectly.) Some of the policies are long term or short term or otherwise medium term according to their feature for addressing.

3. Discussion and conclusion

In the theoretical context, it is needed to identify the thrust areas related to the Food policy. The strategic thrust areas with the food policy are identified shortly as follows.

1. Food availability and accessibility through arrangements in inputs to food production and strengthening the individual level economic power.

2. Food safety and quality control and assurance for each and every parties and levels with knowledge sharing system or through business best practices.
3. Nutrition improvement of food production in scientific manner to all levels of food processing for people.
4. Selection of effective level of society maintaining the food security; such as School level nutrition and nutrition awareness.
5. Making food security and nutrition information system available for required parties through trustworthy and acceptable government bodies.
6. Creating an early warning and emergency management for informing the relevant parties timely to participate actively to solve the problems of food security.
7. Generating and developing an institutional and legal framework, and proper financing systems to solve the problems of food security truly.
8. Strategic approaches for policy implementation, monitoring and evaluation with the policy intervention and updating.

Time to time improvements in the same examination and evaluation system for the food security is needed through continuous basis. It should be functioning independently from all other external factors such as political decision making or superseding by another policies of a territory against food policy and it should be internationally accepted best practices in long run.

The Global Environmental Change and Food System (GECAFS) framework is defined with all four dimensions associated with food security; 1. Food availability, 2. Food access, 3. Food utilization and 4. Food stability. (Ingram & Brklacich, 2002) (Ericksen, Ingram, & Liverman, 2009) There were many scholarly frames works in different studies and some dimensions of the food security are included in those frame works with respect to their study area. The UNICEF is suggesting the Project Cycle Management in Food and Nutrition Programs to improve the efficiency effectiveness. It is very useful to follow to get benefits of the government's programs. (Gross, Schoeneberger, Pfeifer, & Preuss, 2000) Poverty mapping is more commonly used for identifying the geographical settings through various factor analyze purposes because food poverty is also one important factor of poverty modelling. Also, World Bank has had an objective for sustainable development strategy. Some studies were based on Household Income and Expenditure Survey (HIES) data for different countries. The way it follows the World Bank's approach to poverty analysis, which has been used to help users formulate poverty reduction strategies:

1. Measuring Poverty.
2. Analyzing Poverty.
3. Mapping Poverty.

4. Other related information.

Accordingly, poverty variables can be identified in geographical changes and it uses for policy interventions and suggestions and then targeting indices, relevant other matters into food security sector to reach the development achievements through most suitable, efficient and effective manner with sustainable development.

Reference

1. Deitchler, M., Ballard, T., Swindale, A., & Coates, J. (2011). *Introducing a Simple Measure of Household hunger for Cross-Cultural Use.* USA: USAID.
2. Deitchler, M., Ballard, T., Swindale, A., & Coates, J. (2011). *Introducing a Simple Measure of Household hunger for Cross-Cultural Use.* USA: USAID.
3. Ericksen, P. J., Ingram, J. S., & Liverman, D. M. (2009). Food Security and Global Environmental Change: emerging challenges. *Environmental Science & Policy, 12*(4), 373-377.
4. FAO. (1996, June). Food Security. *Policy Brief*, 1-4.
5. Gross, R., Schoeneberger, H., Pfeifer, H., & Preuss, H.-J. A. (2000, April 01). The Four Dimensions of Food and Nutrition: Definitions and Concepts. Rome: FAO.
6. Group, T. A. (2015). *Insights from the Global Food Security Index for Long-Term Planning.* The Economist Intelligence Unit Limited.
7. Ifeoma, J. I., & Agwu, E. A. (2014, March 30). Assessment of Food Security Situation among Farming Households in Rural Areas of Kano State, Nigeria. *Journal of Central European Agriculture, 15*(1), pp. 94-107. doi:10.5513/JCEA01/15.1.1418
8. Ingram, J., & Brklacich, M. (2002). *Global Environmental Change and Food System - GECAFS: ANew Interdisciplinary Research Project.* Ottawa: The GECAFS Project. Retrieved March 01, 2016
9. MRI. (2012). *Assessment of Nutritional status and Associated Factors.* Colombo: MRI, Ministry of Health.
10. Mukherjee, A., Hoare, D., & Hoare, J. (2002, September 2-4). Selectio of Research Methodology for PhD. Researchers Working with an Organization. *18th Annual ARCOM Conference, 2*, 667-76.
11. UNO. (2016). *Sustainable Development Goals.* UNO.
12. Vhurumuku, E. (2014, February 25). Food Security Indicators. Niorbi.

Way ahead of Malaysia's automotive industry: an assessment on national car

Sayeeda Kamaruddin, Masitah Kamaludin
Department of Statistics Malaysia

Abstract

This paper discusses the overview of automotive industry in Malaysia, in particular the sales of motor vehicles. The government became directly involved in the automotive industry in 1983 through the establishment of national car company Proton, followed by Perodua in 1993. As such, the government offered initiatives to encourage the local assembly of vehicles and manufacturing of automobile components. The National Automotive Policy (NAP) was first introduced in 2006 to facilitate the integration of the local automotive industry to regional and global levels and further reviewed in 2009 to focus on enhancing the capabilities of the domestic automotive industry and create a more conducive environment for investments. The local automotive industry has shown a healthy growth, with the capability to fully design, engineer and manufacture cars from the ground up and has also elevated itself to be among the few full-fledged automotive manufacturers in the ASEAN region and dominated Malaysian automotive market with market share of 90 per cent in 1999. However, the market shares of national automobiles for later years have declined obviously whereby for the first time in 2014, non-national car brands took 53 per cent of total volume for the Malaysian automotive market. Therefore, a comparative study is undertaken to assess the development of both national and non-national cars in Malaysia.

Keywords

Automotive; National Automotive Policy

1. Introduction

1.1 Overview of Automotive Industry in Malaysia

Malaysia's automobile industry dates back to the pre-independence era when Ford Malaya was founded in Singapore in 1926 to become the first automobile assembly plant in Southeast Asia as a regional distributor of Ford products. After independence, the automotive industry in Malaysia was established in 1967 to stimulate national industrialisation with the setup of Volvo Car's assembly plant in Shah Alam, Selangor. However, in the early 1980s, due to economic slowdown, the Government realised the need to embark on high-value economic activities that would put Malaysians on the move and spur the industrialization of a country. The Fourth Malaysia Plan, 1981-1985 had emphasized that heavy industries would create new engines of growth and provides strong forward and backward linkages for the development of industries.

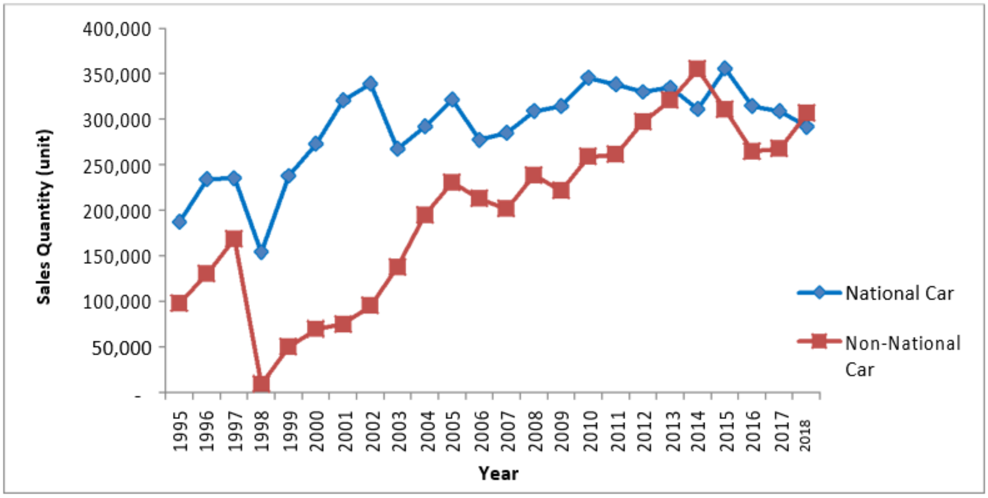
The government became directly involved in the automotive industry in 1983 from a mere motor car assembler into a car manufacturer by establishing a national automotive company Perusahaan Otomobil Nasional Sdn. Bhd (Proton). The first Proton cars were launched in 1985, equipped with the government's protective measures and subsidies in various ways. Subsequently, the national automotive programme also established a small car manufacture, Perodua in 1993, a heavy vehicle company (Malaysian Bus and Truck, MTB) in 1994, a motorcycle manufacturer (Modenas) in 1995 and a light vehicle commercial manufacturer (INOKOM) in 1997.

In order to facilitate the necessary transformation and integration of the local automotive industry into the increasingly competitive regional and global networks, the National Automotive Policy (NAP) was first introduced in 2006 under the Third Industrial Masterplan (IMP3) 2006-2020. In 2009, the policy was reviewed to focus on enhancing the domestic automotive industry's capabilities and creating a more conducive investment environment. Eventually, NAP 2014, a second review of NAP was launched in 2014, focusing on green market expansion initiatives through improvement of the entire automotive ecosystem and technology, human capital and supply chain development to establish Malaysia as a regional Energy-Efficient Vehicle (EEV) hub by 2020.

1.2 Performance of Local Automotive Industry in Malaysia

Although the national automobiles have been sheltered from foreign competition through tariff protection, trade barriers, tax exemptions, rebates, subsides and other government incentives, the market shares of national automobiles have declined slightly in later years. For example, the combined market share of Proton and Perodua declined from 82.6% in 1999 to 58.2% in 2005 (a reduction of 24.3%) and to 46.7% in 2014 (a drop of 11.5%). In contrast, the market shares of foreign automobiles have increased significantly. Therefore, a comparative study is undertaken to assess the development of both national and non-national cars in Malaysia.

Table 1: Sales of National and Non-National Car, Malaysia, 1995 – 2018



Source: Malaysian Automotive Association

2. Methodology

This study investigated the performance of national and non-national car in Malaysia, particularly in terms of the sales quantity. Sample used for this study is the sales quantity of national and non-national car in Malaysia from 1995 to 2018, which gathered from Malaysian Automotive Association (MAA). First, independent sample t-test was used to determine whether there is a statistically significant difference between the means of sales value between the two groups. Subsequently for national car, further independent sample t-test was carried out for their market share to the total to assess whether there exists a statistically significant difference between the means of market share before 2005 (1995 to 2004) and after 2005 (2005-2018). This reference point is used as 2005 was the turning point whereby growth of market shares for national car declined, in contrast with non-national car.

3. Result

3.1 Independent -Samples T-test of National and Non-national Car

Hypothesis for the test are shown as below:

H0: u1 = u2; HA: u1 ≠ u2

Where by:

u_1: Means of sales value for national car; $u_{2:}$ Means of sales value for non-national car

Table 1: Independent Samples Test Analysis of National and Non-national Car

Group	Levene's Test for Equality of Variance		T-test for Equality of Means				
	F	Sig.	t	DF	Sig. (2-tailed)	Mean Diff.	SE Diff.
Equal variances assumed	11.52	0.00142***	4.14	46.0	0.000149***	91799.17	22197.41
Equal variances not assumed			4.14	34.8	0.000212***	91799.17	22197.41

Note. ***, ** and * denote that the statistical test value is significant at the 1%, 5% and 10% level respectively.

The results revealed that there was a significant difference in the mean of sales between national and non-national cars (t46.0 = 4.14, p < .01). The average sales for national car was 91,799 units higher than the average sales for non-national car.

3.2 Independent -Samples T-test of Market Share for National Car Before and After 2005

Hypothesis for the test are shown as below:
H0: u1 = u2; HA: u1 ≠ u2

Where by:

u_1: Market share of national car before 2005 ; $u_{2:}$ Market share of national car after 2005

Table 2: Independent Samples Test Analysis of Market Share for National Car Before and After 2005

Group	Levene's Test for Equality of Variances		T-test for Equality of Means				
	F	Sig.	t	DF	Sig. (2-tailed)	Mean Diff.	SE Diff.
Equal variances assumed	25.60	4.55747E-05***	5.56	22.0	1.4E-05***	0.185	0.033328
Equal variances not assumed			4.80	10.2	0.000678***	0.185	0.038554

Note. ***, ** and * denote that the statistical test value is significant at the 1%, 5% and 10% level respectively.

There was a significant difference in the market shares of national cars (t5.56= 0.000014, p <.01) before and after 2005 whereby market shares before 2005 was 0.185 higher than after 2005.

4. Discussion and Conclusion

The analysis showed that the market shares of national cars before 2005 was significantly difference with that of after 2005, whereby market share before 2005 was 0.185 higher than after 2005.The results consistent with few researchers that associate the downturn of local Malaysian automobile industry with the introduction of ASEAN Free Trade Agreement (AFTA) in 2005. In compliance with the AFTA, the government committed to reduce the import tariff whereby before 2005, all imported vehicles were subject to an import tariff of 60% to 300% of the vehicle's value, varying by engine size. As a result, national cars had greater competition from foreign brands that enjoy lower tariffs.

According to Ariffin and Sahid, (2007), sales of Proton were once dominated the local market in mid 90s and had been really affected with the flourished of foreign cars following the implementation of AFTA in 2005. Drebee at.al (2014) also pointed that the decrease in the market share for Proton and Perodua was attributed to consumers withholding their purchases due to anticipation of lower prices of imported cars with the coming of the AFTA.

Furthermore, from June 2011, imports of used automotive parts and components would no longer be allowed. The policy also called for the provision of incentives for the local assembly and manufacture of hybrid and electric vehicles and their components, with a temporary exemption of excise tax on such vehicles of 2,000 c.c. or below, in the first half of 2011. Toyota and Honda seized this incentive and their hybrid vehicles sold well.

On top of that, there are number of factors which relate the decreasing share of national cars at the later years. Lim, Ngat Chin & Tong, Jane Terpstra, (2013) highlighted that the absence of stiff competition had encouraged Proton to keep on producing outdated designs that neglected basic safety features such as airbags and anti-lock braking systems for most of its domestic models. Wad, P. and Govindaraju, V.G.R.C. (2011) associates the uncompetitive of national car with high dependence on domestic market and technology agreements which has limited their performance at a regional and global scale.

For the local automotive industry to remain significant in the Malaysia automotive industry, the national carmakers should stop being complacent and move forward without the support of the government (Mahidin, Mohd Uzir & R.Kanageswary, 2014). They cannot depend on local market given the Malaysian automotive market is relatively small, with 32 million populations, which is not enough for an automotive company to be overly dependent on domestic sales alone. To expand its sales internationally and domestically, they need to improve its technology and product upgrading especially among parts and component suppliers.

References

1. Abdul Rashid, Mohd Afzanizam. (2018), “Automotive Industries – Selected Markets”, From the Desk of Chief Economist, Bank Islam
2. Ariffin, Aini Suzana & Iskandar Sahid, Mohd Lutfi & Mavak, Mathew (2016), “Factors Potentially Enhancing National Automotive Policy Goals and Industry Innovation”, Journal of Science, Technology and Innovation Policy, Vol. 2 No. 1 (June 2016)
3. Ariffin, Aini Suzana & Iskandar Sahid, Mohd Lutfi (2017), “Competitiveness Analysis of ASEAN Automotive Industry: A Comparison between Malaysia and Thailand”, Journal of Science, Technology and Innovation Policy, Vol. 3 No. 2 (Dec. 2017)
4. APEC (2012), “Overview of Malaysian Automotive Industry”, 17th Automotive Dialogue, St. Petersburg, Russia, 30-31 October 2012
5. Drebee, Hyder Abbas., Abdul Razak, Nor Azam. & Abd Karim, Mohd Zaini. (2014), “Is There an Overlapping Market Between National Car Producers an alaysia?”, Jurnal Ekonomi Malaysia 48(1) 2014 75 – 85
6. Henriksson, J. (2012), “The Malaysia Automotive Sector”, National Institute for Foreign Trade, Italy
7. Lim, Ngat Chin & Tong, Jane Terpstra, (2013), “Proton: Its Rise, Fall, and Future Prospect”, Asian Case Research Journal.
8. Mahidin, Mohd Uzir & R.Kanageswary (2004), “The Development of the Automobile Industry and the Road Ahead”, Journal of the Deaprtment of Statistics, Malaysia
9. Rosli, M. (2006), “The Automobile Industry and Performance of Malaysian Auto Production”, Journal of Economic Cooperation 27,1 (2006) 89-114
10. Sultana, Muneer & Ibrahim, Khairul Amilin (2014), “Challenges and Opportunities for Malaysian Automotive Industry”, American International Journal of Contemporary Research, Vol. 4, No. 9; September 2014
11. Wad, P. and Govindaraju, V.G.R.C. (2011), “Automotive Industry in Malaysia: An Assessment of its Development”, International Journal of Automotive Technology and Management, April 2011

Data preparation: DOSM statistics data warehouse practice

Ahmad Sauqi Haris, Razaman Ridzuan, Zulaikha Kamaruddin,
Asri Shajarah Hassan, Siti Haslinda Mohd Din
Department of Statistics Malaysia

Abstract

Data preparation is a fundamental stage of data analysis. This paper describes the data preparation process as practised by the Department of Statistics Malaysia (DOSM) in archiving data into the data warehouse. Previously, DOSM's data set were stored separately in silo mode by subject at every division, where more than 5,000 micro and aggregate data set were scattered. Starting 2014, DOSM began its data warehouse project, Statistics Data Warehouse (StatsDW) to consolidate micro and aggregated data in the Enterprise Data Warehouse. The most important process in StatsDW is Data Preparation. To prepare the data set, it involves three main division in DOSM which are Subject Matter Division, Data Integration and Management Division and Information Management Division. In this paper, we focus on data preparation in DIMD which are Data Profiling, Data Quality and Extract Transfer Load. We will firstly show the importance of data preparation in data analysis, followed by step-by-step process and the challenges in implementing data preparation. Finally, we will suggest some future directions of data preparation in StatsDW, DOSM.

Keywords

Data Warehouse, Data Profiling, Data Quality, Extract Transform Load, Department of Statistics Malaysia

1. Introduction

Department of Statistics Malaysia (DOSM) was established in 1949 under the Statistics Ordinance 1949 and was then known as Bureau of Statistics. In 1965, the name of Bureau of Statistics was changed to the Department of Statistics, Malaysia and was operating under the provisions of Statistics Act 1965. The Department's responsibility to collect, interpret and disseminate latest and real time statistics in the monitoring of national economic performance and social development.

In accordance with the functions stipulated, the Department collects economic and social data through censuses and surveys conducted regularly. Previously the historical data collected are stored in databases kept separately by the respective Subject Matter Division (SMD). Besides micro data, there are also aggregated data in the publication disseminated and are available both in softcopies and hardcopies. Meanwhile, some other aggregated data are disseminated or obtained in hardcopies.

The Department's ICT Strategic Plan (2005 – 2009) has discovered that the data management system in the department are structured in accordance to the applications and subjects developed in-silo (Department of Statistics Malaysia, 2004). Databases developed in-silos are not uniformed and non-centralised resulting in the databases not able to be accessed and shared electronically with other interested parties. There are also other weaknesses when databases are developed in-silos such as limited harvesting of data, inefficient data management and unorganised data archiving. To overcome this situation, DOSM has started a data warehouse project, the Statistics Data Warehouse (StatsDW) to store the data in one centralized storage in November 2014.

StatsDW is a database used for reporting, conducting data analysis and performing data analytics by integrating data from one or more disparate sources. The main objective of StatsDW is to consolidate all micro data and aggregated data in the Enterprise Data Warehouse (EDW). EDW is a consolidation of various heterogeneous DOSM data sources. It's combines operational data store (ODS) and OLAP data (Star/Snowflake Schema). StatsDW also aims to enhance the data quality and consistency.

2. Importance of Data Preparation

Data preparation is a fundamental aspect of the modelling process and the most important part of the process since it occupies up to 80% of the total time of the project (Refaat, 2018). In StatsDW, Data Preparation was one of the important, time consuming and crucial process. It involves data collecting, cleaning, processing and consolidating the data for use in analysis. Data Preparation is a sub-domain of data integration that can be executed with dedicated tools or traditional tools for data integration like ETL tools, data virtualization or data warehouse automation (Tischler & Grosser, 2017).

Many of the problems which business users and analysts in organizations confront when working with data crop up during the data preparation processes (Stodder, 2016). However, with improved practices and technologies for data preparation, organizations can better deal with current data troubles and prepare for future challenges arising from new data and user requirements. Thus, it is important to ensure that data preparation is being done properly.

3. StatsDW Architecture

Figure 1 shows the architecture for StatsDW. Data preparation started from SMD which are responsible for census and survey in DOSM. The process that involve at SMD are collecting, cleaning and analysing the raw data into data set. Average time for the process is six to twelve months depends on census or survey. The data set that being processes then will be sent to DIMD in various formats such as MS Excel, CSV, MS Access, My SQL, MS SQL Server. These formats are various in nature as there was no standardized format that being used in early stage of data preparation process.

Data Preparation at DIMD started right after receiving the data set from SMD. The detail process involve will be explain in section 4. StatsDW are using two type of server which are Netezza and DB2. Lastly, the processed data set will be disseminated to the user using 7 different platforms namely Mobile Apps, Visualisation, Time Series, eDataBank, MyLab, Location Intelligence, and Analytics.

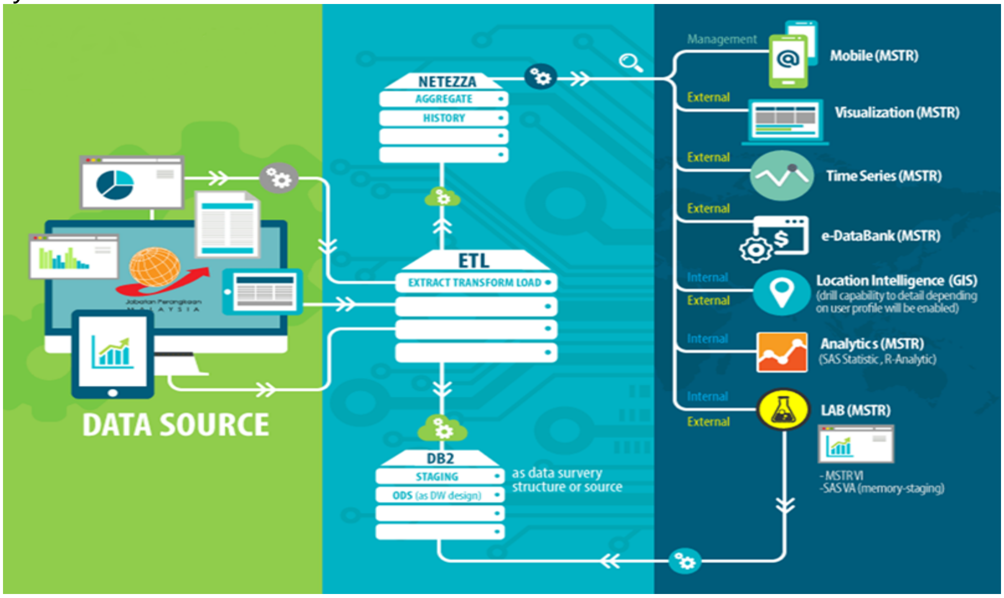


Figure 1: StatsDW Architecture

4. Data Preparation Process

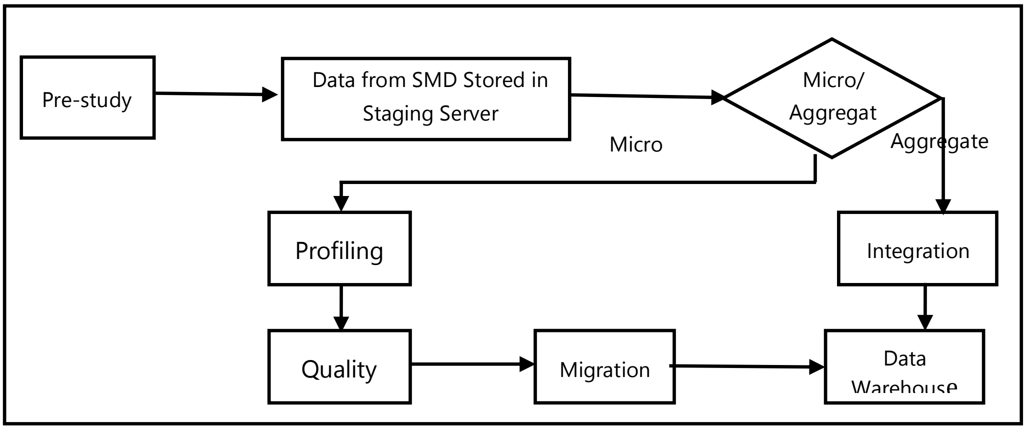


Figure 2: Data Preparation Process

The first step in Data Preparation Process is Pre-Study. The purpose of this process is to understand the census/survey information (questionnaire, coverage, code and classification). This process is important in order to identify the changes that happened between census/survey. Normally the differences occur due to the questionnaire generation as the questions will be improvised between census/survey due to stake holder request, data request trend, changes in industry classification and etc.

The next step is receiving the final and clean micro and aggregate data set from SMD through the Staging Database (server). Data that stored by SMD will not be modified and it remain the origin format. The aggregated data are being check with the publication to ensure that the data set are consistent. Then it will be integrated into the Data Warehouse using standardise format.

Meanwhile for Micro data set, it needs to go through data profiling and data quality process using IBM InfoSphere DataStage & Quality Stage and IBM Data Studio. This process will change the data format to the database file format. After that, it will be migrated into the Data Warehouse through the Extract Transfer Load (ETL) process.

Data Profiling is the process of examining the data available in an existing data source and collecting statistics and information about that data (Abadi, 2007). Profiling data is an important and frequent activity of any IT professional and researcher (Naumann, 2013).

Data Profiling involves the examination of the data set process. The main purpose of data profiling is to improve the ability to search the data by tagging the data with keywords, descriptions, or assigning it to a category. Besides that, data profiling gives metrics on data quality including whether the data conforms to standards or patterns. In this process, it also assesses whether the metadata accurately describes the actual values in the source database. Furthermore, it has an enterprise view of all data, for uses such as master data management where key data is needed, or data governance for improving data quality.

Data Quality is an essential characteristic that determines the reliability of data for further data analysis and data analytics. The main purpose of data quality is to ensure that there are no misspellings, typing error, and random abbreviations in the data set. The data set are being check with publication statistics.

ETL is a migration process from Operational Database and other Historical Data. During the Data Migration activity, data will be extracted from multiple sources and format. Transform activities involve data mapping, verification of process, code generation and data conversion. The final data set will be loaded into the end target in the Enterprise Data Warehouse Environment which are

IBM DB2 for micro data and IBM Netezza for aggregate data. This process is conducted by IMD.

5. Result

From the Data Preparation, the data set can be store in data warehouse with reliable and confident. This is because the data that have being through the data preparation process were being check thoroughly. Besides that, this process will make dissemination a lot easier and faster. As we know, the data quality that being stored in StatsDW is guaranteed clean, thus there is no problem to disseminate any type of allowed data to the user.

Lastly, the main storage that stored plenty of important data are being centralized in one data warehouse. This is the most important part in the Data Warehouse. It is because the problem that we are facing which are data being stored in-silo is now being centralized. We do not have to fear if someone loss the data in their own storage or anything might happen to individual computer anymore.

6. Discussion and Conclusion

There is two type of challenges that we are facing in this Data Preparation process that are internal and external. The internal challenges are lack of competence staff and IT literate to maintain the operation. Besides that, the data set format that being use also are different between each subject. These internal challenges can be resolve by staff training, hiring outsource consultant and internal discussion.

Meanwhile for external challenges, the server that being used are unfriendly because it cannot read any type of database. In the future, it is important to ensure that server that being selected are readable to various type of database. Furthermore, it is also important to choose software that are less complicated to use. As an example, the software should not need to use command to operate it.

References

1. Abadi, D. J. (2007). Column-Stores For Wide and Sparse Data. *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*, (pp. 292–297). California.
2. Department of Statistics Malaysia. (2004). *DOSM Strategic Plan 2005 2009.* Malaysia: Department of Statistics Malaysia.
3. Naumann, F. (2013, December). Data Profiling Revisited. *SIGMOD Record*, pp. 40-49.
4. Refaat, M. (2018, May 2). *https://www.datawatch.com/.* Retrieved from https://www.datawatch.com/2018/11/09/data-preparation-101-the-objective-of-data-preparation/
5. Stodder, D. (2016). *Improving Data Preparation for Business Analytics.*
6. Tischler, R., & Grosser, T. (2017). Data Preparation - Refining Raw Data into Value. 5.

Quarterly Employment Survey & Establishment Approach in Development of Labour Demand Statistics

Mohammad Faris Roslan, Aiman Zakwan Zahari
Department of Statistics Malaysia (DOSM)

Abstract

This paper explores the establishment approach in development of labour demand statistics in Malaysia. It is based on Quarterly Employment Survey (QES). The Quarterly Employment Survey is conducted to collect information on labour demand for all economic sectors. Statistics obtained through the survey are the number of positions, filled positions, vacancies and jobs created by occupation and industry. The statistics is one of the inputs in the planning and formulation of the country's human capital development policy. This paper will highlight the method used in determining the sample size, and the issues and challenges faced in collecting and processing all the data and information which considering and determined by the sample size.

Keywords

Jobs created; Positions; Vacancies; Issues; Challenges

1. Introduction

Labour demand indicates the total labour that the economy is willing to employ at any given point of time. At the microeconomic level, labour demand by individual firm is the positions in the company; and with the information of vacancies, hires and separations, the statistics of jobs created is obtained.

The Quarterly Employment Survey (QES) conducted to collect information on labour demand in formal private sector in Malaysia. The principle statistics obtained through the survey are the number of jobs (positions), filled positions, vacancies and job created by economic activity and category of skill. At the microeconomic level, labour demand by firm refers to total positions in the company and with the information of vacancies, hires and separations; the statistics on jobs created can be estimated.

This paper will highlight the method used in determining the sample size, the response mode and results, and the issues and challenges faced in collecting and processing all the data and information which considering and determined by the sample size.

2. Methodology

The QES is implemented using establishment approach, covering all economic activities which is classified according to Malaysia Standard

Industrial Classification (MSIC) 2008 version 1.0. The reporting unit for this survey is the establishment. Occupation coverage is based on nine categories according to Malaysia Standard Classification of Occupation (MASCO) 2013.

The sampling frame is from the identified population. The sampling design of the survey is a one-stage stratified random sampling. Industries at national level have been classified as stratum and the establishment as the sampling unit. Each stratum has been set up into four sub-stratums to ensure the sample is distributed taking into account the economy characteristics of the industry. The main sub stratum comprise is heterogeneous, was fully covered while other sub stratum that is homogeneous were sampled which is based on Small and Medium Enterprise (SME) category.

The number of employees used to estimate the sample size. The formula used in the estimation of the sample size for a stratum is as follows:

$$n = \frac{\left(\sum N_i S_i\right)^2}{V + \sum N_i S_i^2}$$

where;

n = Sample size

N_i = Population size for stratum i

S_i^2 = Variance for stratum i

V = Desired variance

$$V = RSE^2 \cdot \left(\frac{\hat{Y}_i}{Z}\right)^2$$

where;

$\hat{Y}_i$ = Estimated no.of employee for stratum i

RSE = Relative standard error

Z = Cofindence level

Sample is distributed to sub stratum of the industry using Neyman Allocation Method as follows:

$$n_{hi} = \left(\frac{N_h S_h}{\sum N_h S_h} \right) n_i^{'}$$

$h = 2,3$ and 4
$i = 1,2,\ldots$k

where;

n_{hi} = Sample size for sub stratum h of stratum i
N_h = Population size for sub stratum h
S_h = Standard deviation for sub stratum h
n_j = Sample size for stratum i
h = Sub-stratum
j = Stratum

The sample sizes for this survey in 2018 are 8,722 establishments. Establishments of the large categories were fully covered while establishments of the second to fourth sub stratum were randomly selected using systematic random sampling.

Weighted analysis is done using sampling weight to ensure that the selected sample can reflect the population survey. The weights required are the sampling design weight and non-response weight.

The sampling design weight for the establishment at stratum h is as follows:

$$W_h = \frac{N_h}{n_h} \ , \quad \text{h} = 1,\ldots,4$$

where,

N_h = Total population of sub stratum h; and
n_h = Total sample of sub stratum h

Non-response weight at sub stratum h as below:

$$NRW_h = \frac{1}{n_h^{'} / n_h} \ , \quad \text{h} = 1,\ldots,4$$

where,

$n_h^{'}$ = Numbers of respond sample size for sub stratum h
n_h = Numbers of sample size for sub stratum h

The method of calculating the sampling design weight after the survey (adjusted weight) on sub stratum h as below:

$$W_h^{'} = W_h \times NRW_h \quad , \quad h = 1,\ldots,4$$

where,

W_h = Sampling design weight at sub stratum h

NRW_h = Non response weight at sub stratum h

3. Result

As we go deeper into Malaysia's approach in compiling the labour demand statistics, we make several countries comparison for benchmarking purposes. In United States of America (USA), Bureau of Labour Statistics (BLS) uses the Job Openings and Labour Turnover Survey (JOLTS) program to produce monthly data on job openings, hires, and separations. The JOLTS target sample size is approximately 16,400 establishments and covers all nonfarm establishments in the private sector as well as federal, state, and local governments in the 50 states and the District of Columbia. JOLTS data series are published on a monthly basis and seasonally adjusted data are also published for most JOLTS series.

The Job Vacancy Statistics (JVS) in Canada measures unmet labour demand. It provides a monthly portrait of the level of unoccupied positions, job vacancy rates and unemployment-to-job vacancy ratios. All estimates are produced at various levels of cross-classification of geography province and territories and industry two-digit North American Industry Classification System (NAICS) based on three-month moving averages. These data are useful in assessing the presence and degree of labour shortage and labour market. The JVS contributes to the understanding of trends in filled and unfilled job demand in the labour market and helps identify areas at risk of human resource shortages. Federal departments such as Employment and Social Development Canada as well as provincial and territorial agencies, educational organizations and the private sector are interested in this kind of information.

In Australia, Australia Bureau of Statistics (ABS) through the Job Vacancy Survey is a quarterly survey which collects the number of job vacancies from sample businesses taken from the ABS Register of Business. Job Vacancy Survey uses a sample survey methodology and collects information via online forms and/or telephone interviews. Approximately, 5,000 employers are selected from the ABS Business Register. The Survey of Job Vacancies is used to estimate the number of job vacancies in Australia. Estimates produced from this survey are a main economic indicator of employment growth and are used for monitoring the Australian economy and formulating economic policy. All job vacancies for wage and salary earners are represented in the Job Vacancies Survey, except those in the Australian permanent defence forces, in businesses primarily engaged in agriculture, forestry and fishing, in private households employing staff, in overseas embassies, consulates, etc. located outside Australia.

In New Zealand, the Quarterly Employment Survey estimates the demand for labour by New Zealand businesses. From the survey responses, we estimate the levels and changes in employment, total weekly gross earnings, total weekly paid hours, average hourly and average weekly earnings, and average weekly paid hours in the industries surveyed. Quarterly Employment Survey estimates the number of jobs filled, not the number of people employed. This means a person with multiple jobs during the reference week could be counted multiple times. Data from Quarterly Employment Survey about the total paid hours is used in compiling gross domestic product-economic activity for selected industries. Quarterly Employment Survey average earnings statistics are used in calculating superannuation and paid parental leave.

Table 1: Comparison of employment surveys conducted among selected countries

Country	Collection Frequency	Coverage	Data element	Note
United States of America	Monthly	16,000 establishments (nonfarm, federal, state, local government in 50 states and District of Columbia)	Total employment, job openings, hires, quits, layoff & discharge, other separation and total separation	Estimates by industry, region and establishment date
Canada	Monthly	15,000 establishments (exclude agriculture, Fishing & trapping, private household services, religious organizations, military personnel of defence services and federal, provincial, and territorial public administration	Job vacancy, number occupied position) and unemployed person	
New Zealand	Quarterly	3,500 establishments	Number of employees and gross salaries paid; bonuses paid; overtime payment; and severance, terminate and redundancy payments paid to employees for each month of the reference	
Malaysia	Quarterly	8,722 establishment	the number of positions, filled positions, vacancies and jobs created by occupation and industry	

In 2018, as mentioned before, the sample size covers 8,722 establishments. Figure 1 show the response results received for first and second quarter of 2018 based on different mode of data collections used.

Figure 1: Percentage of QES, Q2 2018.

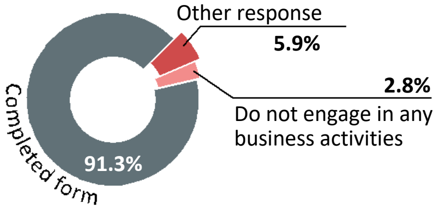
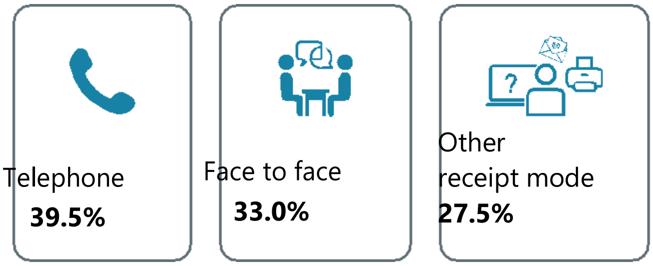


Figure 2: Mode of conducting the surveys in Q2 2018.



Based on result in quarter 2 of 2018, 91.3% respondent had completed the survey achieved the key performance indicator which is 85.0%. The highest receipt mode is by telephone which recorded 39.5% followed by face to face (33.0%).

4. Discussion and Conclusion

Despite a very high number of completed surveys, the respondents raised several issues. Among the issues are respondents find difficulties in understanding the concept and definition of difficulties to categorize the employee based on MASCO. This led too many missing value and at time contribute to non-response. Other than that is additional time needed in data validation process. Overlapping this survey to other establishment survey also partial of non-response.

In terms of definitions, currently there is no specific benchmark that can be used as our reference on the jobs created statistics. The job creation statistics is derived by using variable vacancies, hires and separations. The statistics is only reliable at industry and national level.

To enhance this survey, further research needs to be conducted to identify the right and accurate indicators in compiling job creation data and statistics. It is important as it will lead us for a more comprehensive and better understanding on the labour markets and will be the input to formulate future economic policy. At the end of the day, we would like to drive enhance the labour market and improve Malaysia's competitiveness.

References

1. Australia Bureau of Statistics (2014), Job Vacancy Survey. Canberra, Australia.
2. Bureau of Labour Statistics (2014), Job Openings and Labour Turnover Survey (JOLTS). NE Washington, United States of America.
3. Department of Statistics Malaysia (2018), Employment Statistics. Putrajaya, Malaysia.
4. Department of Statistics Malaysia. (2016). Employment and Salaries & Wages Statistics Report 2015.
   Putrajaya, Malaysia.
5. Department of Statistics Malaysia. (2016). Labour Force Survey Report 2015. Putrajaya, Malaysia.
6. Statistics Canada (2018), Job Vacancy and Wage Survey (JVWS) Statistics. Ottawa, Ontario, Canada.
7. StatsNZ Tatauranga Aotearoa (2018), Quarterly Employment Survey (QES). Wellington, New Zealand.

Bayesian Network approach to the causal influence of socioeconomic status on Infectivity of dengue

Lamidi-Sarumoh Alaba Ajibola1[1,2], Shamarina Shohaimi[1], Mohd Bakri Adam[1], Mohd Noor Hisham Mohd Nadzir[1], Oguntade Emmanuel Segun[1], Nurul Akmar Ghani[1]

[1]Universiti Putra Malaysia, Selangor, Malaysia

[2]Gombe State University, Tudun Wada, Nigeria

Abstract

Low socioeconomic status (SES) is one of the major factors influencing diseases incidence and prevalence all over the world. Dengue fever as one of the neglected tropical diseases is also influenced by varying SES. This study was aimed at evaluating the influence of SES on the declarative incidence and prevalence of dengue fever. A cross-sectional study was conducted among the people living in the state of Selangor, Malaysia between May 2018 and October 2018 which involved 562 participants. Information on socioeconomic status and medical history of dengue fever was collected using a pre-test bilingual questionnaire (Malay and English). Bayesian network was used to classify the variables considered and estimate the approximate inference. The results showed that the family history of dengue fever was influenced by the level of education. The highest probability of respondents who have a family history of dengue was observed among respondents with primary school education followed by vocational training among others. Furthermore, the family history of dengue fever and type of residence influenced been infected with dengue fever. The highest probability of been infected with dengue fever with the family history of dengue was clustered in terrace buildings followed by apartment buildings, flats among others. This study has revealed the magnitude of the causal influence of SES on the infectivity of dengue fever in Selangor, Malaysia. Policies on mediation strategies to curtail the spread of dengue fever should focus more on individuals living in low-income buildings.

Keywords

Socioeconomic; Dengue fever; Bayesian Network; Selangor; Malaysia

1. Introduction

Urban areas are human settlement with social infrastructure and basic amenities. It is usually made up of individuals with middle to high socioeconomic status (SES) due to the cost of housing, cost of food and other essentials of life. Research in Malaysia had shown that dengue fever (DF) is a disease of urban areas (Mahyiddin et al., 2016) because the mosquitoes transmitting dengue virus has the higher chances of striving in an urban environment. Possible breeding sites are often formed in discarded man-made containers which are most common in an urban environment. Cans or discarded tins, old truck tires, plastic bottles, ant traps, gutters and non-degradable containers among others are most common in urban areas than rural areas (Mulligan et al., 2015). Urbanization has been confirmed as one of the social and environmental factors contributing to the number of dengue cases (Egger et al., 2008; Wilcox, Gubler, & Pizer, 2008). The incidence rate of dengue was significantly associated with low SES in a study of dengue transmission carried out in New Caledonia (Raphael et al., 2017). Individuals living in urban areas with low to moderate SES have a higher chance of contracting DF because of their choice of residence type, neighbourhood, and districts. SES can be measured and analyzed based on a combination of three factors; educational status, occupational status and total monthly household income (Winkleby, Jatulis, Frank, & Fortmann, 1992). Typically, SES can be categorized into three different categories namely; high, middle and low which depend on the standard of living of a particular country.

2. Methodology

Study site and design

Selangor is one of the most populous and developing states in Malaysia. The state usually accounts for the more than 50% morbidity rate of the overall dengue cases on yearly basis. In 2015, Malaysia has overall 120,836 reported cases of DF and Selangor accounted for 63,198 cases. In the subsequent year, there were 101,357 reported cases, Selangor had 51,652 cases and in the year 2017, there were 83,848 cases, the state of Selangor reported 44884 cases (Ministry of Health, 2018). The overall number of dengue cases in Malaysia continues to reduce but reported cases from Selangor continue to increase. An intercept cross-sectional study was conducted among the adults living in the state of Selangor, Malaysia between May 2018 and October 2018. Age from 18 years and above were considered as inclusion criteria. Information on SES, residence type and medical history of dengue fever were collected using a pre-test bilingual (Malay and English) structured questionnaire.

Sampling and sample size

In 2017, Selangor has a population of 6.39 million which makes it 20.49% of the Malaysian population, the prevalence of DF was 0.0070. The least percentage of expected sensitivity is 50% (Buderer, 1996), maximum critically acceptable width of 95% confidence interval was chosen. Thus, an approximate value of 550 individuals was the required sample size. A total of 562 respondents voluntarily participated in the study as anonymous.

Statistical analysis

In order to quantify the role of SES on the family history of DF and been infected with DF concurrently, a multinomial Bayesian Network (BN) was set up to propagate causal influence of SES on the aforementioned target variables. The combination of supervised and unsupervised learning was adopted to learn the structure of the network via bnlearn R package. Level of significance was considered at p-value of α <0.05.

Ethical clearance

The ethical protocols were duly approved by the Medical Research and Ethics committee of the Ministry of Health, Malaysia before data collection.

3. Result

Descriptive statistics of the study population

Table 1 presents the descriptive statistics of the respondents. The level of education of the respondents was as follow; 72.6% were university, 20.8% were secondary school, 5.7% were vocational training and lastly, 0.9% were primary school. Regarding occupational status, the majority were students (47.0%), followed by individuals working in a private sector (24.7%), the least of the occupational status were the retired individuals which were 2.0% of the respondents. The highest percentage of total monthly household income is RM1000 to RM3000 which were earned by 41.5% of the respondents and the lowest percentage of total monthly household income is RM9000 and above (4.3%).

The most common type of residence among the respondents are the terrace building (42.9%), followed by flats (15.1%), apartments (12.3%), semi-D (9.4%), bungalow (8.4%), others which was not part of the options given (6.9%), condominium (4.8%). The medical history of DF showed that 13.2% of respondents had been infected with DF and 34.5% had a family history of DF.

Table 1: Features of the respondents

Socioeconomic status	Frequency (%)	p-value
Level of Educational (EDU)		<0.05
Primary school	5(0.9)	
Secondary	117(20.8)	
University	408(72.6)	
Vocational training	32(5.7)	
Occupation status (OCC)		<0.05
Private sector	139(24.7)	
Government sector	65(11.6)	

Self-employed	40(7.1)	
Unemployed	24(4.3)	
Retired	11(2.0)	
Student	264(47.0)	
Housewife	19(3.4)	
Total monthly household income (INC)		<0.05
0.00	12 (2.1)	
RM1000-3000	233(41.5)	
RM3001-5000	163(29.0)	
RM5001-7000	102(18.1)	
RM7001-9000	28(5.0)	
RM9001and above	24(4.3)	
Type of residence (RES)		<0.05
0	1(0.2)	
Flat	85(15.1)	
Apartment	69(12.3)	
Condominium	27(4.8)	
Terrace	241(42.9)	
Semi-D	53(9.4)	
Bungalow	47(8.4)	
Other	39(6.9)	
Medical history of dengue		
Have you been infected with dengue fever before? (IDF)		<0.05
Yes	74(13.2)	
No	488(86.8)	
Do you have a family history of dengue? (FHD)		<0.05
Yes	194(34.5)	
No	368(65.5)	

All the p-values are based on Chi-square analysis to test difference among the groups across each level.
**1USD ≈ 4RM*

From Figure 1, it can be visualized from the BN that level of education is a parent node for occupational status, residence type and family history of DF. Occupational status is a parent node for total monthly household income and residence type. Income, on the other hand, has a child node of residence type which in turn has a child node of been infected with DF. The family history of DF also had been infected with DF as a child node. Notably, all the three factors of SES had an influence on the residence type.
The model of the BN is given by:

$$P[EDU,OCC, INC, RES, FHD, IDF] = P[EDU]\ P[OCC|EDU]\ P[FHD|EDU]\ P[INC|OCC]\ P[RES|EDU,OCC, INC]\ P[IDF|RES, FHD]$$

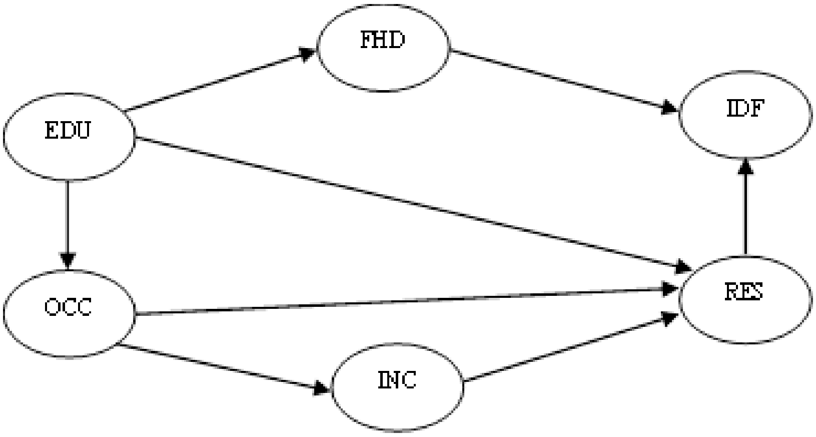


Figure 1: The structure of the influence of socioeconomic status on infectivity of dengue BN learned from data through conditional independent test. The arc represents the dependency between the nodes.

Considering the medical history of DF (IDF and FHD), the conditional probability of respondents who had the family history of dengue and primary school as their educational status was 0.60 which was the highest probability among those who had family history of DF, followed by vocational training (0.56), University (0.36) and lastly by secondary school (0.21) Figure 2.

Given that some respondents were formerly infected with DF and also have the family history of DF, the highest conditional probability was observed in terrace buildings, followed by apartments building, flats among others Figure 3. The strength of arc showed a very high probabilistic dependence exists between IDF and HDF (3.310732e-09).

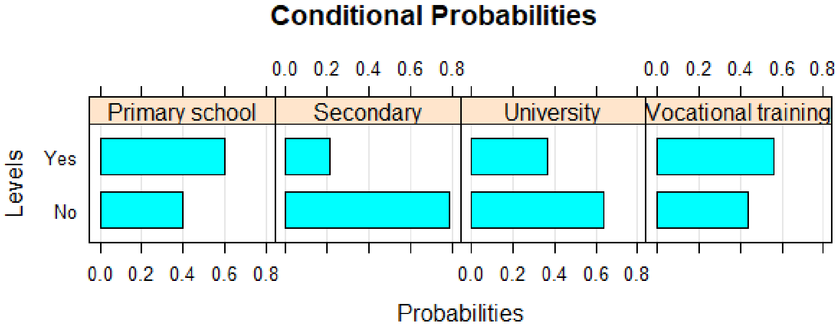


Figure 2: Bar plots of conditional probabilities of family history of DF given educational status

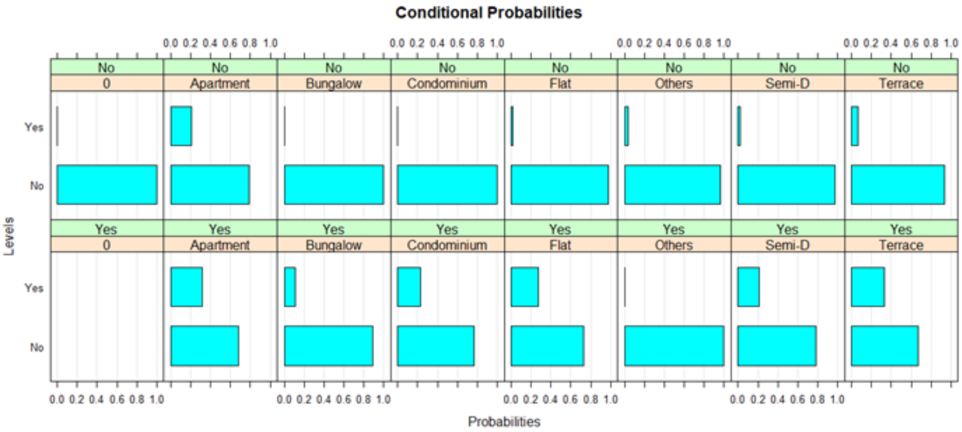


Figure 3: Bar plots of conditional probabilities of been infected with DF given a family history of DF and residence type

The approximate inference from the BN

Approximate inference in a BN uses Monte Carlo simulations, the implementation procedures are known as rejection sampling (Scutari, 2009). BN as an expert system can be queried based on variables of interest, thus, queries were set up to investigate respondents who were formerly infected with DF and family history of dengue fever based on the influence of SES.

Query 1: Given the level of education and total monthly household income and evidence set as formerly infected with DF, the highest probability was observed among the respondents with university education earning between RM1000-RM3000 as total monthly household income.

Query 2: Given the level of education and total monthly household income and evidence set as family history of DF, the highest probability was observed among the respondents with university education earning between RM1000-RM3000. Approximate estimate close to query 1.

Query 3: Given the level of education and occupational status and evidence set as formerly infected with DF, the highest probability was observed among students with university level of education.

Query 4: Given the educational status and occupational status and evidence set as family history of DF, the highest probability was observed among students with university level of education. Approximate estimate close to query 3.

Query 5: Given the occupational status and total monthly household income and evidence set as formerly infected with DF, the highest probability was observed among the respondents who are students earning between RM1000-RM3000.

Query 6: Given the occupational status and total monthly household income and evidence set as family history of DF, DF, the highest probability was observed among the respondents who are students and earning between RM1000-RM3000. Approximate estimate close to query 5.

Conclusion: Respondents who were students with university level of education earning the total monthly household income between RM1000-RM3000 have the family history of DF and were mostly infected with DF.

4. Discussion and Conclusion

This study provides the first description of declarative incidence of DF in Selangor, Malaysia, the influence of SES on DF and novel application of BN. There was no direct influence of SES on infectivity of DF but with the aid of BN, the approximate inference was estimated. This result is consistent with the study carried out in Ampang, Selangor (Mahyiddin et al., 2016) in terms of high knowledge about DF but low preventive practices. It is expected that respondents with university level of education should be less infected due high knowledge about DF compare to other level of education, but total monthly household income influence the choice of residence type which also plays an important role in infectivity of DF. With this discovery, poverty can be considered as a determinant of DF. The link between empirical evidence of DF to poverty was established in a systematic review (Mulligan et al., 2015). Policies on mediation strategies to curtail the further spread of DF should focus more on individuals living in low-income buildings.

In conclusion, infectivity of DF can be attributed to low income, occupational status and level of education which directly influence residence type. The novel application of BN gave improved results of visualization and consideration of variables simultaneously. These findings may contribute to the planning of mediation strategies to limit transmission of DF in Selangor, Malaysia. The expert system of BN on a larger scale can be used to predict the risk factors of mediation strategies for the purpose of maximizing the implementation of health policies.

Strength and limitation of the study

The strength of this study is the novel application of Bayesian Network to the causal influence of infectivity of dengue and the limitation lies in that fact that the research is based on a declarative statement from the respondents. Serological test to affirm seropositivity and seroprevalence of dengue virus among the respondents was not carried out.

Funding

This study was funded by a grant from the Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS) (Project No: 02-02-14-1561 FR).

Conflict of interest

None was declared

References

1. Buderer, N. M. F. (1996). Statistical Methodology: I. Incorporating the Prevalence of Disease into the Sample Size Calculation for Sensitivity and Specificity. *Acad. Emerg. Med.*, **3**(9), 895–900. https://doi.org/10.1111/j.1553-2712.1996.tb03538.x
2. Egger, J. R., Eng, E. O., Kelly, D. W., Woolhouse, M. E., Daviesa, C. R., & Colemana, P. G. (2008). Reconstructing historical changes in the force of infection of dengue fever in Singapore: Implications for surveillance and control. *Bull. World Health Organization*, **86**(3), 187–196. https://doi.org/10.2471/BLT.07.040170
3. Mahyiddin Nur Syakilah, Rosmawati Mohamed, Hamid Jan Jan Mohamed, N. R. (2016). High Knowledge on Dengue But Low Preventive Practices Among Residents in a Low Cost Flat in Ampang, Selangor. *Malaysian J. Nurs*, **8**(1), 39–48. Retrieved from http://www.mjn.com.my/articles_vwful.aspx?transid=29e26934-159d-4767-81a0-ea546736cd85#
4. Ministry of Health. (2018). Dengue cases down in 2017 compared to 2016. *The Star Online*, p. 10894. Retrieved from https://www.thestar.com.my/news/nation/2018/01/10/dengue-cases-down-in-2017-compared-to-2016/
5. Mulligan, K., Dixon, J., Joanna Sinn, C.-L., & Elliott, S. J. (2015). Is dengue a disease of poverty? A systematic review. *Pathogens and Global Health*, **109**(1), 10–18. https://doi.org/10.1179/2047773214Y.0000000168
6. Raphael, M. Z., Cano, J., Mangeas, M., Despinoy, M., Dupont-rouzeyrol, M., Nikolay, B., & Teurlai, M. (2017). Socioeconomic and environmental determinants of dengue transmission in an urban setting : An ecological study in Noume New Caledonia. *PLoS Negl. Trop. Dis.*, **11**(4), 1–18.
7. Scutari, M. (2009). Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Software*. https://doi.org/10.18637/jss.v035.i03
8. Wilcox, B. A., Gubler, D. J., & Pizer, H. F. (2008). Urbanization and the social ecology of emerging infectious diseases. *Social Ecol. Inf. Dis.*, 113–137. https://doi.org/10.1016/B978-012370466-5.50009-1
9. Winkleby, M. A., Jatulis, D. E., Frank, E., & Fortmann, S. P. (1992). Socioeconomic status and health: How education, income, and occupation contribute to risk factors for cardiovascular disease. *Am J. of Public Health*, **82**(6), 816–820. https://doi.org/10.2105/AJPH.82.6.816

Comparison of the Charts Based on Overlapping and Nonoverlapping Subgroup Observations

Jimoh Olawale Ajadi, Inez Maria Zwetsloot
City University of Hong Kong, Kowloon, Hong Kong, China

Abstract

A multivariate dispersion control chart monitors changes in the process variability of multiple correlated quality characteristics. In this article, we developed a multivariate chart based on overlapping subgroup observations and compare it with the most popular method for monitoring nonoverlapping grouped observations-generalized variance chart- proposed by Alt (1984). The effect of subgroup size is also studied. Steady-state average time to signal is used as performance measure. We show that monitoring methods based on overlapping subgroup observations are the quickest in detecting sustained shifts in the process variability. We use a simulation study to obtain our results and illustrated these with a case study.

Keywords

multivariate control chart; dispersion

1. Introduction

Multivariate variability charts monitor the process covariance matrix, Σ to detect changes quickly. Many different methods exist for monitoring multivariate dispersion. Methods are based on monitoring various characterization of the covariance matrix such as the determinant, trace, moving range, entropy or eigenvalues. For an instant, Alt (1984) proposed the generalized variance chart. The monitoring statistic for this chart is the determinant of the estimated covariance matrix. The challenge of using the determinant to monitor dispersion is that different covariance matrices can have the same determinant. Thus, Guerrero-Cusumano (1995) introduced a multivariate control chart based on conditional entropy. Conditional entropy is based on the diagonal elements of the covariance matrix. Yeh and Lin (2002) introduced a box-chart. This chart can detect changes in the process vector mean and covariance matrices simultaneously. The box-chart uses the probability integral transformation to change statistics into the same distribution. Levinson et al. (2002) introduced the G-statistic for monitoring the covariance matrix (named as G chart). The G-statistic tests equality between two covariance matrices. The authors recommended that the G chart should be used together with Hotelling's T^2 chart.

Group observations can either involve overlapping (moving window) or non-overlapping (fixed window) subgroups. For each period in an overlapping subgroup, the oldest observation is removed, and the newest is added to the group. However, the observation is grouped into consecutive non-overlapping subgroups in the fixed window subgroup observations. Table 1 differentiates between the grouped observations by using six bivariate observations of a patient systolic and diastolic blood pressure.

Table 1: A patient systolic and diastolic blood pressure bivariate data with a subgroup size of 2

Date	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Observation	[173, 86]	[176, 87]	[163, 84]	[169, 85]	[153, 82]	[152, 83]
Overlapping		[349, 173]	[339,171]	[332, 169]	[322, 167]	[305, 165]
Non-overlapping		[349, 173]		[332, 169]		[305, 165]

We give an overview of the models and notations that will be used throughout the article in Section 2. In Section 3, we briefly explain the selected methods used for our comparison. We give details about the simulation procedure in Section 4 and study the effect of subgroup size for the charts based on grouped observations in Section 5. We interpret the results of our simulation study in Section 6. In Section 7, a case study is employed to support our findings on the conclusion of the simulation. Conclusion and recommendation are discussed in Section 8.

2. Model and Assumptions

Throughout this paper, we are interested in monitoring the variability of a p-dimensional vector, X_t, representing p-characteristics which may be correlated. For this purpose, we assume that we can observe X_t at times $t =$ 1,2,3, ..., each at equidistance in time. We assume that $\boldsymbol{X_t} \sim \boldsymbol{N_p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a random observation drawn from the multivariate normal process with p the number of correlated quality characteristics. The process vector mean and covariance matrix of $\boldsymbol{X_t}$ are denoted by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. When the process is in-control, we denote $\boldsymbol{\mu} = \boldsymbol{\mu_0}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma_0}$**.** We monitor only changes in the process covariance matrix where the process mean is constant. In this study, we standardize $\boldsymbol{X_t}$ by transforming it to $\boldsymbol{Y_t}$ as in Eq. (1),

$$\boldsymbol{Y_t} = \boldsymbol{\Sigma_0^{-\frac{1}{2}}}(\boldsymbol{X_t} - \boldsymbol{\mu_0}). \tag{1}$$

Thus, $\boldsymbol{Y_t}$ follows a standardized multivariate normal distribution as $\boldsymbol{N}(\boldsymbol{\mu_Y}, \boldsymbol{\Sigma_Y})$, where $\boldsymbol{\mu_Y} = \boldsymbol{\Sigma_0^{-\frac{1}{2}}}(\boldsymbol{\mu} - \boldsymbol{\mu_0})$ and $\boldsymbol{\Sigma_Y} = \boldsymbol{\Sigma_0^{-\frac{1}{2}}}\,\boldsymbol{\Sigma}\,\boldsymbol{\Sigma_0^{-\frac{1}{2}}}$. When the process is in-control, $\boldsymbol{Y_t} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I_p})$, where $\boldsymbol{I_p}$ is a $p \times p$ identity matrix.

3. Compared Methods

Several techniques have been developed in the literature for monitoring variability of a multivariate chart based on grouped observations. In this section, we discuss monitoring methods for grouped observations

a. Monitoring Variability Based on Nonoverlapping Subgroup Data

We choose the generalized variance chart (GVC) proposed by Alt (1984) for monitoring variability of grouped observations because of its popularity. In the GVC chart, the determinant of the sample covariance matrix is plotted against the control limits. The monitoring statistic is equal to

$$det(S_t) = |\boldsymbol{S_t}| \tag{2}$$

where $\boldsymbol{S_t}$ is the sample covariance matrix defined in Eq. (3).

$$\boldsymbol{S_t} = \frac{\boldsymbol{X_{S_t} X'_{S_t}}}{(\boldsymbol{n-1})}, \tag{3}$$

where $\boldsymbol{\bar{X}} = \frac{\sum_{T=1}^{n} \boldsymbol{X_T}}{\boldsymbol{n}}$ and $\boldsymbol{X_{S_t}} = \boldsymbol{X_T} - \boldsymbol{\bar{X}}$ for $T = 1,2,3,\dots,n$. We compare $det(S_t)$ with the control limits;

$$UCL = |\boldsymbol{\Sigma_0}|\left(b_1 + L_1\sqrt{b_2}\right), \tag{4}$$

$$LCL = max\left\{|\boldsymbol{\Sigma_0}|\left(b_1 - L_1\sqrt{b_2}\right), 0\right\}, \tag{5}$$

where b_1 and b_2 are constants (see Montgomery (2013)) and L_1 is a control constant. A signal is obtained whenever $det(S_t) > UCL$ or $det(S_t) < LCL$. We refer to this chart as GVC.

b. Monitoring Variability Based on Overlapping Subgroups

In this section, we modify the MEWMS chart proposed by Huwang et al. (2007) to fit grouped observations and develop a multivariate dispersion charts based on monitoring with overlapping subgroups. Though Sullivan and Woodall (1996) and Holmes and Mergen (1993) employed overlapping subgroups for estimating covariance matrix but we have not seen a chart that applies this technique in Phase II monitoring for the multivariate process variability.

We define the covariance matrix estimator that will be applied for the overlapping subgroups in Eq. (3). The chart we develop uses the trace of the covariance Matrix ($\boldsymbol{S_t}$) for the overlapping subgroup after it is transformed to $\boldsymbol{Y_T}$. The monitoring statistic is defined as

$$tr(\boldsymbol{S_t}) = \sum_{i=1}^{p} y_i^2, \tag{6}$$

where y_i^2 is the diagonal of the matrix. We referred to this chart as TCC. The control limits are given in Eq. (7).

Since $\sum_{t=1}^{p} y_i^2$ follows a chi-squared distribution with p degree of freedom then $E\left(\sum_{t=1}^{p} y_i^2\right) = p$ and $Var\left(\sum_{t=1}^{p} y_i^2\right) = 2p$. We compare the statistic against the upper and lower control limits (UCL and LCL) provided in Eq. (7) to detect signal in the process.

$$LCL/UCL = p \pm L\sqrt{(2p)}, \qquad (7)$$

where L is the control constant.

4. Simulation Study and Performance Criterium

In Section 6, we compare the performance of the selected charts. We simulate for the performance of $p = 2$ (bivariate observations) and $p = 10$ using steady-state ATS as the performance measure. Under this condition, we assume the process is in control state before uniformly random shifts occur within the first 10 and 11 observations for $p = 2$ and $p = 10$ respectively as the start of the shifts before the signal is detected. Monte Carlo simulation is employed to evaluate the average of 50000 time to signals for each method. We represent the shifts in the process variability with δ.

5. Effects of Subgroup Size

Figure 1 displays the ATS curve GVC and TCC charts for various subgroup sizes. Charts are designed for both $p = 2$ with $ATS_0 = 370$. Here, under the out-of-control case, we assume overall shifts in the process variability and the correlation coefficient is unchanged.

We observe from Figure 1 that the ATS_1 value for each of the charts reduces as the subgroup size (n) increases. This indicates that the charts improve as n increases when $p = 2$ for the small and intermediate shifts ($\delta < 3$) in the process. However, we notice that small subgroup size is effective to detect very large shifts in the process ($\delta > 3$).

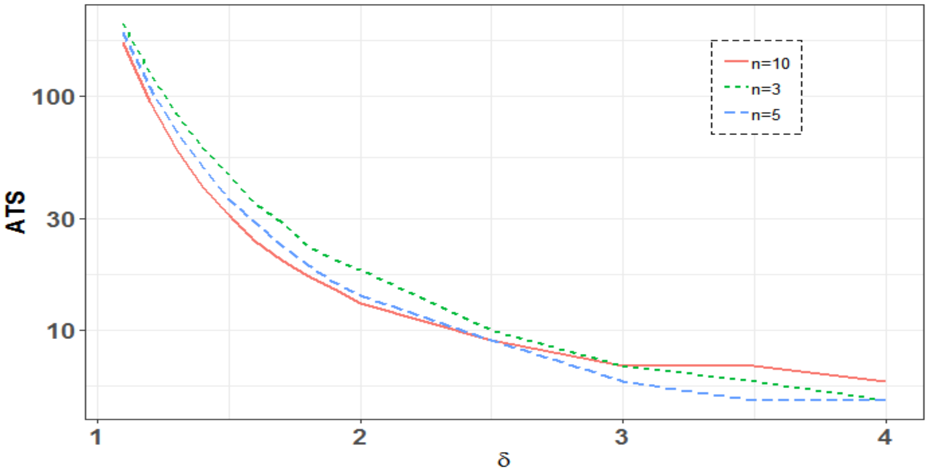


Fig. 1(a): TCC chart for different subgroup size when $p = 2$

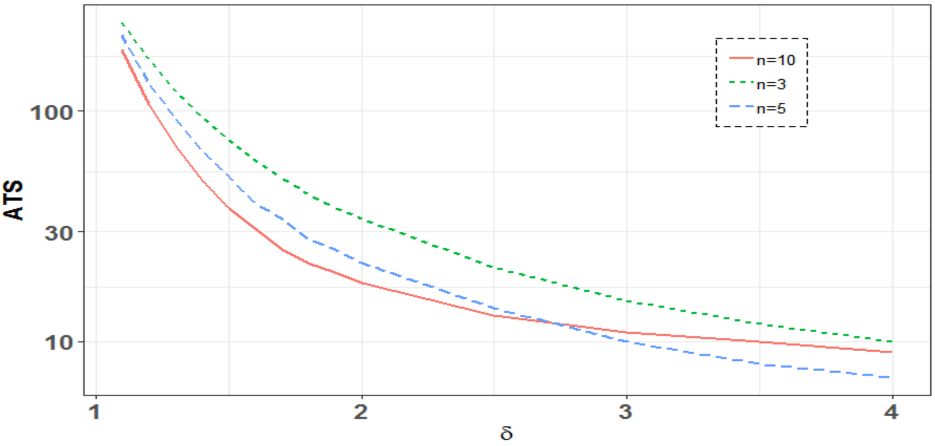


Fig. 1(b): GVC chart for different subgroup size when $p = 2$

6. Performance Comparison

This section compares the performance of the TCC, and GVC charts based on out-of-control ATS values when their $ATS_0 = 370$. We compared the charts based on the subgroup size of 10 and 11 for $p = 2$, and $p = 10$ respectively. We assumed $\rho = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$ when the process is in control for the bivariate normally distributed observations and the covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}. \tag{8}$$

For an out-of-control scenario in Table 4, we simulated an overall increase in the process variances and shifts in the correlation coefficients (ρ =0.6). However, in Table 5, we considered only shifts in the second variance (σ_2^2) of the process. We noticed in Table 4&5 that TCC chart performs better than the GVC. Surprisingly, the ATS values of GVC chart is extremely high when ρ increases. We denote the values of $ATS > 10000$ with $*$ in Tables 4 through 6.

Table 4: ATS for multivariate dispersion charts under an overall process shifts for $p = 2$ ***where*** $\delta = \sigma_1^2 = \sigma_2^2$***,*** $L = 0.97$ ***and*** $L_1 = 2.55$***.***

		TCC	GVC			TCC	GVC
ρ	δ	n =10	n =10	ρ	δ	n =10	n =10
0	1.0	**374**	372	0.6	1.0	**192**	3208
	1.2	**94**	106		1.2	**69**	522
	1.4	**41**	49		1.4	**36**	161
	1.6	**24**	30		1.6	**23**	74
	2.0	**13**	17		2.0	**14**	30
	2.5	**9**	12		2.5	**10**	17
	3.5	**7**	10		3.5	**7**	11
	4.0	**6**	10		4.0	**6**	11

Table 5: ***ATS for multivariate dispersion charts under a single process shifts for*** $p = 2$ ***where***
$\delta = \sigma_1^2$***,*** $L = 0.97$ ***and*** $L_1 = 2.55$

		TCC	GVC			TCC	GVC
ρ	δ	n =10	n =10	ρ	δ	n =10	n =10
0	1	**370**	373	0.6	1	**191**	3210
	1.2	**167**	191		1.2	**107**	1211
	1.4	**90**	117		1.4	**68**	581
	1.6	**57**	80		1.6	**47**	335
	2	**30**	48		2	**28**	150
	3	**13**	24		3	**14**	50
	3.5	**11**	20		3.5	**11**	36
	4	**9**	17		4	**10**	29

We also considered the performance of the charts when $p = 10$ in Tables 6&7 where we assumed $\rho = 0$, $\sigma_i^2 = 1$, $\forall\, i = 1,2,3,\dots,p$ for $ATS_0 = 370$ .We next simulated the ATS_1 values for an overall shifts in variances of the quality characteristics in Table 6. We also simulated the ATS_1values in Table 7 for the partial variability shifts in only the first 3 quality characteristics with either $\rho = 0$ or $\rho = 0.6$ but we considered the last 7 variables to have stable variances with $\rho = 0$.

We present the simulation result for the overall shifts and partial shifts in the process variability for $p = 10$ in Table 6 &7 respectively. In both tables we also noticed that GVC chart performs worse than TCC.

Table 6: ATS for multivariate dispersion charts under an overall process shifts for $p = 10$ *where* $\delta = \sigma_i^2\ \forall\ i = 1,2,3,\dots,p$, $L = 0.8655$ *and* $L_1 = 0.66$.

		TCC	GVC			TCC	GVC
ρ	δ	n =11	n =11	ρ	δ	n =11	n =11
0	1.0	**371**	370	0.6	1.0	**27**	*
	1.2	**36**	85		1.2	**21**	*
	1.4	**13**	38		1.4	**14**	*
	1.6	**9**	24		1.6	**11**	2825
	2.0	**7**	15		2.0	**8**	217
	2.5	**6**	12		2.5	**6**	47
	3.5	**5**	10		3.5	**6**	17
	4.0	**5**	10		4.0	**5**	13

Table 7: ATS for multivariate dispersion charts under partial process shifts for $p = 10$ *where* $\delta = \sigma_1^2 = \sigma_2^2 = \sigma_3^2$, $L = 0.8655$ *and* $L_1 = 0.66$

		TCC	GVC			TCC	GVC
ρ	δ	$n = 11$	n =11	ρ	δ	$n = 11$	n =11
0	1.0	**367**	368	0.6	1.0	**207**	1185
	1.2	**155**	221		1.2	**101**	624
	1.4	**72**	150		1.4	**55**	382
	1.6	**41**	111		1.6	**35**	259
	2.0	**19**	72		2.0	**19**	146
	2.5	**12**	49		2.5	**12**	90
	3.0	**9**	38		3.0	**10**	64
	4.0	**7**	27		4.0	**7**	41

7. Real Life Example

In this section, an example is presented to show how the compared charts should be applied. We apply the bivariate datasets from an industrial process to monitor each of the compared charts. The dataset is available from Santos-Ferna´ndez (2013) as well as MSQC package in R. The indust1 and indust2 represent the Phase I and II observations respectively. From indust1, we estimate the process vector mean (μ), sample covariance matrix ($\boldsymbol{S_t}$) and the sample MSSD covariance matrix ($\boldsymbol{MSD_t}$) as $\mu = [4.09804 \quad 7.16239]$, and $\boldsymbol{S_t} = \begin{pmatrix} 0.10779 & 0.07600 \\ 0.07600 & 0.12580 \end{pmatrix}$ respectively. We assume the subgroup size is 5.

Figure 2 depicts that TCC chart detects signals at observations 29 and 30, signals at observation 1-8, 29 and 30. No alarm was triggered for both GVC chart.

When the subgroup size is increased to 10 (the result is omitted), TCC chart signals at observation 29, 30, 31 and 32. This validates our findings in Section 5 that as the subgroup size increases for $p = 2$, there is an improvement in the performance of the TCC chart. Note that it will be difficult to detect signals in the process before the first 10 observation in this case since the monitoring starts at the 10th observation.

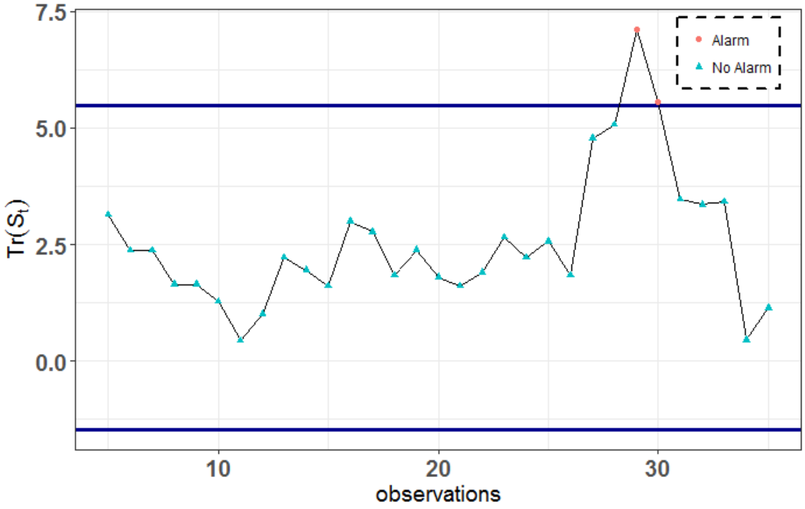


Fig. 2(a): TCC control chart

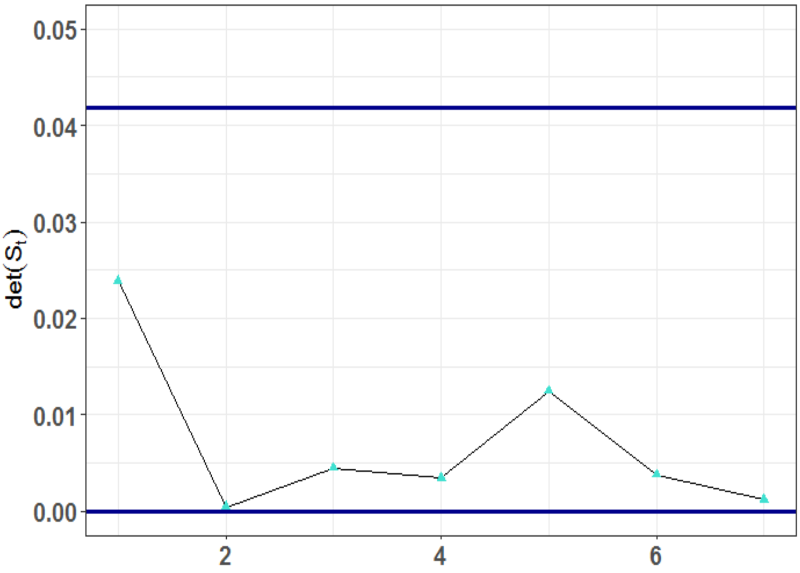


Fig. 2(b): GVC control chart

8. Conclusion and Recommendation

In this paper, we compared the performance of the multivariate charts based on monitoring with overlapping and nonoverlapping subgroups. For the charts based on monitoring with nonoverlapping chart observation, we used the generalized variance chart proposed by (Alt 1984). We developed TCC chart for the chart based on overlapping subgroup. Our comparison shows that the TCC chart has better performance in all the cases we considered We recommend that the practitioner should consider monitoring with a multivariate chart based on overlapping because it shows consistently better performance in respect of the number correlated quality characteristics.

References

1. Alt FB (1984) Multivariate quality control, in The Encyclopedia of Statistical Sciences. Encycl Stat Sci Kotz, S, Johnson, NL Read, CR (eds), Wiley, New York NY 110–122
2. Guerrero-Cusumano JL (1995) Testing variability in multivariate quality control: A conditional entropy measure approach. Inf Sci (Ny) 86:179–202. doi: 10.1016/0020-0255(95)00098-A
3. Holmes DS, Mergen AE (1993) Improving the Performance of the T 2 Control Chart. Qual Eng 5:619–625. doi: 10.1080/08982119308919004
4. Huwang L, Yeh AB, Wu C (2007) Monitoring Multivariate Process Variability for Individual Observations. J Qual Technol 39:258–278
5. Levinson WA, Holmes DS, Mergen AE (2002) Variation charts for multivariate processes. Qual Eng 14:539–545. doi: 10.1081/QEN-120003556
6. Montgomery DC (2013) Introduction to Statistical Quality Control, 7th edn. John Wiley & Sons, Hoboken, NJ.
7. Santos-Ferna´ndez E (2013) Multivariate Statistical Quality Control Using R. Springer-Verlag New York
8. Sullivan JH, Woodall WH (1996) A Comparison of Multivariate Control Charts for Individual Observations. J Qual Technol 28:398–408. doi: 10.1080/00224065.1996.11979698
9. Yeh AB, Lin DKJ (2002) Simultaneously Monitoring Multivariate Process Mean and Variability. Int J Reliab Qual Saf Eng 9:41–59. doi: 10.1142/S0218539302000652

Dengue surveillance using functional data analysis

Wang Zezhong, Inez Maria Zwetsloot
City University of Hong Kong

Abstract

Dengue, a mosquito-borne viral disease, has become a global problem since it affects more than 100 countries, and the incidence of dengue has grown dramatically around the world in recent decades. In previous studies, both statistical and non-statistical methods had been used to analyze dengue related data to do surveillance. The availability of monthly AOI (Area Ovitrap Index for Aedes albopictus) which indicates the extensiveness of the distribution of Aedine mosquitoes in a particular area in Hong Kong offers the opportunity to monitor the behaviour of dengue vector. Since AOI data is collected intermittently at several discrete time points, it can be treated as functional data, so that functional data analysis (FDA) can be applied to display the pattern and detect any abnormal value of AOI data. In this talk, we will review the literature on dengue surveillance. We will give a brief introduction and illustrate the usefulness of FDA. We will elaborate on using functional data to do bio surveillance and use an example from Hong Kong to showcase this methodology.

Keywords

dengue surveillance; functional data analysis; statistical process monitoring; bio surveillance

1. Introduction

Dengue is an arboviral disease, which is transmitted by female mosquito's. According to the statistics from the World Health Organization, the number of cases reported increased from 2.2 million in 2010 to 3.2 million in 2015, and the number of the affected country increased from 9 in 1970 to 100 in 2018 (WHO.2018). Since dengue is an infectious disease without an effective vaccine, real-time surveillance and reliable prediction of an outbreak become crucial.

Dengue surveillance is an application of a bio surveillance system which has two main objectives: to provide outbreak situational awareness (SA) and to enhance outbreak early event detection (EED) (Fricker et al. 2018). The Unites States Centers for Disease Control and Prevention (CDC) defines SA as the ability to use detailed, real-time health data to confirm or refute, and to provide an effective response to, the existence of an outbreak. And EED is the ability to detect, at the earliest possible time, events that may signal a public health emergency.

Some research had been done in dengue surveillance focusing on the first objective of situational awareness. For example, Buczak et al. (2012, 2014) used fuzzy association rules to analyze the relationship between dengue incidences and meteorological data in Philippines and Peru, and to predict the outbreak. Ramadona et al. (2016) used a generalized linear regression model to predict dengue outbreaks based on dengue cases and climate data. However, go to et al. (2013) analyzed the effects of meteorological factors on dengue incidence in Sri Lanka by using time series data and showed temperatures and the rainfall did not significantly affect dengue incidences. But the conclusions based on incidence and meteorological data are inconsistent.

Most of the current research focussed on prediction, which belongs to the first objective of bio surveillance; situational awareness. In our research, we will focus on another objective of bio surveillance; early event detection. Early event detection can help in containing a potential outbreak of dengue. We use a new type of data, monthly AOI (Area Ovitrap Index for Aedes albopictus), which indicates the extensiveness of the distribution of Aedine mosquitoes in a predefined area of Hong Kong (Fehd.gov.hk. 2019).

AOI data, which is collected intermittently at several discrete time points, can be treated as functional data (Wang, J. 2016). Therefore functional data analysis (FDA), which models data using functions or functional parameters (Ramsay and Silverman, 2005), can be applied to represent AOI data so that the pattern and variation of the data can be studied and explained. We combine FDA modelling with tools from statistical process monitoring (SPM) to design an early event detection system for signalling abnormal increases in the AOI level in Hong Kong.

In the flowing, we will first introduce and describe the data, next in Section 3 we describe our methodology based on use FDA to model the AOI data and design the early event detection system based on a control chart to detect abnormal values in the modelled data. In Section 4, we illustrate this early event detection system by applying it to the 2018 AOI data and detect an increase in the vector one month before the dengue outbreak that hit Hong Kong in the summer of 2018.

2. Dengue and AOI in Hong Kong – data description

According to statistics from the Hong Kong Department of Health, there were 163 cases of dengue fever in 2018, of which 29 were local cases in August (Department of Health, 2019). All 29 local cases were reported in August 2018. To monitor indicators related to the dengue epidemic, the Food and Environmental Hygiene Department in Hong Kong has been using Oviposition Trap (Ovitrap) to detect the presence of adult Aedine mosquitoes, the most

common infectious vector in Hong Kong, in selected areas for vector surveillance since 2000 (Fehd.gov.hk. 2019).
The monthly AOI per area in Hong Kong is calculated as:

$$Ovitrap\ Index\ for\ Aedes\ Albopictus\ in\ a\ selected\ area$$
$$= \frac{Number\ of\ the\ Aedes\ positive\ ovitraps\ in\ this\ area}{Total^{\textit{number of ovitraps retrieved from this area}}} * 100\%$$

For our research purpose, we use MOI, the average AOI over all areas within a month, to reflect the territory-wide situation of Aedes albopictus (Fehd.gov.hk. 2019).

Figure 1 displays the MOI data and we observe a seasonal pattern; during the winter the MOI is (near) zero, it starts to rise in spring and hits is maximum in the summer when Hong Kong is very hot and humid after which the MOI starts to decline in the fall.

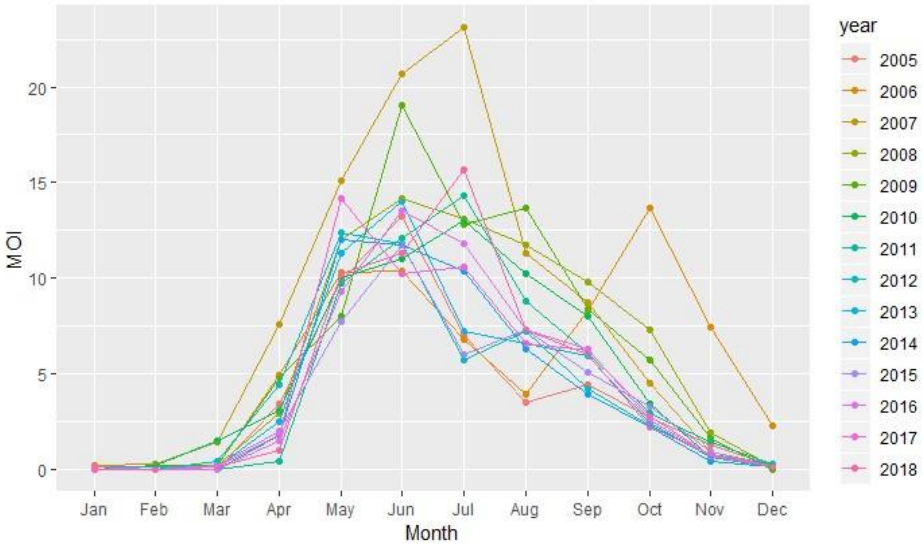


Figure 1: MOI Data in Hong Kong in the period 2005-2018

3. Methodology

To detect any abnormal values in the MOI data as they come in month by month, we first need to model the general seasonal pattern. The MOI can be represented by a function of the months, hence will use FDA to model the MOI (section 3.1). Next in section 3.2, we design a control chart to monitor the difference (residuals) between new incoming data month and the estimate FDA model.

3.1 Model MOI using Functional Data Analysis

As we mentioned before, functional data is often recorded continuously over a selected time interval or intermittently at several discrete time points (Wang, J. 2016), which has an ordering on time. Define y_{ij} as the MOI value in year i = 2005,… ,2018 and month j = 1,2, … ,12 where j = 1 represents January. We wish to use the year 2005 up to 2017 to estimate the seasonal pattern. To reduce the noise we work with the average MOI for each month, defined as $y_j = \frac{1}{13}\sum_{i=2005}^{2017} y_{ij}$.

The observations y_j can be represented as a function of time with an error ϵ_j which follow normal distribution with 0 mean; $y_j = x(t_j) + \epsilon_j$. The functions $x(t)$ is the functional data, which can be represented as:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = C'\phi(t) \qquad (1)$$

where K is the number of basis functions, C is the vector of K coefficient, and $\phi(t)$ is the vector of basis functions. Since the MOI data is seasonal data we use a Fourier basis for $\emptyset(t)$. The Fourier series is: $\emptyset_1(t) = 1, \emptyset_2(t) = \sin(\omega t), \emptyset_3(t) = \cos \omega t$, … where $\omega = \frac{2\pi}{T}$, is the time period. In the Fourier basis the number of basis functions K must be an odd integer. For details see Ramsay and Silverman (Chapter 5, 2005).

In order to control for overfitting, we add a roughness penalty to the least square least square criterion:

$$F(C) = \sum_j [y_j - x(t_j)]^2 + \lambda \int [Lx(t)]^2 dt \qquad (2)$$

where λ is smoothing parameter and L is the differential operator. For the Fourier basis, it is standard to choose the harmonic acceleration operator such that we have $L = \omega^2 D + D^3$. We set $\lambda = exp(-3)$ as this value minimizes the cross validation term (see Chapter 5, Ramsay and Silverman (2005) for details). Now we use the least square estimate to obtain estimates of C by minimizing Eq (2) subject to Eq (1). This gives us the following fitted fuction:

$$x(t) = \hat{C}'\phi(t) \qquad (3)$$

Figure 2 shows the fitted function $x(t)$ as well as the average MOI data y_j (the dots).

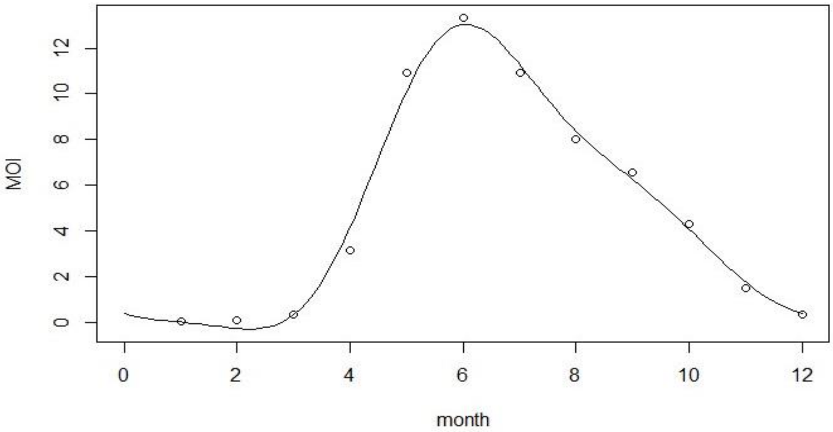


Figure 2: *Fitted FDA model. Line is x(t) and the dots are the average MOI data points.*

3.2. Control chart to detect abnormal values

Next we need a tool to compare online new observations with the modeled seasonal patter, for this we use a control chart, one popular tool from statistical process monitoring (SPM). SPM has historically been applied to online monitor of production processes in manufacturing industries. The shewhart control chart is the most basic chart, it plots the data together with two bounds: an upper and lower control limit. Once the data exceeds these bounds a signal is observed.

One active subfield is the monitoring of so-called profiles. A profile is characterization of a relationship between a response variable and series of values of an explanatory variable, profiles can also be seen as functional data. Woodall et al. (2004) review the literature and discuss general issues involved in using control charts to monitor profiles.

Control charts have also been adapted to and applied in bio surveillance. Overviews can be found in Woodall et al (2010) and Fricker (2011). Because bio surveillance data is not stationary, the control charts have to be applied in two steps: first we model the systematic effects in the data, the purpose of this step is to remove the systematic effects from the data and create data which are approximately stationary. The second step is to monitor the (standardized) residuals and/or the model parameters (Fricker, 2013). A Shewhart control chart is often used in this step, following the "one-sided" scheme for monitoring the mean incidence. This entails choosing a threshold h and plotting it together with the sequence of residuals (z_t) over time. As long as $z_t < h$, we assume that there is no evidence of an outbreak. However, if $z_t \geq h$, the chart signals that an outbreak may be occurring.

We take this latter approach and calculate residuals according to $res_j = y_{ij} - x(t_j)$. Where $y_{ij}$is the new data point coming in in year j and month i and $x(t_j)$ is the expected value based on Eq (3). We plot these residuals on a chart with the upper control limit $UCL = 3 * S_{res}$. Where we compute the standard deviation of the residuals as the pooled samples standard deviation of the original data:

$$S_{res} = \sqrt{\frac{1}{12}\sum_{i=1}^{12} S_i^2}$$

Where S_i^2 is the sample variance of the MOI data in month i in the years 2005 to 2017.

4. Result detecting abnormal MOI values

We know that an outbreak of dengue occurred in Hong Kong in August 2018. This outbreak in August was partially local (29 cases) and partially imported (134 cases). The local cases are due to the spreading of dengue through mosquito bites. This mosquito population is measured through measuring the AOI in multiple regions throughout Hong Kong. We compare the average AOI (MOI) to the estimated functional model and plot the residuals on a control chart as designed in section 3.2. Figure 3 shows the control chart. We observe a signal in July, indicating a higher than usual MOI level. This is the month before the dengue outbreak in Hong Kong in August 2018.

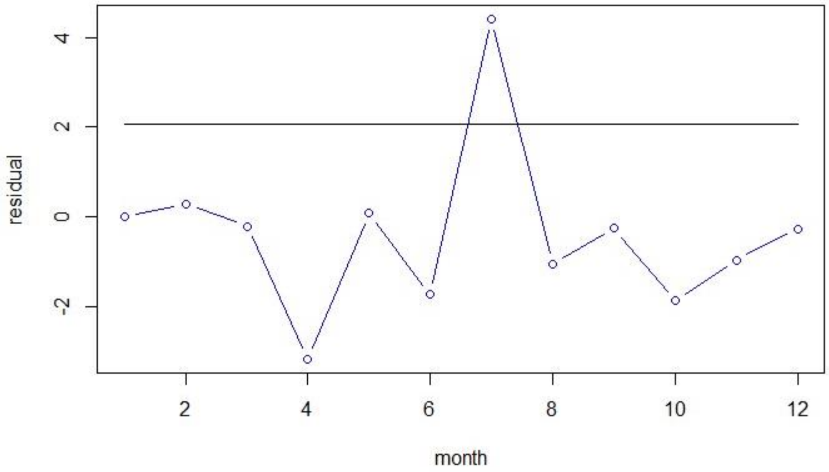


Figure 3 *Control Chart for MOI in 2018. A signal is observed in July.*

5. Conclusion

We develop a early outbreak detection system for monitoring online the MOI data in Hong Kong. The method described above is based on functional data analysis and control charting. We use aggregated AOI data to model dengue outbreaks and achieve real-time monitoring, with an outbreak defined as a signal of out of control. Effective methodologies to predict and monitor disease outbreaks may allow preventive interventions to avert large epidemics.

As AOI data are spatially located, the use of disaggregate AOI data could benefit the systems so that we could know the time of dengue outbreak as well as the particular area. A limitation of this method is that we focus on AOI, which only helps detect local outbreaks. Any people carrying dengue from overseas will not be included.

References

1. Aditya Lia Ramadona, Lutfan Lazuardi, Yien Ling Hii, Åsa Holmner, Hari Kusnanto, & Joacim Rocklöv. (2016). Prediction of dengue outbreaks based on disease surveillance and meteorological data. PLoS ONE, 11(3).
2. Buczak, A., Koshute, P., Babin, S., Feighner, B., & Lewis, S. (2012). A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Medical Informatics and Decision Making, 12(1), 124.
3. Buczak, A., Baugher, B., Babin, S., Ramac-Thomas, L., Guven, E., Elbert, Y., Carvalho, M. (2014).
4. Prediction of high incidence of dengue in the Philippines (Prediction of high incidence of dengue). 8(4).
5. Department of Health. (2019). Statistics on dengue fever, 2018 (English). Retrieved from https://data.gov.hk/en-data/dataset/hk-dh-chpsebcdde-dengue-fever-cases/resource/898d1406-3451-4c5ab297-7e8f732e7fac
6. Fehd.gov.hk. (2019). Dengue fever. Retrieved from Food and Environmental Hygiene Department https://www.fehd.gov.hk/english/pestcontrol/dengue_fever/index.html.
7. Fricker Jr, R. D. (2011). Some methodological issues in biosurveillance. Statistics in Medicine, 30(5), 403-415.
8. Fricker, R. (2013). Introduction to statistical methods for biosurveillance: With an Emphasis on Syndromic Surveillance. Cambridge: Cambridge University Press.
9. Fricker, R., & Rigdon, S. (2018). Disease surveillance: detecting and tracking outbreaks using statistics. CHANCE, 31(2), 12-22.
10. Goto, K., Mettananda, S., Gunasekara, D., Fujii, Y., & Kaneko, S. (2013). Analysis of effects of meteorological factors on dengue incidence in Sri Lanka using time series data. PLoS One, 8(5).
11. Megahed, F. M., & Jones-Farmer, L. A. (2015). Statistical perspectives on “big data”. In Frontiers in statistical quality control 11 (pp. 29-47). Springer, Cham.
12. Ramsay, J. and Silverman, B. W. (2005). Functional data analysis. Springer-Verlag.
13. Wang, J.L, Chiou, J.M. & Muller, H.G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1), pp. 257–295.

14. Woodall, W. H., Spitzner, D. J., Montgomery, D. C., & Gupta, S. (2004). Using control charts to monitor process and product quality profiles. Journal of Quality Technology, 36(3), 309-320.
15. World Health Organization (2018), Dengue and severe dengue. Retrieved from World Health Organization: www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue.
16. Woodall, W. H., Grigg, O. A., & Burkom, H. S. (2010). Research issues and ideas on health-related surveillance. In Frontiers in Statistical Quality Control 9 (pp. 145-155). Physica-Verlag HD.

Research of Better Life Index and Empirical Analysis of Beijing

He Jingwei, Zhang Lu, Zhang Jiaojiao
Beijing Municipal Bureau of Statistics, Beijing, China

Abstract

With the rapid development of the economy and society, GDP is no longer the key factor for measuring the life quality of residents. For the statistical work in the new era, it is desperately needed to build a comprehensive indicator that reflects people's livelihood and well-being more comprehensively in excess of GDP. By taking the high-quality development as the entry point, this paper starts with the development objective and process, selects 49 indicators that reflect the current development situation and development features of Beijing to build an indicator system from the perspective of realizing people's better life and then takes the entire-array-polygon evaluation method to objectively measure the non-empowerment of indicator system, thus obtaining the better life index of Beijing City. The measurement results show that the better life index of Beijing City steadily increased from 2010 to 2017, the people's needs for better life were constantly satisfied, the development foundation was continuously strengthened, and the development potential was also sufficient. However, there are still shortcomings in tourist environment, human settlement, people's health and external attraction, so various efforts are still needed for the development.

Keywords

High-quality development; Indicator system; Entire-array-polygon evaluation method

1. Introduction

The constant progress of human development idea is symbolized by "economic growth" to "economic development" and then by "social development" to "all-round development of human beings", which is also an inevitable requirement for realizing the all-round development of human beings. It is shown in the reports of the 19th National Congress of the Communist Party of China that Chinese economy has turned into a high-quality development stage from the high-speed development stage, and the major social contradiction in China has been transformed into the contradiction between the increasing needs of people for a better life and the unbalanced and inadequate development. The past extensive development model and "only GDP" development concept are no longer suitable for the new era of socialism with Chinese characteristics, so it is urgent to build an indicator system that reflects the demand of people for a better life so as to correctly guide the whole society to pay more attention to the fundamental objective of development (i.e.improving people's livelihood and well-being and achieving higher-quality, higher-efficiency, fairer and more sustainable growth).

2. Methodology

2.1 Establishment of Indicator System

With reference to the research on life quality, sustainable development and high-quality development at home and abroad, in conjunction with Maslow's hierarchy of needs theory, we manifest the development achievements that people finally enjoy with "life quality" and the necessary support for realizing a better life with "solid foundation" according to the spirit in the reports of the 19th National Congress of the Communist Party of China as well as the urban function positioning of "four centers" in Beijing City[1] so as to build a better life index system for Beijing City.

Table 1 Better Life Index System for Beijing City

First-class indicator	Second-class indicator	Third-class indicator
Life quality	Food, housing and transportation	Per capita protein intake
		Per capita number of eating out
		Proportion of organic food consumption
		Per capita rental expenditure ★
		Per capita housing floor area for urban residents
		Per capita commuting time ★
		Average daily congestion duration ★
	Medical treatment, endowment, health and security	Number of licensed practicing physicians (assistants) per thousand
		Number of beds in nursing institutions for the aged per thousand
		Juvenile myopia rate★
		Number of patients with hypertension, hyperlipidemia and hyperglycemia★

[1] Beijing is the political center, cultural center, international exchange center and science and technology innovation center of the country.

First-class indicator	Second-class indicator	Third-class indicator
		Urban minimum living guarantee standard
	Culture, education, and life quality	Number of visitors to the museums
		Number of people borrowing books and literature from libraries
		Number of people who watch theatrical performances
		Per capita expenditure on education, culture and entertainment for urban residents
		Greenway length
		Per capita park green space
	Subjective feeling	Urban security
		Harmony and livable satisfaction
		Social governance satisfaction
Solid foundation	Innovation, improving efficiency, and	R&D investment strength
		Number of overseas students and returnees in Zhongguancun national independent innovation zone
First-class indicator	Second-class indicator	Third-class indicator
	adjusting structure	Proportion of added value of sophisticated industry
		Proportion of added value of modern service industry
		Proportion of import and export of high-tech products
		Proportion of service consumption in total consumption
		Energy consumption of GDP in ten thousand yuan areas★
		Water consumption of GDP in ten thousand yuan areas★
		Land consumption of GDP in one hundred million yuan areas★
	Enhancing influence	Number of headquarters of multinational enterprises

First-class indicator	Second-class indicator	Third-class indicator
		Number of international conferences
		Number of cultural exchanges with other countries
		Number of inbound tourists to Beijing
		Number of QS world universities
		Number of state key laboratories
		Number of 5A tourist attractions
		Number of group series events
		Total enrollment of students in 985 universities
	Improving people's livelihood	Proportion of senior teachers in basic education
		Student and teacher ratio in basic education stage★
		Community coverage of garbage classification
		Length of composite pipe gallery
		Number of parking spaces filed
		Coverage of 15 minutes' community service circle
		Number of elevators installed in old residential areas
	Improving environment	Number of days for air quality compliance
		Forest coverage
		Proportion of water bodies better than Class III

Note: the item marked with "★" refers to an inverse indicator, and the remaining are positive indicators.

2. 2 Model and Method

2.2.1 Data source: The research data mainly come from the following 4 platforms: (1) the statistical publications such as China Statistical Yearbook and Beijing Statistical Yearbook; (2) official reports on departmental websites such as eBeijing, Beijing Municipal Bureau of Culture and Beijing Municipal Bureau of Sports; (3) independent statistical survey data from Beijing Municipal Bureau of Statistics (subjective perception survey); and (4) data published by international official websites such as ICCA and QS.

2.2.2 Measurement model: As various indicators in the indicator system are not mutually independent, priority should be given to multiplicative model in terms of measurement model. Therefore, the entire-array-polygon evaluation method is introduced for this research, in which the traditional addition is changed into the multi-dimensional multiplication. By setting the critical value, amplification and contraction effects are exerted on the comprehensive indicator.

The normalization of the entire-array-polygon evaluation method with fully arranged polygons is processed into the hyperbolic normalized functions:

$$\mathrm{F(x)} = \frac{a}{\mathrm{b}x + \mathrm{c}}$$ Formula (1)

$$\mathrm{F(x)} = \begin{cases} -1, & \mathrm{x = L} \\ 0, & \mathrm{x = T} \\ 1, & \mathrm{x = U} \end{cases}$$ Formula (2)

In the above formula, U is the upper limit of the indicator x, L is the lower limit of the indicator x and T is the critical value of the indicator A (usually the average value). According to the above three conditions, the following can be obtained: $\mathrm{F(x)} = \frac{(U-L)(x-T)}{(U+L-2T)\,x+(U+L)T-2UL} x \geq 0$
Formula (3)
For the indicator x_i, the normalized calculation formula is:

$$S_i = \frac{(U_i-L_i)(x_i-T_i)}{(U_i+L_i-2T_i)\,x_i+(U_i+L_i)T_i-2U_iL_i}$$ Formula (4)

In the above formula, L_i, T_i and U_i are respectively the lower limit, critical value and upper limit of the indicator x_i. According to related research, the lower limit, critical value and upper limit mentioned in this paper are respectively the minimum value, average value and maximum value of the indicator x_i.

The calculation formula for the comprehensive indicator of entire-array-polygon evaluation method is provided as follows:

$$S = \frac{\sum_{\mathrm{i \neq j}}^{\mathrm{i,j}} (S_i+1)(S_j+1)}{\mathrm{2n(n-1)}}$$ Formula (5)

In the above formula, S_i and S_j respectively refer to the normalized value of the indicator i and the indicator j under the same indicator directory, and S is the comprehensive indicator value and its value range is [0, 1].

3. Results

Based on the entire-array-polygon evaluation method and the calculation result of each indicator, the following three conclusions are made:

The first is the continuous improvement of better life index. The better life index of Beijing City was increased from 0.03 in 2010 to 0.51 in 2017, and the development speed has shown an accelerating trend since 2013. In recent years, great efforts have been made in terms of urban environmental governance, ecological improvement and livelihood projects in Beijing City. For example, in the action of "landscape engineering on the vacated land", the green public leisure space with different characteristics is formed through the reasonable planning of idle and vacated land so as to effectively improve the environmental quality and also improve and beautify the human settlement environment; in the construction of "15 minutes' community service circle", the public service stations such as community service station, police affairs station and health service station are established by integrating and optimizing various convenience service stations such as vegetable market, breakfast shop and convenience store according to the residents' actual demand, and the online and offline services are carried out by using the "Internet+" technology so as to promote the high-quality life with high-quality service and effectively improve people's sense of gain.

The second is the excellent performance in "solid foundation". According to the measurement results of comprehensive graphic method with fully arranged polygons, the "life quality" indicators in 2010 and in 2011 were higher than the "solid foundation" indicator; and the indicators in 2012 and in 2013 were identical; since 2013, the "solid foundation" indicator has risen dramatically to exceed the "life quality" indicator. To lay a solid foundation, the new urban governance plans such as continuous improvement of ecological environment and garbage classification of comprehensive underground pipe gallery and communities not only aim to improve the development quality in the current stage but also serve the future sustainable development.

The third is the slight regression of some aspects, which is specifically manifested as follows:

(1) the "per capita commuting time" and "average daily congestion duration" are both very long, respectively 50 minutes and 160 minutes, and the "per capita rental expenditure" continuously rises, almost doubled in eight years; and (2) the external attraction is slightly weakened, and the "number of inbound tourists to Beijing" continuously decreases. To continuously meet the people's needs for a better life and promote the high-quality development of the whole city, it is recommended to increase the input in public transportation and promote the job-housing balance, develop lease and purchase simultaneously to promote the supply and standardize the market

to maintain stability, and perfect the international service environment and improve the urban attraction.

Figure 1 Changes in Better Life Index of Beijing City from 2010 to 2017

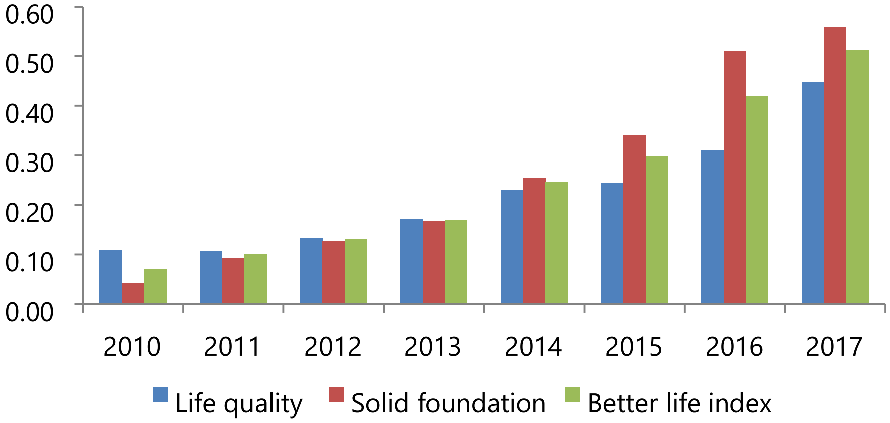


Table 2 Decrease of Selected Indicators for the Better Life Index of Beijing City from 2010 to 2017 under the Comprehensive Graphic Method with Fully Arranged Polygons

Three-level indicator	2010	2011	2012	2013	2014	2015	2016	2017
Per capital rental expenditure	1.00	0.74	0.48	0.20	-0.09	-0.38	-0.65	-1.00
Per capita commuting time	1.00	1.00	1.00	1.00	1.00	1.00	-0.45	-1.00
Average daily congestion duration	1.00	1.00	0.60	0.13	0.13	-1.00	-0.92	-0.67
Juvenile myopia rate	1.00	0.69	0.28	-0.29	-0.46	-0.66	-0.82	-1.00
Proportion of import and export of high-tech products	1.00	0.09	-0.48	-1.00	-0.94	0.00	0.26	-0.82
Number of inbound tourists to Beijing	0.57	1.00	0.73	-0.03	-0.40	-0.52	-0.58	-1.00

4. Discussion and Conclusion

In comparison with the former research, we take the following methods: first, combine the achievement of people's better life with the high-quality development, and embody the new era development concept in the indicator system; second, highlight the urban positioning of Beijing City during the selection of indicators, and make the indicator system faithfully reflect the progress of better life and development weakness according to the current economic and social development issues of Beijing City; third, place emphasis on "people-centered", and select the prospective indicators that reflect the life quality and manifest "people first" in an exploratory way to provide reference for future relevant statistical systems; fourth, introduce subjective indicators such as satisfaction, and comprehensively measure the intuitive perception of people for a better life through the combination of objective and subjective indicators; and fifth, take the entire-array-polygon evaluation method, expand the applicable range of model, and enhance the measurement accuracy and reliability.

For the statistical work in the new era, it is required to ensure the statistical quality, effectively expand the statistic range and focus on human development; increase the input, make full use of existing investigation methods, and extensively collect existing indicator data; promote the statistics reform and transformation of related areas, organically combine the big data thinking with the traditional investigation, and effectively obtain characteristic and prospective indicator data, so as to monitor the people's needs for a better life in a more scientific and reasonable manner and meet the people's needs for a better life.

References

1. Xi Jinping. Strive for Great Victory for the Socialism with Chinese Characteristics for a New Era in Building a Moderately Prosperous Society in all Respects. A Report Given at the 19th National Congress of the Communist Party of China [EB/OL]. http://cpc.people.com.cn/n1/2017/1028/c64094-29613660.html,2017-10-28.
2. Pan Jiancheng. How to Monitor the Better Life and Imbalance Insufficiency [J]. China Statistics, 2018(5): 4 - 6.
3. Li Hongjie, Zhang Yuanzhao. Evaluation of National Life Quality by OECD and Its Enlightenment to China [J]. FUBBS.CN (humanistic and social science edition), 2018 (02): 5 - 11.
4. OECD. OECD better life initiative. Compendium of OECD well-being indicators. Organization of Economic Cooperation and Development, 2011.
5. Wu Qiong, Wang Rusong, Li Hongqing, Xu Xiaobo. Ecological City Indicator System and Evaluation Method [J]. Acta Oecologica, 2005 (08): 2090 - 2095.
6. Sun Xiao, Liu Xusheng, Li Feng, Tao Yu. Comprehensive Evaluation of Sustainable Development of Cities of Different Sizes in China [J]. Acta Oecologica, 2016, 36 (17): 5590 - 5600.

Research on Building a Modern Statistical Investigation System

Liu Lufang
Beijing Municipal Bureau of Statistics, Beijing, China

Abstract

With the accelerated development of the new economy, the economic forms and social phenomena that need to be reflected and disclosed by statistics are complicated and changeable. Big data and other new information technologies have put forward higher requirements for the completeness, promptness and accuracy of official statistics, and the official statistics have entered a new stage of promoting development through reform and promoting transformation through innovation. Based on the understanding of modern statistical investigation system, this paper sorts out the problems in traditional statistical investigation system. According to the economic and social development of Beijing City, this paper puts forward a main frame for the modern statistical investigation system to lay a solid foundation for the constant development of statistical work in Beijing.

Keywords

Official statistics; modern; statistical investigation method

1. Introduction

In recent years, the development of new economy and new drivers has been accelerated. New industries and new business forms such as sharing economy and collaborative economy have been constant emerged. The economic forms and social phenomena that need to be reflected and disclosed by statistics are complicated and changeable. The modern technologies such as big data and cloud computing result in a lot of objective real-time data. This raises a higher requirement for completeness, promptness and accuracy of official statistics. It is urgent to solve how to integrate the traditional data production modes with the modern information technologies such as big data. Therefore, it is of great significance to build a modern statistical investigation system based on the actual development of economy and society.

2. Knowledge about the Modern Statistical Investigation System

Modernization is a worldwide trend of continuous development and progress, and a historical change in the transformation from traditional society into modern society. Seen from the Chinese context of modernization, it is a comprehensive systematic project, including modernization of science and technology, economic modernization, political modernization, social modernization and human modernization. The statistical investigation system is the basis and carrier for carrying out the statistical work and generating the statistical data. It is a whole set of organization and management system and institutional method system for management and organization of the statistical investigation, including methodological systems, organizational systems, regulatory systems, and resource allocation. The modern statistical investigation system is the modernization of statistical investigation system. By taking the comprehensive, accurate and prompt reflection of economic and social development situations as an objective, and the perfect statistical management system, statistical legal system and efficient statistical production process as a guarantee, the modern statistical investigation system supported by the modern information technology aims to realize the modernization of statistical system, statistical indicator, statistical method, statistical means, statistical product, statistical service and statistical guarantee.

Perfect the statistical classification, expand the statistical content and realize the full coverage of statistics over economy, politics, culture, society and ecological civilization. Better reflect the statistical indicator system involving structure, quality, benefit and sustainable development, and give full play to the effect of statistical indicator system as "weather vane" and "navigator". Reallygive full play to the basic role of the census, further strengthen the dominant position of sample investigation and comprehensively take such methods as consolidated statement, key-point investigation, administrative record and scientific reckoning to reduce the statistical investigation cost, decrease the burden at the grass-roots level and continuously improve the statistical efficiency. Improve the statistical means, combine the traditional data production mode with the modern information technology, and make great efforts to form an "internet+" statistical production mode based on egovernment, e-commerce record and extensive use of big data. Make full use of the statistical data as "gold mine" and promote the transformation of statistical resources fromsegmentation to information sharing and the diversification, visualization and facilitation of statistical products.

3. Problems in the Traditional Statistical Investigation System

3.1 Advantages and Disadvantages in Statistical Management System

According to the *National Statistical Organization Manual* issued by the United Nations Statistics Division, the government-level statistical management system is divided into three types, namely centralized type, decentralized type and mixed type. Under the centralized type statistical system, the official statistical system is complete and unified, and the business division with the government departments is specific and not mutually overlapped, but statistics are divorced from administration due to huge central bureau of statistics and heavy burden on internal management and coordination. The decentralized type statistical system can make statistics closely combined with the administrative departments, reduce the statistical demand contradiction, and make full use of statistics in state administrative decision-making, but its disadvantages are manifested by repeated statistics, data coming from various sources, inconsistent statistical specifications and poor cohesion of statistical data between the departments.

3.2 Hysteresis of Statistical Investigation Methods

Currently, the statistical management system under unified leadership and hierarchical responsibility is implemented in China, and the local statistics follow the methodological system, indicator system, data management system, management system and legal system of national statistics, etc. With Beijing's economic and social development entering a period of new normal as well as the constant emerging of new economic forms and social phenomena, the new requirements and formulations for economic and social development from central to local are increasing, and the challenges are enormous from respondents to statistical personnel and from data provider to data demander. However, the current statistical investigation system is still quite unfit in dealing with new challenges and moving towards a new era, etc.

First, the statistical content is hard to satisfy the development characteristics in the new era, and the statistical data lay emphasis on economy over society, total over structure and speed over efficiency. There is no efficient statistics for new economic forms in the economic field, the statistics in the biological field is still in the research and development stage, and there is still great space to develop statistics in the social field.The update speed and efficiency of statistical content lag behind the pace of economic and social development, and the statistical content is hard to accurately reflect the new situations and problems such as development of emerging industries and replacement of old kinetic energy with new kinetic energy and is also hard to reflectthe real conditions about cross-border operation and mixed operation of enterprises.

Second, the data sharing mechanism is still insufficient in terms of the support from policies and regulations, business coordination and linkage, technical standards and specifications and sharing platform construction, etc. Various government departments have established their own data reporting platforms and databases, but the data between the departments is hard to be shared, so many "isolated information islands" have formed to result in low use efficiency of statistical data and grievous waste of resources. Besides the imperfect data sharing mechanism, thenon-unified technical standards and regulations and the unestablished"soft technologies" such as data sharing scope, metadata standard, data sharing process and management rules, each department still does things in its own way in the sorting process of data resources, there is a great variety of storage formats,the interface parameters are changeable, the statistical standards of some indicators are different, and the query positioning efficiency is low.

Finally, the application of modern information technology is insufficient. There is technical weakness in data mining, development, analysis and utilization, the networked data collection methods fail to cover all the statistical business, and some investigation still adopts the traditional methods such as interview instead of networked, integrated and automatic processing and analysis software. The statistical investigation methods are unable to keep up with changes in the enterprise, the monitoring of small and medium-sized enterprises involving new economy is insufficient in the non-census years, many small and micro businesses show such features as broad body distribution, small scale, quick change and difficult monitoring, and it is hard to follow up the management of directory information base in time. Some technical research and development encounters bottlenecks. For example, the remote sensing technology is hard to meet the statistical requirements of small-variety crops and its application has certain limitation in agricultural statistics. The extraction of remote sensing information mainly relies on manual visual interpretation, and the experience of interpretation personnel always affects the measurement results. The technical methods such as intelligent identification still require further research and development.

4. Main Frame for Building a Modern Statistical Investigation System in Beijing City

Based on such basic content as perfect and efficient statistical organization system, comprehensive and scientific statistical institutional system, standardized and fair statistical legal system, convenient and shared data management system, optimized and advanced technical guarantee system as well as flexible and convenient statistical service system, the building of modern statistical investigation system in Beijing should be supported by modern information technology means, focusing on actual development of economy and society and aiming at high-efficiency, all-around, multi-angle and deep-rooted services.

4.1 Perfect and efficient statistical organization system

The modern statistical investigation system should be based on perfect and efficient statistical organization system. The statistical department should integrate the functions of statistical agencies, rationalize the division of responsibilities and optimize the business process according to the objective needs of economic and social development so as to provide an institutional guarantee for the high-quality accomplishment of statistical work. Specifically, based on the business process under the model of "unified enterprise reporting", the functions of statistical agencies are integrated, and the same or similar responsibilities that are dispersed in different departments are integrated and reconstructed to improve the systematic and collaborative properties of statistical business process. In terms of the statistics of relevant departments, it is required to establish a departmental data review and evaluation system to effectively improve the quality of departmental statistics; establish a departmental coordination mechanism to specify the division of responsibilities between governmental statistics and departmental statistics as well as the data sharing mode; perfect the working mechanism between various departments, standardize and unify the statistical reporting system, indicator system, statistical standard and basic unit, etc. and finally make the comprehensive governmental statistics and the departmental statistics form an organic whole, mutually supported and promoted, complement each other and make concerted efforts.

4.2 Comprehensive and scientific statistical institutional system

The modern statistical investigation system should be guided by a comprehensive and scientific institutional method system. For this purpose, it is required to enhance the top-level design and establish a systematic and standardized statistical institutional system; oriented by demand, enrich and perfect the statistical investigation system and gradually realize the full coverage of statistical system over economy, politics, culture, society and ecological civilization construction; oriented by result, fully plan various kinds of statistical investigation and enable the statistical information resources to complete each other according to the actual development of economy and social; establish a sound statistical investigation evaluation mechanism to evaluate the feasibility and applicability of investigation items; fully consider the linkage between general investigation and conventional investigation and between general investigations to improve the overall effectiveness of statistical investigation; and make full use of departmental administration records and data to be mutually complemented with the conventional statistical data.

It is also required to perfect the statistical standard and statistical indicator system and improve the applicability of statistical standard; and highlight the concept of comprehensive statistics at the macro level, take specific measures for the respondents classified at the micro level, optimize the report design and reduce the burden at the grass-roots level and the data non-sampling error. Centered on the main line of economic restructuring and development mode transformation, a scientific, unified, complete and applicable statistical standard system that is applicable for the new direction of economic and social development and manifests the new trend and new layout should be established. By integrating the statistical indicator systems, an accessible and forward-looking statistical indicator system that fully reflects the current features of economic and social development and effectively manifests the current situation of respondents should be established.

4.3 Standardized and fair statistical legal system

The modern statistical investigation system must be guaranteed by a standardized and fair legal system. Centered on the approval of statistical investigation items, quality control of statistical data, management of statistical data, sharing of statistical data and issuing of statistical information, etc., relevant laws and regulations should be revised and perfected according to the Statistics Law. For example, a sound information sharing mechanism should be established for the current bottlenecks in sharing of statistical information, and a data sharing laboratory should be established in an exploratory way so as to realize the sharing of data resources conditionally and gradually. It is required to strengthen the building of statistical credit system, establish the basic standards and specifications for the credit evaluation of statistical investigation organizations, and establish the integrity evaluation system for enterprise statistics and the integrity filing system for statistical personnel.

4.4 Convenient and shared data management system

The modern statistical investigation system should rely on a convenient and shared data management system. Establish a hierarchical sharing mechanism for data in an exploratory way. Perfect the data query function, and realize the multi-dimensional data query based on the "chain data" related to statistical object and the "block data" related to the event theme. Perfect the open platform for statistical data, and provide an open service platform for data that has convenient path and is easily operated, visually displayed and interactive with customization summarized for the public.

4.5 Optimized and advanced technical guarantee system

The modern statistical investigation system should be supported by an optimized and advanced technical guarantee system. For this purpose, it is required to innovate the statistical investigation method and use the modern information technologies to reform the statistical production modes; promote the application of modern information technologies such as mobile internet, big data and cloud computing in the statistical work and take such big data as electronic administrative records and various transactions, interactions and sensing as important sources for the basic data of governmental statistics; explore a production mode of "Internet+ statistics", promote the expanded application of mobile intelligent terminal (Personal Digital Assistant) in the statistical data acquisition links such as special investigation and one-shot investigation and promote the electronization of statistical investigation data acquisition; explore the application potential of remote sensing-based information technology in data acquisition and further research and explore the application in such fields as investment, ecological environment quality monitoring, alleviation monitoring of farmer's wholesale market; and promote the application of such technological means as face recognition and administrative record in large general and special investigation activities.

4.6 Innovative, flexible, convenient and friendly statistical service system

The modern statistical investigation system should be based on a flexible and convenient statistical service system. For this purpose, it is required to accurately meet the needs of different service objects and break down the service content; accurately master the development trend of economy and society and provide prompt, effective and insightful statistical analysis products; strengthen and perfect the provision of extensive statistical data and information service for the public, enterprises and public institutions and institutions for academic research; provide the systematic, serialized and personalized statistical products to meet the statistical needs at different levels and improve the quality of statistical service; actively adopt the modern information technologies to spread the statistical data and products, expand the statistical service channels and provide the statistical service for the society easily and quickly; and provide the straightaway statistical products for the public such as cartoon’s promotional video of statistical knowledge and visual statistical product.

References

1. Li Baoqing. Thoughts about Building a Modern Statistical System [J]. China Statistics, 2003 (10): 8 - 10.
2. Lu Junhai, Lv Huan. Discussion on the Characteristics of China’s Modernization and Significance in the World [J]. Economic & Trade Update, 2011 (11):1 - 2+11.
3. Li Qiang. Strengthen the design consciousness of large period statistics and promote the establishment of large period design system [J]. Statistical Research, 2013 (11): 3 - 6.
4. Liu Jianping. Suggestions on deepening the reform of China's governmental statistical investigation system [J]. Statistics & Information Forum, 2016, 31 (11): 7 - 8.
5. Yu Fangdong. New discussion about operation frontier of foreign governmental statistics and international comparison [M]. China Statistics Press, 2012.
6. Huang Yinghui. International experience of statistical system and its enlightenment to China [J]. On Economic Problems, 2009 (5): 53 - 55.
7. Chen Guanghui, Liu Jianping. Research on the construction of modern statistical investigation system in the new era [J]. Statistical Research, 2018 (6): 11 - 17.

Research on Monitoring and Evaluation System for the Long-term Mechanism of Beijing Real Estate Market

Xu Shuming
Beijing Municipal Bureau of Statstics, China

Abstract

Under the background of "Housing is for living in, not for speculation" as proposed by the Central Government, to reflect the development of real estate market and the achievement of control objective, this paper, based on related index data of various departments and research institutions, establishes a monitoring and evaluation index system for the long-term mechanism of real estate market from four aspects such as economy, society, stability and sustainability. The results show that the total index has risen as a whole since the introduction of the property purchase quota policy in 2010, and the total index in 2017 reached 118.8, the highest value over the years. Next, what needs to be done is to improve the basic system settings, build a precise and efficient housing security system and optimize the long-term supply system.

Keywords

real estate; monitoring and evaluation; long-term mechanism

1. Introduction

At the end of 2016, under the background of a sharp rise of housing prices again, the real estate market was specifically re-positioned at the Central Economic Working Conference, and it was also required to accelerate the research and establishment of basic systems and long-term mechanisms that are in line with the national conditions and adapt to the market rules. To reflect the general trend and major characteristic of market development and change more comprehensively and accurately, improve the effect of short-term regulation and control policies, and promote the sound, sustainable and stable development of the real estate market, this paper aims to establish a monitoring and evaluation system for the long-term mechanism of the real estate market from multiple stakeholders such as government, resident, market and enterprise.

2. Methodology

The comprehensive evaluation index of the real estate market uses the comprehensive index method, covers 4 dimensions and 27 indicators. The evaluation system is calculated based on three operations, namely standardization, empowerment and combination to dynamically reflect the development and change of the real estate market from 2006 to 2017.

Table 1 Monitoring and Evaluation System for the Long-term Mechanism of Real Estate Market

First-level objective	Second-level dimension	Third-level indicator	Unit	Data source
Economy	Growth support	Contribution rate of real estate industry to GDP	%	Beijing Municipal Bureau of Statistics
		Amount of investment in construction and installation engineering and growth rate	RMB 100 million, %	Beijing Municipal Bureau of Statistics
	Fiscal and taxation support	Various taxes and fees of real estate industry and growth rate	RMB 100 million, %	Beijing Local Taxation Bureau
		Transaction price of land and growth rate	RMB 100 million, %	Beijing Municipal Commission of Planning, Land and Resources
Society	Government guarantee	Proportion of the number of houses allocated for leasing and sale in the number of households approved after application	%	Beijing Municipal Commission of Housing and Urban-Rural Development
		Proportion of low-income housing investment in real estate development investment	%	Beijing Municipal Bureau of Statistics
		New construction area of low-income housing and growth rate	10,000 m^2, %	Beijing Municipal Bureau of Statistics
	Residents' purchasing power	Housing-price-to-income ratio	%	Shanghai E-house Real Estate Research Institute
		Individual housing loan and growth rate	RMB 100 million, %	People's Bank of China, Beijing Branch
		Per capita living space	m^2	Beijing Municipal Bureau of Statistics

Stability	New housing market	Proportion of the number of houses available for sale in the number of new houses endorsed on the internet	%	Beijing Municipal Commission of Housing and Urban-Rural Development
		Year-on-year price index of new commercial residential buildings	%	Survey Office of the National Bureau of Statistics in Beijing
		Cycle of reducing inventory	Month	Beijing Municipal Bureau of Statistics
		Unmarketable rate	%	Beijing Municipal Bureau of Statistics
	pre-owned housing market	Year-on-year price index of second-hand residential buildings	%	Survey Office of the National Bureau of Statistics in Beijing
		Number of increased customers for pre-owned housing and growth rate	Person, %	Homelink Research Institute
		Number of second-hand houses endorsed on the internet and growth rate	m^2, %	Beijing Municipal Commission of Housing and Urban-Rural Development
	Leasing market	Turnover and growth rate of leasing market	Set, %	Homelink Research Institute
		Rent index	%	Survey Office of the National Bureau of Statistics in Beijing
Sustainability	Industry development	Number of employees in real estate development industry and growth rate	Person, %	Beijing Municipal Bureau of Statistics
		Total assets of real estate development enterprises and growth rate	RMB 100 million, %	Beijing Municipal Bureau of Statistics
		Asset-liability ratio	%	Beijing Municipal Bureau of Statistics
		Business activity index of real estate industry	%	Survey Office of the National Bureau of Statistics in Beijing

		Profits of real estate development enterprises and growth rate	RMB 100 million, %	Beijing Municipal Bureau of Statistics
	Restrictive factor	Growth of permanent population	Person	Beijing Municipal Bureau of Statistics
		Paid-in investment of real estate development enterprises and growth rate	RMB 100 million, %	Beijing Municipal Bureau of Statistics
		Land supply area and growth rate	Hectare, %	Beijing Municipal Commission of Planning, Land and Resources

3. Result

(1) Comprehensive and healthy development of the real estate market

Since 2006, the comprehensive index of monitoring and evaluation over Beijing real estate market has risen in fluctuations. In 2017, the total index was 118.8, 12.4 points higher than that in 2016 and the highest value over the years.

Figure 1 Index Change Chart of Monitoring and Evaluation over Beijing Real Estate Market from 2006 to 2017

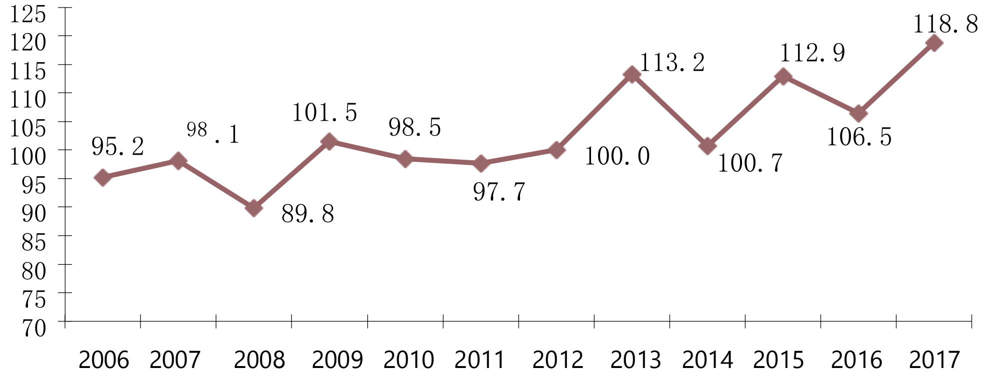


Table 2 Index Changes in Development of Beijing Real Estate Market at Different Levels from 2006 to 2017

Year	Comprehensive index	Sub-index			
		Economy	Society	Stability	Sustainability
2006	95.2	108.3	88.6	93.1	90.7
2007	98.1	108.5	101.2	88.9	93.9
2008	89.8	86.2	97.6	102.0	73.4

2009	101.5	128.8	98.1	80.7	98.3
2010	98.5	85.8	117.6	103.0	87.7
2011	97.7	91.8	113.7	89.7	95.8
2012	100.0	100.0	100.0	100.0	100.0
2013	113.2	156.7	95.1	92.3	108.9
2014	100.7	114.8	86.3	100.8	101.0
2015	112.9	121.6	109.8	112.0	108.4
2016	106.5	89.2	129.2	90.2	117.2
2017	118.8	129.4	114.9	108.7	122.2

(2) Real estate having significant support for the economy and taxation

Affected by the increase in land supply, the support of land transaction area and price for finance and taxation was increased over the previous year. The supply of land for residential and business purposes increased 70.7% over the previous year, the land transaction price increased 88.9%, and the real estate tax increased 34.9%, resulting in the increase of economy sub-index to 129.4 and reversing the decline trend of index in 2016.

(3) Effective increase of residents' housing demand and guarantee level

With the continuous expansion of construction scale of low-income housing, the housing demand of medium and low-income families is constantly satisfied, which has played an important role in stabilizing the real estate market. In 2017, the proportion of investment in low-income housing in real estate development investment increased to 22.7%. The proportion of the number of houses allocated for leasing and sale in the number of households approved after application in the current year was increased from 0.76 in 2009 to 1.03 in 2017. Moreover, the living environment has improved, and the per capita living space of residents was increased from 23.6 m^2 in 2006 to 34.2 m^2 in 2017.

It should be noted that the construction speed of low-income housing has declined. Moreover, with the rise of housing price, the income from housing price was increased from 9.8 in 2006 to 24.2 in 2017, higher than other key cities.

(4) Market stability needing to be strengthened

The regulatory changes of real estate policy results in the fluctuation of stability sub-index in recent years. In 2017, with the issuing of tightened regulatory policy in Beijing City, the market sustained an obvious slow growth, and the stability sub-index recovered to 108.7. Specifically:

The new housing market tends to be stable. In 2017, the stability sub-index of new housing market recovered to 92.1. Since October 2016, the year-on-year price index of new commodity housing continuously dropped, with 99.8 in February 2017. Moreover, the market supply and demand matching was slightly improved, and the number of new houses available for sale exceeded that of new houses endorsed on the internet for the first time. But with the slow growth of the market, the cycle of reducing inventory was extended from 15.5 months in 2016 to 28.7 months in 2017; the unmarketable rate (more than three years' area for sale / area for sale) reached 26.1%, which was the highest value over the years and was not favorable for the continuation of the market stability.

The pre-owned housing market has poor stability. In 2017, the stability sub-index of the pre-owned housing market was 83.5, 28.1 points higher than the previous year. After the "3·17" new policy was promulgated, the slow growth of pre-owned housing market sensitive to the policy was obvious, and the year-on-year price index dropped for 15 consecutive months, resulting in the recovery of the stability sub-index. But the area of pre-owned houses endorsed on the internet dropped 50.8 over the previous year; and the number of increased customers also substantially dropped nearly 30%, thus affecting the further improvement of the pre-owned housing market stability.

The stability of the leasing market gradually increases. In 2017, the stability sub-index of the leasing market was 150.6. Compared with the primary and secondary markets, the price index of the leasing market was relatively stable, and the year-on-year index of rent in recent two years was 103.8. The continuous rise of the housing price has prompted some new populations to choose to rent housing.

The data provided by Homelink Research Institute shows that the transaction volume of rental housing broke through 100,000 in 2015 and then was maintained at around 120,000, and the transaction volume of rental housing increased 7.7% in 2017.

(5) Sustainable development capacity of real estate market improved

In the sub-index, the sustainability index is less volatile and steadily increases. The index was 122.2 in 2017, the highest value in recent years. One of the reasons is that the industry development capability is improved. In recent ten years, the average annual growth rate of the enterprises' total assets has exceeded 10%, and the asset-liability ratio is quite stable. The second reason is that the development sustainability of elements is good. From the elements that affect the industry development, with the advance of relief and rectification work, the new permanent population in Beijing City decreases year by year. The new permanent population decreased by 22,000 for the first time in 2017, relieving the supply and demand contradiction of the real estate market to a certain extent. But the scarcity of urban land resources is gradually prominent, and the supply of new land was gradually decreased to 2,826.5 hectares in 2017 from 6,509 hectares in 2006. Due to the restriction of future land elements, the quantitative expansion needs to be transformed into the improvement of utilization efficiency.

4. Discussion and Conclusion

(1) Perfecting the setting of basic systems

The role of the government as a system maker and market regulator should be fully utilized to promote a sustainable development of the market. First, make a long-term plan for the development of Beijing real estate market according to the population strategy. Second, perfect related systems. Perfect the land system by taking such methods as encouraging the revitalization and reuse of land on hand and improving the efficiency of land use; perfect the financial system by taking such methods as establishing the regulatory agencies for housing finance and supervising and controlling the financial risk in the real estate market; and perfect the fiscal and tax system by taking such methods as maintaining a stable tax framework, simplifying the tax system in the transaction link, reducing the tax burden in the transaction link and increasing the tax in the holding link based on the principle of protecting and encouraging the housing consumption. Third, strengthen the industry supervision and credit construction by taking such methods as establishing cross-department and cross-industry joint punishment mechanism.

(2) Building a precise and efficient housing security system

We should do well the housing security work to ensure that the housing demand of medium and low-income population can be satisfied. First, conduct hierarchical security. When the eligibility for the application for different types of low-income housing is formulated, the division of revenue standard can be properly overlapped to avoid the emerging of sectional phenomenon. Second, expand the security scope. Increase the support for low-income housing according to the needs of medium and low-income population; and at the same time gradually expand the security scope to the housing demand of various kinds of personnel who meet the “four-center” positioning requirements of the capital according to the new development trend of the capital. Third, strictly review and supervise such links as construction, application and use of low-income housing, gradually establish a closed application and exit mechanism, and avoid the slow growth of inventory while great efforts are made in the construction of low-income housing.

(3) Establishing a long-term diversified supply structure

Within the allowable scope of planning, adjust the land supply in a moderately elastic manner according to the situations of real estate market, increase the proportion of residential land and the construction proportion of commodity housing for ordinary home buyers, and adjust the supply structure. Guide the residential layout with industrial layout, increase the public resource supporting services such as transportation, medical treatment and education, promote work-life balance, and reasonably plan the layout. Encourage the leasing service enterprises to integrate the residents' vacant houses and other social idle housing resources by such means as bulk leasing and conduct centralized management; encourage the commodity housing in transit or unmarketable commodity housing to be transformed into the rental housing; encourage the rural collective economic organizations to construct the rental housing by means of the collective construction land and promote diversified supply. Guide the demand gradually released by perfecting the supply system and stabilize the real estate market by stabilizing the market expectations.

References

1. Xu Pei, Wen Yong. Building of Statistical Indicator System for the Real Estate Industry [J]. Statistics and Decision,2014(21).
2. Chen Xiaochuan, Yang Haiyan. Research on Comprehensive Evaluation of Real Estate Development Level in Different Provinces of China [J]. Business Review, 2009 (1).
3. Chen Hongyan, Wang Qiushi. Evaluation Standard for Sound Development of Real Estate Market and Connotation [J]. Jiangxi Social Sciences, 2013 (5).
4. Ji Hanlin, Zhao Qingbin. Research on Sustainable Development Evaluation of Real Estate Industry Based on Principal Component Analysis [J]. Resource Development & Market, 2013 (29).

Analysis of the impact of demographic shift on housing demand in Beijing

Zhang Lin
Beijing Municipal Bureau of Statistics, Beijing, China

Abstract

In the medium to long term, the development cycle of housing market is closely linked with regional demographic shift, with changes in population size and structure directly affecting development trend of the housing market. As demographic shift is highly predictable, this paper examines the impact of changes in population size and structure on housing demand, preferences and spatial distribution of properties in Beijing, which can help local government predict its housing market's development trend, make decisions and adjust related policies. This paper concludes that the population scale and structure in Beijing can still sustain its housing demand but slowing population growth and ongoing demographic shift will have an impact on housing demand and preferences in the future. In the future, the housing planning in Beijing should look ahead and take into account population policies and optimize the supply of various types of houses.

Keywords

Real estate market; Population size; Age structure; Housing preference

1. Introduction

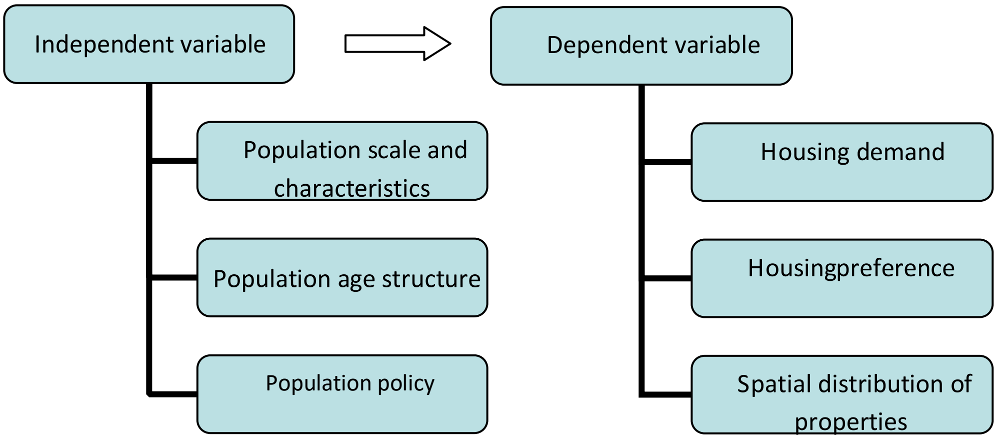
The housing market is influenced by economic situation, monetary policy and other related policies. And it is also closely linked with regional demographic shift. The changes in population size and structure affect the current and future development trend of regional housing market. China further deepened the reform of urban housing system in 1998 when the welfare-oriented public housing distribution system was cancelled, and most of the Chinese city dwellers started to buy houses in the housing market, which was very different from the past, Judging from the developmental characteristics of and changes in population and the housing market in past twenty years, population has become an important factor affecting the development trend of the housing market.

As demographic shift is highly predictable, predicting housing demand based on the changes in population can help predetermine the development trend of the housing market, which can help local government make decisions and adjust related policies.

2. Methodology

Using analytical methods such as qualitative inference and quantitative inference and using the population size of Beijing as the independent variable, this paper investigates the impact of population size and structure of Beijing on housing demand, preference and spatial distribution in the capital city of China.

Figure 1 Research Framework



3. Analysis of the Impact of Population Size and Structure on Housing Demand

3.1 Main Factors Driving Housing Demand

3.1.1 Ratio between the housing stock and the number of households (RHH) in Beijing The ratio[1] between the housing stock and the number of households in Beijing is used to measure the balance between supply and demand. According to the data of the fifth and sixth censuses of Beijing, this ratio of urban residential housing in Beijing was respectively 0.73 and 0.79 in 2000 and in 2010. Seen from the international experience, one of the essential conditions for ensuring the sound and steady development of housing market is to ensure that the housing stock is in line with the population size and the number of households (or slightly exceeding it). For example, there were about 3,033,000 households 3,328,000 houses in New York in 2008, and higher housing stock in the city guaranteed a better living condition for its residents. Currently, the RHH is relatively low in Beijing, so there is still room for increasing the housing stock.

[1] Ratio between the total houses and the total family households. If this ratio reaches 1, it shows that the problem concerning the absolute short supply has been alleviated.

3.1.2 Household scale

As for the household scale, the proportion of one-person households and two-person households in Beijing was 51.9% in 2015, which was 12.3 percentage points higher than in 2004(39.6%). Driven by the household miniaturization and urbanization, the number of urban households increased by 51.1% in Beijing from 2000 to 2010, which in turn resulted in rigid demand in the housing market.

3.1.3 Population without "Hukou"[2]

From the structure of housing purchasers, before the purchase quota [3] policy was implemented in Beijing, the proportion of the population without Hukou (non-local households) purchasing houses in the city accounted for 1/3, higher than that of the migrant population living here. After the purchase quota policy was implemented in 2011, the proportion of the population without Hukou purchasing houses in Beijing once fell below 10%; the proportion of purchasing the houses was 20% lower than that of the migrant population from 2011 to 2015. The purchase quota policy inhibited the purchase of housing by the external population. However, as more and more external population meet the purchase conditions over time, the potential demand of the migrant population purchasing the houses will likely turn into the actual demand.

3.1.4 Population dependency ratio

The population dependency ratio in Beijing was 28.2% in 2000 and 25.6% in 2015. The child dependency ratio decreased by 4.7%, which is the main factor for driving the decrease of the total dependency ratio. It indicates that the reduction of the dependency burden (mainly the reduction of the child dependency burden) is accompanied by the stronger household savings and purchasing power and the stronger willingness and ability to maintain and increase the value through housing purchase, which is one of the reasons for effectively driving the demand of the housing market.

[2] Hukou is a system of household registration in China. A household registration record officially identifies a person as a resident of an area and includes identifying information such as name, parents, spouse, and date of birth.

[3] In 2011, the purchase quota policy was formally implemented in Beijing that the local households that own two or more houses, the non-local households that own more than one houses as well as the non-local households that cannot provide the temporary residence permit for Beijing and have paid the social insurance or personal income tax for more than five consecutive years were suspended to sell the houses to them in Beijing.

3.2 Analysis of Restrictive Factors for Demand

3.2.1 Population growth

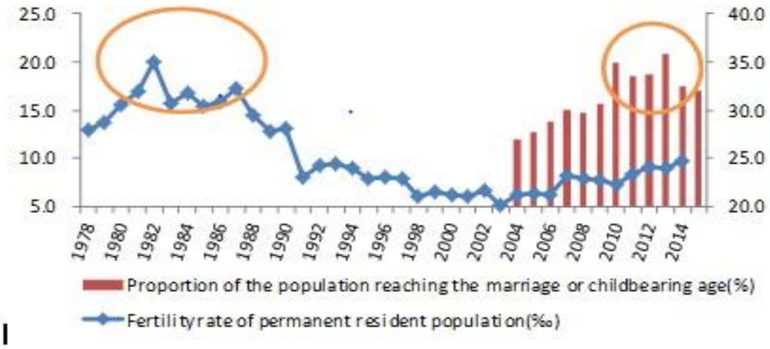
Since 2011, the increment and growth rate of the permanent residents in Beijing City have gradually declined, and especially the growth rate of the permanent external population has significantly dropped from 14.7% in 2010 to 0.5% at present. In recent two years, as great efforts are made in population decentralization, the total population has begun to decrease.

3.2.2 Future marriage and childbearing age population (20-34 years old)

During the period of baby boom from 1980 to 1987, the fertility rate in Beijing exceeded 15‰ and entered the downlink channel after 1987. The fertility rate has been below 10‰ since 1991 and has fallen to a minimum of 5.06 in 2003 but has begun to rise since then. It is reckoned by considering the first marriage age as around 27 years old that the baby boom in the 1980s made the population reaching the marriage or childbearing age in Beijing reach the peak value from 2007 to 2014 (see Figure 2), but the population reaching the marriage or childbearing age will substantially drop in around 2018 and decrease to the low value in 2030. This is basically consistent with the situation that the proportion of the population reaching the marriage or childbearing age in Beijing began to drop after reaching the highest value (35.8%) in 2013. It can be predicted that this drop trend will continue for a long period of time (from 2018 to 2030), and the decrease of the population reaching the marriage or childbearing age will drive the demand decrease of the real estate market.

Figure 2 Fertility Rate of Permanent Resident Population and Proportion of the Population Reaching the Marriage or Childbearing Age in Beijing

Unit: ‰, %



3.3 Analysis on the Impact on Housing Preference due to the Population Age Structure

The populations of different ages have different housing demands and purchasing abilities. In Beijing, the population at the age of 21 to 35 years old is the group that has the strongest housing demand, and the proportion of this group purchasing the houses reached 51.2% in 2015, mainly centred on first-time purchase and new houses; the population at the age of 36 to 50 years old is the second largest group of house purchase, accounting for 29.9%, they have the demand for better housing and buy the largest house on average (the average area of the new houses purchased is 135 m^2); the population at the age of 51 to 60 years old accounts for 8.5% in house purchase, which mostly requires the improved type and endowment type housing and has an average purchase area of new houses slightly smaller than that of the population at the age of 36 to 50 years old; and the population at the age of more than 60 years old accounts for 4.7% in house purchase, which mostly requires the endowment type housing and has an average area reduced to at least about 120 m^2.

It is estimated by taking the age time-translation method that first the proportion of the population reaching the marriage or childbearing age (20 to 34 years old) will decrease, and the demand for small houses will also decrease in the next ten years. Second, the demand for improved type housing will become the market subject with the increase of the population at the age of 35 to 64 years old. Seen from the changes in unit design in recent five years, the preference for improved type housing has been shown. It is shown in the investigation over the household composition of urban residents in 2014 that the one-bedroom and two-bedroom households accounted for 60.1%, 6.5% lower than that in 2010; and the three-bedroom and four-bedroom households accounted for 28%, 1.3% higher than that in 2010. In addition, with the relaxation of the two-child policy, the demand for improved type multi-bedroom houses will be further increased in the future.

3.4 Analysis on the Impact on Spatial Distribution of Real Estate due to the Population Decentralization Policy

According to the urban function positioning of the capital and the coordinated development requirement of Beijing, Tianjin and Hebei, in 2020 the permanent resident population in six districts of Beijing City will decrease by 15% on the basis of 2014. in 2015, the growth rate of the permanent resident population and the migrant population respectively dropped below 1%; the proportion of construction area in six districts decreased by 15.5% than 2010, and the proportion of the construction area in the new development areas such as Tongzhou and Daxing increased by 12.4%. The population decentralization, industrial function transfer and new town planning and construction, etc. drive the spatial distribution of the real estate to be spread from the centre to the outside.

5. Research Conclusion and Discussion

5.1 Research conclusion

This paper finds that in the short term the housing supply in Beijing still fails the demand. There are several main factors driving housing demand in Beijing. First, the ratio between housing stock and the number of households is relatively low in Beijing, and more houses are needed. Second, the household miniaturization resulted in the rigid housing demand. Third, the population without Hukou poses potential demand for housing. Fourth, the low population dependency ratio boosts the effective demand for housing.

But in the medium and long term, with population structure shifting and population policy having effect, the rigid demand for housing coming from a huge population will gradually decrease. It is estimated by taking the age-moving method that impact of the factors restricting the housing demand will appear and increase in the influence in the next ten years. First, the slowdown of the population growth will limit the expansion of demand. Second, the decrease in the proportion of the population reaching the marriage or childbearing age (20 to 34 years old) will affect the housing demand in the future. The change in the age structure of the population will have a greater impact on the demand structure and housing preferences. With the decrease in the proportion of the population reaching the marriage or childbearing age (20 to 34 years old), the demand for small houses will decrease to a certain extent. With the increase of the population between 35 to 64 years old, the demand for better housing will become the main demand. In addition, with the relaxation of the one-child policy, the demand for better housing, such as multi-bedroom houses, will further increase in the future. The changes in city positioning and population policy will also directly drive the spatial distribution of housing to spread from city centre to its outskirts.

5.2 Discussion and suggestion

Due to the predictability of the demographic shift, in the policy making processes, the Beijing government should take into consideration housing supply and demographic shift as a whole. The government should plan and adjust the number and spatial distribution of newly built houses according to such factors as fluctuation of population, population policy, and two-child policy effect. This paper suggests that the government should adjust the supply structure of newly built houses in a timely manner. Based on the house types and the changing population structure, the government should gradually adjust the house types in the housing market, appropriately increase the supply of large-sized and multi-bedroom houses, and gradually meet the demand for better housing. Second, the government should develop the welfare-orientated housing for older people, and plan and build the house types, communities and nursing organizations suitable for older people to live. Third, the government should closely combine the spatial distribution and planning of the urban housing with the population policy.

References

1. Beijing Municipal Bureau of Statistics, Survey Office of the National Bureau of Statistics in Beijing. Beijing Statistical Yearbook (1999 - 2016). China Statistics Press, 1999 - 2016
2. Beijing Municipal Commission of Housing and Urban-Rural Development. Beijing Real Estate Yearbook (2016). China Development Press, 2016
3. Yang Hualei, Wen Xingchun, He Lingyun. Baby Boom, Population Structure and Housing Market. Population Research, 2015: 5: 87 - 99
4. Dai Guohai. Influence of Demographic shift on Cyclical Fluctuation of Real Estate Market. Jinan Finance, Issue 7: 25 - 30

Absolute partial mean curve: An alternative to receiver operating characteristic curve

Pooja Bansal, Pramod K. Gupta
Department of Biostatistics, Post-Graduate Institute of Medical Education and Research, Chandigarh 160012 India

Abstract

Rational: A medical diagnostic test in medical science/research serves a critical purpose in diagnosing a medical condition of a person with or without a disease, correctly. Receiver operating characteristic (ROC) curve is one of the most extensively used statistical methods in medical science research for evaluation of a diagnostic test performance. The practical difficulty/drawback in constructing ROC curve is selection of different cut-off points of a medical diagnostic test measure. The practice for selection of different cut-off points is a random choice between minimum and maximum value of a medical diagnostic test measure thus precision of critical values can be compromised. **Aim:** We aim here to develop a procedure similar to equal variance normal ROC curve based on Absolute Partial Mean (*APM*) curve. The *APM* curve construction does not depend on random choice of cut-off rather the critical values of *APM* curve is derived as a function of grand mean and subsequent partial mean derived from a medical diagnostic test measure. **Result:** Thereby, implying that the cutoff point obtained from Youden index for equal variance ROC curve is equivalent to the point where the rotated lognormal *APM* curve (APM_R) attains its maximum. The calculation of (APM_R) curve is much simpler in comparison to ROC curve and thereby can be used as effective alternative to ROC curve for diagnostic test especially.

Keywords

Medical Diagnostic Test; ROC curve; AUC; APM curve

1. Introduction

A medical diagnostic test in medical science/research serves a critical purpose in diagnosing medical condition of a person with or without diseases, correctly. Receiver operating characteristic (ROC) curve is one of the most extensively used statistical methods in medical science research for evaluation of diagnostic test performance, Aphinyanaphongs et al. (2004). ROC graph, in practice is constructed using data driven approach based on sensitivity and specificity, (Shapiro (1999). Sensitivity and specificity of medical diagnostic test are two independent indicators that tell whether a person is with and without disease, respectively. However, the computation of sensitivity and specificity depends upon critical (optimal cut off) values of a medical diagnostic test measure. The ROC curve is thus an approach to find equilibrium between sensitivity and specificity of a medical diagnostic test. The ROC curve is a graphical representation and the quantitative performance of a medical diagnostic test is obtained in terms of area under the curve (AUC), Hanley et al. (1982) at optimum cut-off point, which is also known as Youden J-Statistics or Youden Index, Youden (1950). The value of AUC lies between 0 and 1 and higher values of AUC are translated into better medical diagnostic test. The AUC thus is also used in statistical tests to compare the performance of two medical diagnostic tests.

The practical difficulty/drawback in constructing ROC curve is selection of different cut-off points of a medical diagnostic test measure. The practice for selection of different cut-off points is random choice between minimum and maximum value of a medical diagnostic test measure thus precision of critical values can be compromised.

We would like to bring in here the theoretical background of ROC curve that depends on bivariate normal distribution. The most popularly used ROC curve is meant for discriminating populations that follow normal distributions and is also termed as bi-normal ROC curve, Hanley and McNeil (1996), Metz (1976). However, it can be noted through the literature that there are at least few other curves that are being used other than medical field for evaluation of diagnostic test performance. However, the theory for finding the critical value is based on theoretical premises rather on random choice. Ordinal dominance curve in the field of economics is noted to be related to the Normal model of ROC, (Bamber (1975). The Lorenz curve used for assessing relative income inequality, Lorenz (1905) and normal ROC model are closely related. Irwin and Hautus (2015) has also mentioned that equal variance normal ROC curve (the fundamental model of diagnostic systems) and the lognormal Lorenz curves are mirror images of each other and hence, the lognormal Lorenz can be used to evaluate the diagnostic systems.

We aim here to develop a procedure similar to equal variance normal ROC curve based on Absolute Partial Mean (*APM*) curve, Arora et al (2011). The *APM* curve construction does not depend on random choice of cut-off rather the critical values of *APM* curve is derived as a function of grand mean and subsequent partial mean derived from a medical diagnostic test measure. This is an ongoing work and thus we are discussing here the special case of *APM* curve for lognormal distribution and its comparison with ROC curve under similar set-up.

2. Methodology

The *APM* curve: Let $X_1, X_2, \dots, X_n$ be a medical diagnostic measure of n-individuals, which is ordered $X_{(1)} < X_2 < \cdots < X_{(n)}$. Henceforth, *APM* curve for given observations is defined in equation (1) according to Arora et al (2011).

$$APM\left(F(x)\right) = \mu_x - \mu \qquad (1)$$

$F(x)$ represents the distribution function, μ is grand mean and

$\mu_x = \frac{\int_o^x tf(t)dt}{\int_o^x f(t)dt}$ is the partial mean.

Development of New *APM* curve for Lognormal Distribution: Let us assume $X_1, X_2, \dots, X_n$ be a medical diagnostic measure of $n-$individuals which follows log-normal distribution with μ and σ^2. Hence, the APM curve for lognormal distribution is derived in equation (2).

$$APM\left(F(x)\right) = \frac{\mu(\varphi(\varphi^{-1}(F(x)-\sigma)-F(x))}{f(x)} \qquad (2)$$

$\varphi(.\,)$ in the equation (2) is the distribution function of standard normal distribution. Further, the equation for *APM* curve for a quantile distribution function, $F^{-1}(p) = \inf\{x\colon F(x) \geq p\}$, given in (2) can be redefined as:

$$APM\ (p) = \frac{\mu*(\varphi\left(\varphi^{-1}(p)-\sigma\right)-p)}{p} = \frac{\mu*\mu^{-1}(p)-\sigma)}{p} - \mu, 0 \leq p \leq 1 \qquad (3)$$

It is pertinent to mention that the ordinates of the *APM* curve given in (1) are negative as depicted in Fig-1A. Therefore, it is required to rotate the lognormal *APM* curve through 180^0 around x-axis to make it comparable to ROC curve. The general equations of rotating a point (x, y) through 180^0 around x-axis are $x^{'} = x$ and $y^{'} = -y$. We used equation (3) to obtain the desired new curve APM_{R} given in equation (4).

$$APM_R(p) = -APM\ (p) = \mu - \mu_x \qquad (4)$$

The rotated new *APM* curve ($APM_{\$}$) is shown in Fig-1B.

Development of New Data Driven *APM* curve: Let $X_1, X_2, \dots, X_n$ be a set of values of n individuals following lognormal distribution with mean μ and variance σ^2. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ *be* ordered values from lowest to highest and $\bar{X}$ be the corresponding sample mean. Then the sample estimate of lognormal APM curve is given as:

$$APM_R(p) = \hat{\mu} - \hat{\mu}_x = \bar{X} - \bar{X}_p$$

Where $\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ (overall sample mean) and $\hat{\mu}_x = \sum_{i=1}^{k} x_i$, i.e. mean of first k observations and $p = \frac{k}{n}$ $with\ k = (1, 2, \dots, n) and\ 0 \leq p \leq 1p$

The ROC Curve: Let us assume $X_1, X_2, \dots, X_n$ be a medical diagnostic measure of n –individuals follows Normal distribution with μ and σ^2. The sensitivity and the specificity of the test for the given cut off, say "c" are given in (6) and (7), respectively for given "*d*", which is the mean of the diseased distribution specified in standard units.

$$se = 1 - \varphi(c - d) \qquad (6)$$

$$sp = \varphi(c) \qquad (7)$$

Henceforth, the ROC curve for equal variance normal distribution in accordance to Irwin and Hautus (2015) is defined in equation (8).

$$R(p') = \varphi(\varphi^{-1}(p') + (p') + d) \qquad (8)$$

where $p' = 1 = sp = 1 - \varphi(c), 0 \leq p' \leq 1.$

Development of Relationship Between Two Curves: Let us say that "c" and "z" are the cut-off point of said two curves respectively for ROC curve (R) and APM_R. The two said curves to have common cut-off point it is necessary to hold condition given in equation (9).

$$c = z \Rightarrow \varphi(c) = \varphi(z) \Rightarrow ,1 - \varphi(c) = 1 - \varphi(z) \Rightarrow p' = 1 - p \qquad (9)$$

Where $z = \frac{\ln x - \mu}{\sigma}$ and $\varphi(.)$ is the *CDF* of standard normal distribution. The relationship between two thus is represented by equation (10).

$$R(p') = \left(\frac{1-p'}{\mu}\right) * APM_R(p') + p' \qquad (10)$$

The relationship derived in equation (10) satisfies the following conditions.

(i) $APM_R(0) = 0, APM_R(1) = 0\ as\ R(0) = 0\ and\ R(1) = 1$

(ii) $0 \leq APM_R(p) \leq \mu$

(iii) $APM_R'(p) \geq 0\ and\ APM_R''(p)\ \leq 0$

Development of Relationship Between AUC of Two Curves: The AUC for equal variance normal ROC curve given by Hanley and McNeil (1982) is shown in equation (11).

$$AUC_R = \int_0^1 R(p')dp' = \int_0^1 1 - \varphi(c-d)dp' = \varphi\left(\frac{d}{\sqrt{2}}\right) = \varphi\left(\frac{\sigma}{\sqrt{2}}\right) \tag{11}$$

The AUC for APM_R curve and its line of equality is derived by us and given in equation (12) and thus relationship between them are derived using (10) and given in equation (13).

$$AUC_{APM} = \int_0^1 APM_R(p')\,dp' = \int_0^1 \mu * \left(\frac{R(p')-p'}{1-p'}\right) dp' \tag{12}$$

$$AUC_{APM} = \mu * (AUC_R - 1/2) \tag{13}$$

Development of Cut-Off of Two Curves: The cut-off point for ROC curve is obtained through Youden index, Youden (1950), given in equation (14).

$$J_R = \max\{se(c) + sp(c) - 1\} = \max\{R(p') - p'\} \tag{14}$$

Similarly, the cut-off point for APM_R curve and its line of equality i.e. $p' = 0$ is derived by us and given in equation (15).

$$J_{APM} = max\{APM_R(p') - (p' = 0)\} = max\{APM_R(p')\} = \; = max\left[\mu * \left(\frac{R(p')-p'}{(1-p')}\right)\right] \tag{15}$$

The relationship between cut-off points is thus derived with help of equations (14) and (15)

$$J_R\ (Youden\ Index) = \frac{J_{APM}}{\mu}$$

3. Results

A set of 50-observations were simulated from log-Normal distribution with mean = 4 and standard deviation = 2. We did preliminary analysis for theoretical findings developed in previous section. Fig1 to Fig-3 and Table-1 shows that results from both approaches are same.

Indices	J_{APM}	J_R	Relationship
Youden	1.14	0.27	$J_R(Youden\ Index) = \frac{J_{APM}}{\mu}$
Cut-off	1.17	4.99	$\mu * J_R = 4*0.27 = 1.1$

4. Discussion and Conclusion

The above result forms an important basis for the calculation of cut-off point which is the main aspect of using ROC curve. Thereby, implying that the cut-off point obtained from Youden index for equal variance ROC curve is equivalent to the point where the rotated lognormal APM curve attains its maximum. The calculation of APM curve is much simpler in comparison to ROC curve and thereby can be used as effective alternative to ROC curve for diagnostic test especially in the case when data is of quantitative nature. The findings developed herein is the intermediate as the work is ongoing.

References

1. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis C. F. (2004). Text categorization models for high quality articleretrieval in internal medicine. J Am Med Inform Assoc.
2. Arora S., Mahajan K. K, Bansal P. (2011). Absolute Partial Mean curve. Advances and Applications in Statistics 24(2), 139-156.
3. Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Mathematical Psychology 12, 387–415.
4. Hanley J. A., McNeil B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36.
5. Hanley J. A. (1996). The use of the “binormal” model for parametric ROC analysis of quantitative diagnostic tests. Stat. Med. 15:1575–1585.
6. Irwin, R. J., & Hautus, M. J. (2015). Lognormal Lorenz and normal receiver operating characteristic curves as mirror images. Royal Society open science, 2(2), 140280.
7. Lorenz M. O. (1905). Methods for measuring concentration of wealth. J. Am. Stat. Assoc. 209-19.
8. Metz C. E. (1978). Basic principles of ROC analysis. Semin. Nucl. Med. 8:283–298.
9. Shapiro D. E. (1999). The interpretation of diagnostic tests. Stat Methods Med Res 8(2):113–34.
10. Youden, W.J. (1950). Index for rating diagnostic tests. Cancer, 3, 32–35.

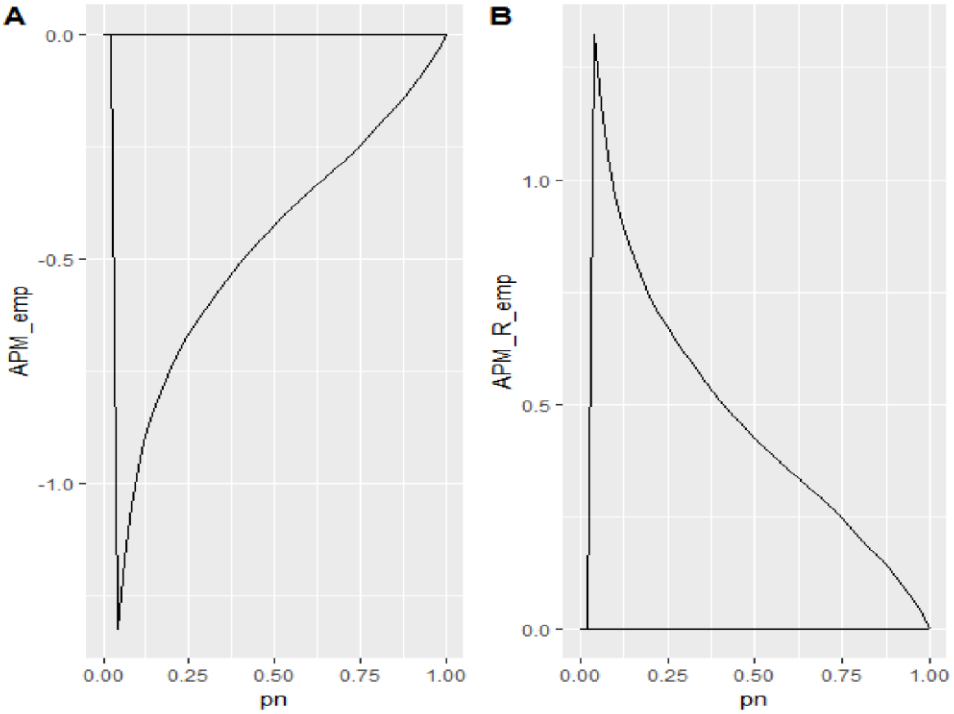


Fig-1: Data Driven *APM* Curve and its 180 Degree Rotation View (B)

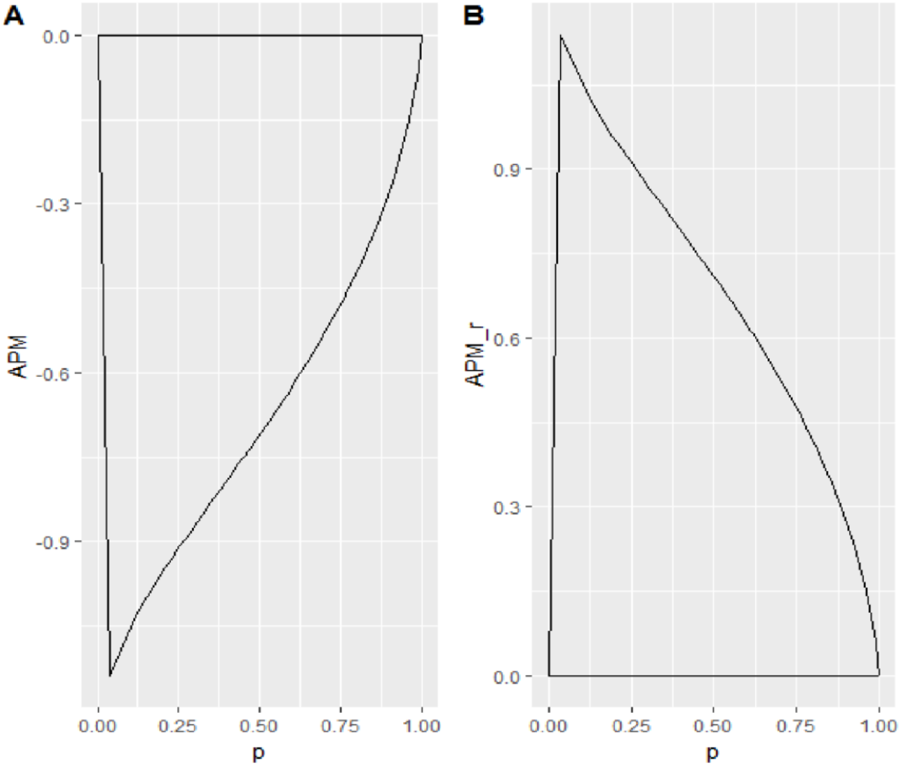


Fig-2: Log-Normal *APM* Curve and its 180 Degree Rotation View (B)

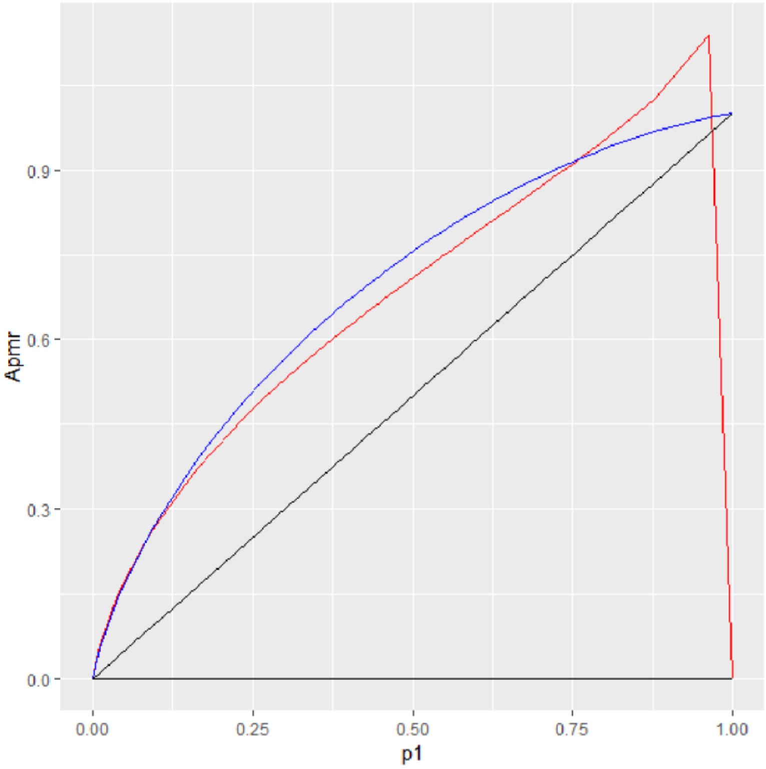


Fig-3: ROC and *APM* Curve for Same Data Set

Ranking Based Variable Selection for Censored Data Using AFT Models

Md Hasinur Rahaman Khan, Marzan Akhter
University of Dhaka, Dhaka-1000, Bangladesh

Abstract

Many variable selection techniques are developed for complete data but when censoring is present, this techniques must be modified because of the complex structure of the likelihood function. In this paper, we consider variable selection technique for accelerated failure time models by extending the ranking based variable selection (RBVS) algorithm and its iterative procedure IRBVS as proposed in Baranowski and Fryzlewicz (2017). The method is developed for the the right-censored data through estimating the Stute's weighted least squares technique. Extensive simulation studies are conducted to demonstrate the performance of the methods. We further illustrate the performance of the method using a real microarray data analysis. The overall performances of the proposed methods, in comparison with particularly the iterative sure independence screening and the stability selection methods, are found as very impressive particularly for low dimensional data but reasonably very good for high-dimensional data when there was no correlation among the covariates. The real data analysis also suggests that our methods are enable to select genes that are significantly associated with the survival of the patients.

Keywords

Censoring; High-dimensional data; Variable selection; Variable ranking; Subset selection; Stability selection

1. Introduction

In modern statistical applications, variable selection for high dimensional data has been the focus of many researches. In case of high dimensional data, there are many variables that have no impact on the response variable. When the number of covariates (say p) is larger than the sample size n, the traditional methods (say, the least square estimation) become difficult and in turn the precise statistical inference is not possible. This leads to growing interest for identifying a subset of covariates that affects the response variable significantly. There are many algorithms and methods that have been developed for variable selection for complete data. For example, Fan and Lv (2008) ([5]) consider marginal correlation ranking for sure independence screening (SIS) in linear model. Cho and Fryzlewicz (2012) ([4]) suggest a variable screening procedure which is known as tilting procedure where an adaptive choice is performed between the use of marginal correlation and tilted correlation for each variable. The ranking procedures are constructed by ranking of covariates according to the measure of association between the covariates and the response. Several ranking based techniques are developed for variable selection because of simplicity and computational gains. A ranking procedure that depends on the marginal quantile utility is suggested by He et al. (2013) ([7]). Li et al. (2012) ([9]) suggest another method for ranking the covariates based on their distance correlation to the response.

As a variable ranking procedure, Hall and Miller (2009)[6] suggest a method which has focussed on position of each variable in the ranking. The bootstrap confidence intervals are computed to find out these positions. The covariates for which the right end of the confidence interval is lower than the chosen cutoff point are selected. This cut off point is used to be$\frac{p}{2}$. Meinshausen and Buhlmann (2010)[10] propose another approach called Stability Selection (StabSel) which is based on sub-sampling. It depends on the choice of cutoff point. It is important to concentrate on the appropriate choice of the threshold for better result. Recently, Baranowski and Fryzlewicz (2017)[1] develope a variable selection procedure for high dimensional data based on ranking, which is known as the Ranking Based Variable Selection (RBVS). It is a sub-sampling procedure and the ranking of covariates are performed according to their impact on the response. The main concept of the procedure is that there are some subsets which contain irrelevant covariates in high-dimensional data and those may appear to have very high impact on the response variable. But they are not consistently related to the response variable over the subsamples where the truly important variables will be related to the response variable consistently, both over the entire sample and the sub-samples. The main goal of this study is to extend the ranking based variable selection variable selection algorithm to censored data using the AFT models that are estimated with the Stute's least square techniques ([11]).

2. Methodology

Let us consider, $T_i(i = 1,...,n,)$ be the logarithm of the failure time and $X_i(i = 1,...,n,)$ is a covariate vector for the ith subject in a random sample of size n with length p. The AFT model can then be defined as

$$T_i = \beta_0 + \boldsymbol{X}_i{}'\boldsymbol{\beta} + \epsilon_i; \ \ i = 1, \ldots, n, \tag{1}$$

where β_0 is the intercept, $\beta \in R^p$ is the unknown $p \times 1$ vector of regression coefficients and ${}_i$'s are independent and identically distributed random variables whose common distribution may take a parametric form, or may be unspecified, with zero mean and bounded variance. In the AFT model, T_i is subject to right censoring, C_i be the logarithm of the censoring time. So, we only observe (Y_i, δ_i, X_i) where $Y_i = \min\{T_i, C_i\}$ and the censoring indicator is denoted by $\delta_i = I\{T_i \leq C_i\}$. The AFT models cannot be solved using ordinary least squares method because it is difficult to handle censored data by this model. To overcome this problem, the Stute's weighted least squares method (Stute, 1996) is considered where weights are needed to account for censoring in the least square criterion. The weighted least square for AFT model is used in many studies [8]). Let, $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the ordered logarithm of survival times. The weighted least square estimator defined by Stute (1996) can be writen as

$$l(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\left(Y_{\pi(i)} - \boldsymbol{X}'_{\pi(i)}\boldsymbol{\beta}\right)^{2} . \qquad (2)$$

But for simplicity we still denote the weighted data as (Y_i, X_i). The objective function in equation (2) can easily be estimated under any techniques. We consider the lasso technique ([13]) that minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients is less than a constant. The lasso estimator, based on the estimating equation as defined in equation (2), is given by

$$\hat{\beta}_{pen} = \text{argmin}^{\beta}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \sum_{j=1}^{p} pen(|\beta_j|)\right),$$

where $pen(M) = \lambda M$ and λ is the tuning parameter. The MC+ estimator [14]. based on the estimating equation as defined in equation (2) can be expressed by,

$$\hat{\beta}_{pen} = \text{argmin}^{\beta}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \sum_{j=1}^{p} pen(|\beta_j|)\right)\_ \quad ,$$

where $pen(M) = \lambda\int_0^M \max\{0, (1 - x/(\gamma\lambda))\}dx$. Here $\lambda > 0$ and $\gamma > 0$ are tuning parameters.

a. The Ranking-Based Variable Selection (RBVS)

Under the RBVS, at first we pick some data randomly to assess the impact of each variable on response Y repeatedly. In next stage, we sort the covariates in decreasing order for each random draw according to their level of impact on Y. From the ranking of variables, we have to identify the sets of covariates appearing at the top of the rankings frequently and record the corresponding frequencies. Then this process leads to finding the variables that has the highest probability of selection in the final model. According to Baranowski and Fryzlewicz (2017), suppose we observe $Z_i = (Y_i, X_{i1}, \ldots, X_{ip})$; $i = 1, \ldots, n$, as the n independent copies of a random vector $Z = (Y,X_1,...,X_p)$ which including how this can be obtained for censored data is explained in previous section. For variable selection, our interest is to identify the subset of $\{X_1,...,X_p\}$ which truly affects Y. Consider $\hat{w}_j = \hat{w}_j(Z_1,...,Z_n), j = 1,...,p$ as a measure that gives the estimates under different techniques including penalization. Based on the measures $\hat{w}_1,...,\hat{w}_p$, the variable ranking $R_n = (R_{n1},...,R_{np})$ is a permutation of $\{1,...,p\}$ which satisfies $\hat{w}_{R_{n1}} \geq \cdots \geq \hat{w}_{R_{np}}$. If the ties are present, then they are taken at random basis.

Let us define, $\Omega_k = \{A \subset \{1,...,p\} : |A| = k\}$ for any $k = 0,...,p$, where $|A|$ is the number of elements in set A. For every $A \in \Omega_k, k = 1,...,p$, the probability of its being ranked at the top can be expressed as

$$\pi_n(A) = P(\{R_{n1}(Z_1,...,Z_n),...,R_{nk}(Z_1,...,Z_n)\} = A).$$

We can set $\pi_n(A) = \pi_n(\varphi) = 1$ for $k = 0$. Now, for any integer m which satisfy $1 \leq m \leq n$, it can be written as

$$\pi_{m,n}(A) = P(\{R_{n1}(Z_1,...,Z_m),...,R_{nk}(Z_1,...,Z_m)\} = A).$$

By the use of a variant of the m-out-of-n bootstrap ([2]), the estimators of $\pi_{m,n}(A)$ are obtained to compute an estimate of S. For n sufficiently large, S will be one of the following sets:

$$A_{k,m} = \text{argmax}_{A\in\Omega_k}\pi_{m,n}(A), k = 0,...,p.$$

So, it can be written as

$$\hat{A}_{k,m} = \text{argmax}_{A\in\Omega_k}\hat{\pi}_{m,n}(A).$$

In practice, the number of elements in S is typically unknown. So, it should be estimated. From Baranowski and Fryzlewicz (2017), s is estimated by

$$\hat{s} = \text{argmin}_{\text{k}=0,\dots,\text{p}-1} \frac{\hat{\pi}_{m,n}(\hat{A}_{k+1,m})}{\hat{\pi}_{m,n}(\hat{A}_{k,m})} .$$

The important advantage of the estimator is that it does'nt require any parameters. The algorithm of RBVS consists of four steps.

i. The RBVS Algorithm

- Step 1: *Subsampling:* For each $b = 1,\cdots,B$, draw uniformly without replacement m-element subsets $I_{b1}, I_{b2}, \cdots, I_{br} \subset \{1,\cdots,n\}$ for $r = \left[\frac{n}{m}\right]$ times.

- Step 2: *Ranking evaluation:* Evaluate $\hat{w}_j(Z_i \in I_{bl})$, $j = 1,\dots,p$ for each $b = 1,\cdots,B$ and $l = 1,\dots,r$ and sort $\hat{w}_j(Z_i \in I_{bl})$ in non-increasing order. Then record corresponding variable ranking $R_n(Z_i \in I_{bl})$.

- Step 3: *Estimation of k-top-ranked sets:* For each $k = 1,\dots,k_{max}$, find the k-element set $\hat{A}_{k,m}$ which is the most frequently occurring in the top of the rankings $R_n(Z_i \in I_{bl})$, $b = 1,\dots,B$, $l = 1,\dots,r$. Then write down the probabilities ˆ $\pi_{m,n}(\hat{A}_{k,m}) = \frac{1}{Br}\sum_{b=1}^{B}\sum_{j=1}^{r} I\left(\hat{A}_{k,m} = \boldsymbol{R}_{n,1:k}(\boldsymbol{Z}_i, i \in I_{bj})\right)$

- Step 4: *Estimation of the top-ranked set:* Find $\hat{s} = argmin_{k=0,\dots,k_{max}-1}\frac{\hat{\pi}_{m,n}(\hat{A}_{k+1,m})}{\hat{\pi}_{m,n}(\hat{A}_{k,m})}$ as estimate of the size of the top-ranked set. Return $\hat{S} = \hat{A}_{\hat{s},m}$ which is the final estimate of S.

Here, we consider $\hat{\pi}_{m,n}(\hat{A}_{k,m})$ for $k \leq k_{max}$. When the RBVS is used for the absolute values of the regression coefficients estimated via the lasso and MC+ variable selection methods we then term them as RBVS-lasso and RBVS-MC+ respectively. An important part for RBVS method is to choice three parameters–m,B,k_{max}. In our simulation studies$m = [\frac{n}{2}]$ is considered as in Baranowski and Fryzlewicz (2017) ([1]). We consider $B = 500$ for our simulation studies. We take the value of k_{max} equals to p for our study. We use the 10-fold cross-validation for tuning parameter λ for RBVS and Stability Selection methods. We also set $\gamma = 3$ for MC+ according to Breheny and Huang (2011) (citebreheny2011coordinate).

b. IRBVS: An Iterative Extension of RBVS

We used the iterative procedure of RBVS algorithm called IRBVS as proposed in Baranowski and Fryzlewicz (2017) ([1]). The concept of the iterative procedure is the measure w_j which is computed to select the variables, may unable to detect some important variables where a strong dependence between covariates is observed. It can be happened that a covariate may be jointly related but marginally uncorrelated to the response variable ([5]). In this case some important variables can be missed by the RBVS algorithm. Under the algorithm firstly need to set $\hat{S} = \emptyset$ and to repeat (1) define $Z_i^* = (Y_i^*, X_{ij}^*, j \in \{1,\dots,p\} \backslash \hat{S})$, $i = 1,\dots,n$, where Y_i^*, X_{ij}^* are the residuals left after projecting Y, X_j onto the space spanned by the covariates whose indices are in $\hat{S}$, (2) find $\hat{S}^*$ and $\hat{s}$, which are the output of the RBVS algorithm, (3) set $\hat{S} = \hat{S} \cup \hat{S}^*$ until $\hat{s} = 0$, where $\hat{S}$ is the estimate of the set of active covariates. This study attempts to incorporate the lasso, MC+ and Pearson correlation coefficient (PC) through the RBVS and IRBVS algorithms for variable selection under right-censored data. The implementation of these methods are clearly stated in **rbvs** R package.

c. The Competitors of RBVS/IRBVS Algorithms: SIS and StabSel

Sure Independence Screening ([5]) is a sure screening method to select important variable from high dimensional data. This method is developed to reduce dimension from high to a relatively lower scale preferably lower than the sample size. Stability selection is another popular method for variable selection taken by Meinshausen and Buhlmann (2010) (citemeinshausen2010stability). In StabSel algorithm, firstly a variable selection technique is chosen and then applying this technique, a sub-samples of the data of size $[\frac{n}{2}]$ is picked up randomly. Then, a set of important variables is selected by an initial procedure. As the initial procedure, the selection probabilities of variables are computed and select those variable whose selection probabilities exceed a pre-specified threshold.

3. Results

a. Simulation Studies

The logarithm of the survival time is generated from the true AFT model as defined in equation (1), where σ is set as 1 in the simulation studies. The correlated data sets were generated using the Cholesky decomposition. The censoring times are generated from a particular distributions so that it can maintain the desired censoring rate, say $P\%$. Here we consider log-normal AFT models and three level of censoring, $P\%-$ 10, 30, and 50. For each method in this simulation setting, the probability of being selected for each variable are recorded and the average number of relevant variables selected in the final model and the following error measures are also recorded: the number of False Positives (FP, the number of irrelevant variables incorrectly identified as the relevant variables), the number of False Negatives (FN, the number of relevant variables incorrectly identified as the irrelevant variables) all over 100 simulated data sets. These error measures are often adopted for evaluating the performance of variable selection. The number of sample splits B=500 is considered for the Tables. In the figures, q indicates the number of significant variables which is imputed in simulation study.

Table 1: Average number of relevant variables selected in the final model, False Positive (FP) and False Negative (FN) rates for methods RBVS, IRBVS, ISIS and StabSel from 100 simulation runs. Both RBVS and IRBVS have used B = 500 and m = $n/2$.

$P\%$	Parameter	RBVS			IRBVS			ISIS		StabSel	
		PC	Lasso	MC+	PC	Lasso	MC+	Lasso	MC+	Lasso	MC+
						$n=50,\ p=100,\ r=0$					
10%	variable count	3	5	5	5	5	5	8	6	5	4
	FP	0.006	0.003	0.004	0.015	0.009	0.008	0.037	0.018	0.011	0.000
	FN	0.427	0.000	0.000	0.140	0.000	0.000	0.015	0.057	0.000	0.017
	FP+FN	0.433	0.003	0.004	0.155	0.009	0.008	0.052	0.075	0.011	0.017
30%	variable count	3	5	5	4	5	6	11	7	5	4
	FP	0.012	0.005	0.003	0.013	0.010	0.013	0.075	0.039	0.011	0.001
	FN	0.650	0.047	0.017	0.440	0.065	0.005	0.055	0.245	0.050	0.222
	FP+FN	0.662	0.052	0.020	0.453	0.075	0.018	0.130	0.284	0.061	0.223
50%	variable count	2	4	4	2	5	6	12	8	4	2
	FP	0.004	0.007	0.004	0.008	0.014	0.016	0.081	0.051	0.007	0.001
	FN	0.745	0.257	0.190	0.695	0.175	0.062	0.087	0.317	0.277	0.617
	FP+FN	0.749	0.264	0.194	0.703	0.189	0.078	0.168	0.368	0.284	0.618
						$n=50,\ p=100,\ r=0.25$					
10%	variable count	2	5	5	4	5	5	10	5	6	4
	FP	0.005	0.006	0.005	0.016	0.008	0.010	0.060	0.014	0.017	0.001
	FN	0.647	0.002	0.017	0.435	0.000	0.000	0.037	0.260	0.003	0.075
	FP+FN	0.652	0.008	0.022	0.451	0.008	0.010	0.097	0.274	0.020	0.076
30%	variable count	4	5	4	5	6	5	9	3	6	4
	FP	0.029	0.006	0.005	0.029	0.014	0.013	0.053	0.008	0.016	0.001
	FN	0.742	0.067	0.142	0.547	0.017	0.110	0.117	0.437	0.017	0.240
	FP+FN	0.771	0.073	0.147	0.576	0.031	0.123	0.170	0.445	0.033	0.241
50%	variable count	3	3	3	3	4	4	9	2	5	2
	FP	0.020	0.006	0.006	0.015	0.015	0.014	0.060	0.007	0.014	0.002
	FN	0.785	0.452	0.412	0.730	0.362	0.342	0.292	0.660	0.242	0.642
	FP+FN	0.805	0.458	0.418	0.745	0.377	0.356	0.352	0.667	0.256	0.644

i. Example 1: Case-I ($p<n$)

We consider p = 40 with two blocks: β coefficients for $j \in$ 1,...,4 are set to be 5 and the remaining β coefficients (i.e. $j \in$ 5,...,40) are set to be zero. We also set $X \sim U(0,1)$ for n = 50. We consider log-normal AFT models and based on these models, the survival time is generated by using the Eq. (1) with $\epsilon \sim N(0,$ 1). The censoring time is generated from the Uniform distribution.

The finding reveals that for uncorrelated data the FP+FN rates are very low for lasso and MC+ methods under the RBVS and IRBVS methods at all level of censoring. For correlated data set (when r = 0.25) the performance of the lasso and MC+ methods under RBVS and IRBVS are reasonably very good under 10% and 30% censoring. But for 50% censoring, the FP+FN rates are lower for the StabSel than the other methods, but it provides the higher error rates under lower level of censoring. The most striking result is that the FP+FN rates increases for RBVS, IRBVS and ISIS methods when the censoring increases, but the rates decrease under stability selection method. For the RBVS-PC and IRBVS-PC, the results are not so impressive since the error rates are high compared to the other methods. So, it reveals that the RBVS and IRBVS based methods except PC perform very good under low dimensional settings.

ii. Example 1: Case-II (*p*>*n*)

Simulation I: We consider *p* = 100 covariates with two blocks: *β* = 5 for *j* ∈ 1,...,4 (i.e. *q* = 4) and *β* = 0 for *j* ∈ 5,...,100. We keep everything else similar to the Case-I of Example 1.

The results of Table 1 show that for uncorrelated data all methods except for the PC give low rates of FP+FN at 10% censoring, but among them, the lasso and MC+ under RBVS and IRBVS provide very small rates i.e. their FN rates are zero. At 30% censoring, the StabSel-lasso gives impressive result. The IRBVS-MC+ performs very well among all the methods under all level of censoring. For correlated data, the FP+FN rates of PC under both RBVS and IRBVS are very high as their false negative rates are high. This indicates that, they can not detect all the significant variables. For lasso and MC+, these methods give small error rates under 10% and 30% censoring. The rates are low for the StabSel-lasso at all level of censoring.

Simulation II: We consider *p* = 200 covariates with two blocks: *β* coefficients for *j* ∈ 1,...,4 (i.e. *q* = 4) are set to be 5 and the remaining *β* coefficients (i.e. *j* ∈ 5,...,200) are set to be zero. We keep everything similar to the previous simulation example. The results are shown in figures instead of tables. The Figure 1 shows that for the StabSel-MC+ with *r*= 0, the FP+FN rates are very high for 10%, 30% and 50% censoring. This rates are smaller in StabSel-lasso for *r* = 0.25 than RBVS-lasso and IRBVS-lasso. All methods with *r* = 0 give good

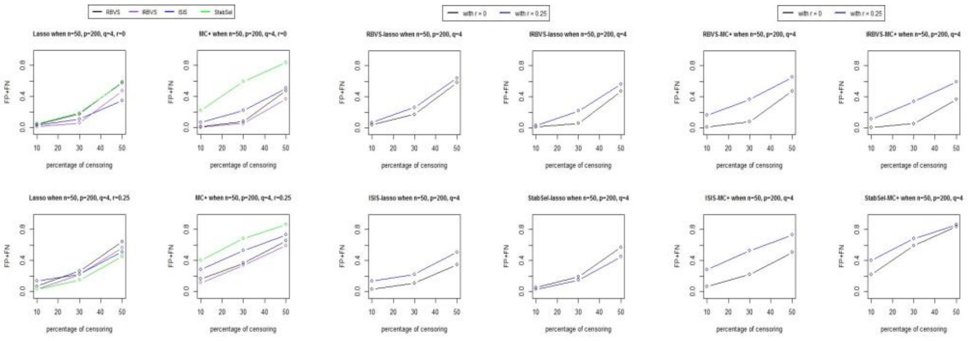


Figure 1: The percentage of censoring vs FP+FN for all methods implemented for lasso and MC+ when n=50, p=200, q=4.

result (middle four plots) compared to the methods with r = 0.25 except for the stability selection. But the four right plots show that all methods for r = 0 provides small FP+FN rates than the methods for r = 0.25.

3.2 Mantle Cell Lymphoma Data Analysis

The MCL comprises about 6% of all non-Hodgkins lymphomas as well as a higher fraction of deaths from lymphoma, given that it is an incurable malignancy ([12]). The median survival is approximately 3 years. Our concern is to select the gene expressions which are correlated with survival in this MCL patients. The expression values of 8810 cDNA elements are considered as the explanatory variables in the gene expression datasets. The data set is available at http://llmpp.nih.gov/MCL/. For simplicity, we have directly used the pre-processed data as used in Khan and Shaw (b2016) ([8]). All the techniques are employed for this dataset. Here we also use 10fold cross validation to select tuning parameters and for iterative sure independence screening, BIC is considered. We are interested here to select the most important genes which are highly related to the response variable. The results are reported in Table 2 which is conducted for the subsample size$m = \frac{n}{2}$ for ranking based variable selection and stability selection methods. The diagonal elements of this table indicate the number of genes selected by the methods and there are some common genes (off diagonal elements) that are found between the methods. The Table shows that, the lasso and MC+ procedure under RBVS and IRBVS as well as the StabSel-MC+ select smallest number of genes (1). ISIS-lasso select the maximum number of genes (21). The ISIS-MC+, RBVS-PC and IRBVS-PC select 20, 13, 14 genes, respectively. There are 4, 2, 1 common genes are found when we match RBVS-PC with ISIS-lasso, ISIS-MC+, and StabSel-lasso, respectively. If we match the IRBVS-PC+ with the RBVS-PC, ISIS-lasso, ISIS-MC+, then 11, 4, 2 common genes are found respectively. Finally, we select those genes which are selected by most of the variable selection techniques used here. According to this Table, there are four genes with uniqid 15936 [checkpoint homolog (S. pombe)], 24376 [proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin)], 27762 [cell division cycle 20 homolog (S. cerevisiae)], 28343 [Guanine nucleotide binding protein (G protein), gamma 10] are selected by the methods.

Table 2: Number of genes selected by the methods (diagonal elements) and number of common genes found between the methods (off diagonal elements) when $m = \frac{n}{2}$.

Methods	RBVS-PC	RBVS-lasso	RBVS-MC+	IRBVS-PC	IRBVS-lasso	IRBVS-MC+	ISIS-lasso	ISIS-MC+	StabSel-lasso	StabSel-MC+
RBVS-PC	13	1	0	11	1	0	4	2	1	0
RBVS-lasso	1	1	0	0	1	0	1	1	1	0
RBVS-MC+	0	0	1	1	0	0	1	1	0	0
IRBVS-PC	11	0	1	14	0	0	4	2	0	0
IRBVS-lasso	1	1	0	0	1	0	1	1	1	0
IRBVS-MC+	0	0	0	0	0	1	0	0	0	0
ISIS-lasso	4	1	1	4	1	0	21	8	1	0
ISIS-MC+	2	1	1	2	1	0	8	20	1	0
StabSel-lasso	1	1	0	0	1	0	1	1	2	0
StabSel-MC+	0	0	0	0	0	0	0	0	0	1

4. Discussion and Conclusion

The four variable selection methods–RBVS, IRBVS, ISIS and StabSel are considered to implement with the accelerated failure time models (AFT) directly to right-censored data under its high or low dimensionality. Among the methods, two ranking based methods, RBVS, IRBVS, are developed based on the original ranking based variable selection approach as suggested in Baranowski and Fryzlewicz (2017) ([1]). Other two methods are extended based on two greedy variable approaches in literature. The AFT models under these approaches are estimated by the well known Stute's weighted least squares technique. The Pearson's correlation coefficient, the regression coefficients estimated via the lasso ([13]) and the regression coefficients estimated via the MC+ algorithm ([14]) are also considered as variable selection techniques under both RBVS and IRBVS algorithms. For iterative procedure of sure independent screening and stability selection (StabSel), the lasso and MC+ algorithms are applied to both the low and high dimensional censored data. The simulation study results revealed that the methods based on RBVS and IRBVS except the PC method performed reasonably well under low dimensional data. The findings suggest that for low dimensional correlated and uncorrelated datasets, both the RBVS and IRBVS techniques give good results. For correlated high-dimensional censored datasets, the modified RBVS and IRBVS methods perform reasonably very well at both low and medium censoring levels but slightly poorer at high censoring level. The overall performance of the lasso and MC+ procedure under RBVS and IRBVS is considerably good for both low and high dimensional data sets. It is noticed that as a variable selection technique, ranking based variable selection algorithm and its iterative version performed reasonably well for low dimensional data and also for high dimensional data particularly when there is no correlation among the covariates. If the correlation is considered, the performance of RBVS and IRBVS is not found that much impressive particularly at high censoring levels. This also suggests that new techniques are required to introduce to improve the results under this situation.

The Mantle Cell Lymphoma (MCL) real data is used to identify the genes that are significantly associated with the lifetime data. We implemented the methods by considering subsample size $m = \frac{n}{2}$ for RBVS, IRBVS, ISIS, and StabSel methods. Four genes with uniqid 15936 [checkpoint homolog (S. pombe)], 24376 [proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin)], 27762 [cell division cycle 20 homolog (S. cerevisiae)], 28343 [Guanine nucleotide binding protein (G protein), gamma 10] are selected by most of methods from the MCL data.

References

1. Baranowski, R. and Fryzlewicz, P. (2017). Ranking-based variable selection for high-dimensional data. *Working Paper: URL http://personal.lse.ac.uk/baranows/rbvs.html*.
2. Bickel, P. J., G¨otze, F., and van Zwet, W. R. (2012). Resampling fewer than n observations: gains, losses, and remedies for losses. In *Selected Works of Willem van Zwet: Part of the Series Selected Works in Probability and Statistics*, pages 267–297. Springer.
3. Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232.
4. Cho, H. and Fryzlewicz, P. (2012). High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: Series B*, 74(3):593–622.
5. Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5):849–911.
6. Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3):533–550.

7. He, X., Wang, L., Hong, H. G., et al. (2013). Correction: Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(5):2699–2699.
8. Khan, M. H. R. and Shaw, J. E. H. (2016). Variable selection for survival data with a class of adaptive elastic net techniques. *Statistics and Computing*, 26(3):725–741.
9. Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.
10. Meinshausen, N. and Bu¨hlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473.
11. Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics*, 23(4):461–471.
12. Swerdlow, S. H. and Williams, M. E. (2002). From centrocytic to mantle cell lymphoma: a clinicopathologic and molecular review of 3 decades. *Human Pathology*, 33(1):7–20.
13. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.
14. Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Bivariate smooths within GAMLSS for spatially correlated count data

F. De Bastiani[1], R. A. Rigby[2], D. M. Stasinopoulos[2], M. A. Uribe-Opazo[3], L. M. Oliveira[4]

[1]Universidade Federal de Pernambuco
[2]London Metropolitan University
[3]Universidade Estadual do Oeste do Parana
[4]Universidade Federal de Pernambuco

Abstract

This paper analyzed the number of adolescent workers in Pernambuco region of Brazil, people aged from ten to seventeen-year-old. It explores the possibilities of fitting bivariate smoothing within the GAMLSS framework to take account the spatial dependence between the observations. GAMLSS also allows to model any or all of the parameters of a non-exponential family distribution for the response variable. The focus of this paper is to consider three different approaches for spatial modeling: kriging, tensor product and thin plate splines. The potential of using spatial analysis within GAMLSS is discussed.

Keywords

kriging; negative binomial; tensor product splines; thin plate splines

1. Introduction

The methods for spatial analysis have rapidly become popular due to demand from a wide range of fields. In Geostatistics, the methodology developed to predict values in non-sampled sites is referred in the literature as “kriging", as presented by Matheron (1963). Alternatives to kriging, are for instance, multivariate versions of smoothing techniques popularized by Hastie and Tibshirani (1990) and of the P-spline approach of Eilers and Marx (1996). Fahrmeir et al. (2013) show details about these three smoothings techiniques, kriging, thin plate splines and tensor product. The main goal of this work is to provide spatial modelling facilities within the GAMLSS framework, Rigby and Stasinopoulos (2005) to analyze the number of adolescent workers aged from ten to seventeen years old in Pernambuco region of Brazil, related to some potential explanatory variables. This would allow the fitting of a variety of different distributions for the response variable, rather than restricting to the exponential family distribution used within the generalized linear models framework, and also allow spatial modelling of any or all parameters of the response variable distribution. In this paper we are concentrating on a thin plate splines, tensor product splines and kriging within GAMLSS, for a count type of data set.

2. Methodology

2.1 Description of dataset

There are 184 observations, of count response variable together with some explanatory variables. The number of adolescent workers aged from ten to seventeen years old are from each city of Pernambuco State in Brazil in the year 2010. We modelled the response variable considering one of the count distributions available in GAMLSS. More details about available distributions for count type of data available in the R package gamlss are presented in Stasinopoulos et al. (2017). Our aim is to investigate variables that effect the number of adolescent workers and also to understand the spatial distribution of labor. The variables are: Labor Inf: number of infant, adolescent from ten to seventeen years working in the week that the data where collected in 2010; CVLI: log of an indicator constituted for the crimes of homicide, felony, murder and corporal injury followed by death; GINI: index that measures the degree of inequality in the distribution of individuals according to per capita household income. Its value varies from 0, when there is no inequality and 1, when the inequality is maximal; illiteracy: log of illiteracy rate of the population aged 15 years or over; GDP: log of the gross domestic product; POP: the population living in each city, POP Young: log of young population divided by age groups; Poor log of proportion of individuals with household income per capita is equal to or less than BRL 140:00 monthly, in August 2010. The universe of individuals is limited to those living in permanent private households; IFDM: Firjan index development of a municipality; Infant death: number of infant deaths. Figure 1 shows the spatial coordinates corresponding to the centroide for each city. The shading indicates the number of adolescent workers in each city. In the plot it is evident where is sited the city with highest number of adolescent workers, which is Recife, the capital of Pernambuco region.

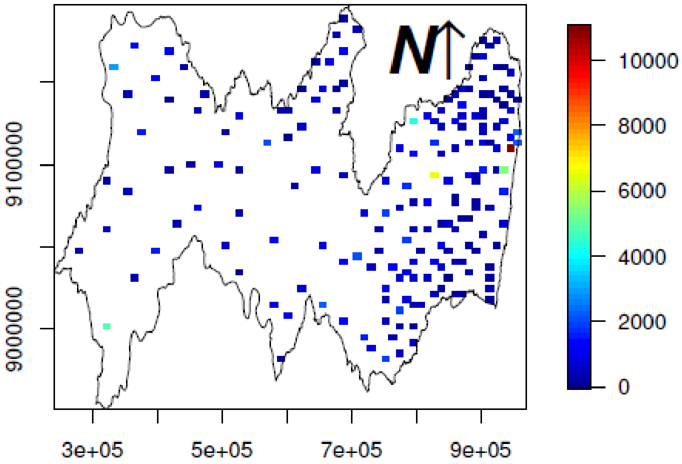


Figure 1: The number of adolescent workers in cities of Pernambuco

2.2 The GAMLSS

GAMLSS is a distribution based regression model. The distribution of the response variable can be selected by from a very wide range of distributions available in the gamlss package in R including highly skewed and kurtotic continuous and discrete distributions.

Specifically, a GAMLSS model assumes that, for i = 1, 2, ... , n, observations Y_i have independent probability (density) function $f_Y(y_i \mid \theta_i)$ conditional on $\theta_i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i})^T = (\mu_i, \sigma_i, \nu_i, \tau_i)^T$ a vector of four distribution parameters, each of which can be a function to the explanatory variables. Rigby and Stasinopoulos (2005) define the original formulation of a GAMLSS model as follows. Let $g_k(.)$ be a known monotonic link function relating the distribution parameter $\theta_k = (\theta_{k1},\dots, \theta_{kn})^T$ to predictor $\eta_k = (\eta_{k1}, \dots, \eta_{kn})^T$, for k = 1, 2, 3, 4.

To include random effect in the model for each parameter, see Rigby and Stasinopoulos (2005):

$$Y|\gamma \overset{ind}{\sim} D(\mu, \sigma, \nu, \tau)$$

$$\begin{aligned}
g_1(\mu) &= \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} Z_{j1}\gamma_{j1} \\
g_2(\sigma) &= \eta_2 = X_2\beta_2 + \textstyle\sum_{j=1}^{J_2} Z_{j2}\gamma_{j2} \qquad (1) \\
g_3(\nu) &= \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} Z_{j3}\gamma_{j3} \\
g_4(\tau) &= \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} Z_{j4}\gamma_{j4}
\end{aligned}$$

where X_k is a known design matrix and the x_{jk}'s are vectors of length n, s_{jk} is a smooth nonparametric function of variable $X_{jk}, \beta_k = (\beta_{1k}, \dots, \beta_{j'_k k})^{\top}$ is a parameter vector of length J'_k, for k = 1, 2, 3, 4 and j = 1, . . . , J_k, D is any distribution with up to four distribution parameters, γ_{jk} have independent (prior) normal distributions with $\gamma_{jk} \sim N_{qjk}$ (0, G_{jk}^{-1}) and G_{jk}^{-1} is the (generalized) inverse of a a q_{jk}xq_{jk} symmetric matrix $G_{jk} = G_{jk}(\lambda_{jk})$ which may depend on a vector of hyper-parameters λ_{jk}. Different formulations of the Z's and the G's result in different types of additive terms, for example, random effects terms, smoothing terms, time series terms or spatial terms. The parametric vectors β_k and the random effects parameters γ_{jk}, for $j = 1, 2, \dots, j_k$ and $k = 1, 2, 3, 4$ are estimated within the GAMLSS framework (for fixed values of the smoothing hyper-parameters λ_{jk}) by maximizing a penalized likelihood function ℓ_p given by

$$\ell_p = \ell - \frac{1}{2} \sum_{k=1}^{p} \sum_{j=1}^{J_k} \gamma_{jk}^{\top} \mathbf{G}_{jk} \gamma_{jk}$$

where $\ell = \sum_{i=1}^{n} \log f(y_i|\theta^i)$ is the log-likelihood function. Estimation of the smoothing hyper-parameters (i.e. λ's) can be achieved by a local maximum likelihood method, see Rigby and Stasinopoulos (2013) for more details.

i. Spatial models within GAMLSS

De Bastiani et al. (2018) presented Gaussian Markov random field spatial models in GAMLSS, to analyzed lattice data. The following describes spatial models for data with spatial continuity (different of lattice data) and the way that they are or could be implemented within the GAMLSS framework.

In thin plate spline the basis are given by $Z_{im} = B(||s_i - s_m||)$ where $s_i = (s_{i1}, s_{i2}), s_m = (s_{m1}, s_{m2})$ and $||x_i - x_m|| = (s_{i1} - s_{m1})^2 + (s_{i2} - s_{m2})^2$ is the distance between s_i and s_m and B is a radial function (given by Wood (2017)).

Tensor product can also be included within the GAMLSS model (1), where the ith row Z_i of the $n\ x\, (q_1 q_2)$ combined Z basis matrix is obtained by a Kronecker product of the ith rows Z_{i1} and Z_{i2} of the individual basis matrices $Z_1$and Z_2 for the two variables in the tensor product bivariate spline, i.e. $Z_i = Z_{1i} \otimes Z_{2i}$for i= 1, 2, ... , n, and γ are the corresponding coefficients. The $(q_1 q_2)$ x $(q_1 q_2)$ combined bivariate penalty G can be constructed from a Kronecker product of the two univariate penalty matrices:

$$G = G_1 \otimes I_{q2} + I_{q1} \otimes G_2$$

where G1 and G2 are the individual univariate penalty matrices. This results in a quadratic penalty $\gamma^{\top} G \gamma$ like in equation (2).

In kriging the spatial effects are described in terms of stochastic process, a probabilistic modeling framework. Let consider a Gaussian field γ(s), $s \in R^2$ characterized by the expectation function $E(\gamma(s)) = 0$, the variance function Var$(\gamma(s))$, and the correlation function $C(\gamma(s_i), \gamma(s_m)) = C(\gamma(s_i), \gamma(s_m))$, for $i, m = 1 \dots, n$, where the spatial random effect γ_i for observation i is a function γ_{si} of the spatial coordinate s_i for observation i, i.e. $\gamma_i = \gamma(s_i)$. For stationary Gaussian fields, the expected value and variance are spatially constant, and the correlation function only depends on the difference $s_i - s_m$. For the special case of isotropic correlations functions

$$C(\gamma(s_i), \gamma(s_m)) = C(||s_i - s_m||) = C(d_{m)} = \emptyset_{2^r}{}_{im'}$$

with $d_{im} = ||s_i - s_m||$ is the distance between s_i and s_m , and $\mathrm{R} = [(r_{im})]$ is an $\mathrm{n \; x \; n}$ matrix with each entry being the correlation between the value of γ at two points, and this matrix may have different forms depending on the parametric form assumed for the variance-covariance matrix.

3. Result

We used a forward stepwise selection with both linear and smooth terms in all explanatory variables for each parameter of each of Poisson, *PO(μ)*, negative binomilal, $NBI(\mu, \sigma)$ and beta negative binomial, $BNB(\mu, \sigma, v)$ models. According to $GAIC(k = 4)$ the 'best' model was the negative binomial, $NBI(\mu, \sigma)$ model. The resulting model was:

$$\begin{aligned} &Y \sim NBI\ (\mu, \sigma) \\ &\log \mu = POP + illiteracy + Poor \\ &\qquad\quad + IFDM + Infant\_death \\ &\log \sigma = GINI \end{aligned}$$

which gave $GAIC(k = 2) = 2589.96$, $GAIC(k = 4) = 2603.96$ and $GAIC(k = \log(184)) = 2612.465$. Then we start adding spatial effects (kriging, thin plate spline or tensor product) for each parameter of the distribution.

Comparing the values of the GAIC for the models with spatial effect (ommited here) and the GAIC for the model given in Equation (2), the best choice according to $GAIC(k = 2)$, $GAIC(k = 4)$ and $GAIC(k = \log(184))$ is the negative binomial model (2) with additional tensor product bivariate smooths for both μ and σ, i.e. using the spatial information to model both the parameters that control the mean and variance of the distribution.

Figure 2 displays a worm plot, van Buuren and Fredriks (2001), for the residuals of the chosen fitted model. The worm plot is a detrended normal QQ plot of the residuals, which here indicates a reasonable fit to the data, since over 95% of the points lie within the elliptical (dashed) 95% pointwise interval bands.

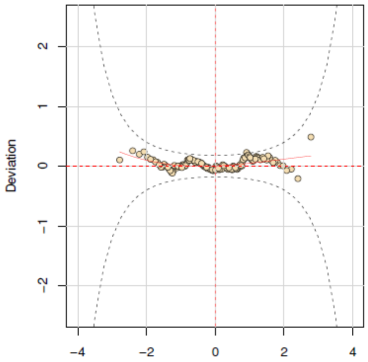


Figure 2: Worm plot of the residuals for the chosen negative binomial model, with spatial effect for 1a and a fitted with tensor product

Figure 3 (left) shows the site effect on $\log(\hat{\mu})$ where we can see that the number of adolescent workers is highest in the eastern region, which is on the coast of the state including the capital Recife. The second plot of Figure 3 (right) shows that the effect of the spatial information in the $\log(\hat{\sigma})$ is in the north-south direction.

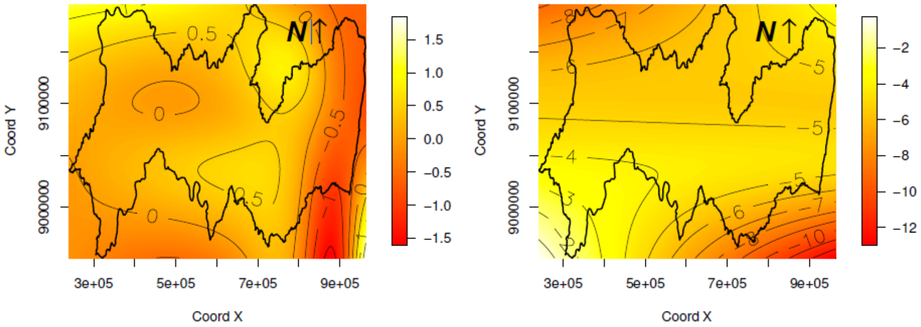


Figure 3: Fitted spatial effect for $\log(\mu)$ (left) and $\log(\sigma)$ (right) for the chosen negative binomial model using tensor product spatial effects

4. Discussion and Conclusion

It is possible to model spatial data within the GAMLSS. The resulting methodology allows fitting of non exponential family response variable distributions and also allows spatial modelling not only of the location parameter of the response variable distribution, but also other parameters of the distribution. Spatial effects must be considered for both μ and σ in the negative binomial, $NBI(\mu, \sigma)$, distribution. Spatial analysis allows public policies to be more strategic. Indeed, if the adolescent worker cases are spatially concentrated, it is more efficient to direct policies towards these places.

Acknowledgments
The authors would like to thank you the support of World Bank Trust Fund for Statistical Capacity Building (WB TFSCB), National Council for Scientific and Technological Development (CNPq) and Coordination for the Improvement of Higher Education Personnel (CAPES).

References

1. F. De Bastiani, D. M. Stasinopoulos, R. A. Rigby, A. H. M. A. Cysneiros, and M. A. Uribe-Opazo. Gaussian markov random field spatial models in gamlss. *Journal of Applied Statistics,* 45(1):168–186, 2018.
2. P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science,* 11:89–121, 1996.
3. L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: models, methods and applications.* Springer, 2013.
4. T. J. Hastie and R. J. Tibshirani. *Generalized additive models.* Chapman and Hall, London, 1990.
5. G. Matheron. Principles of geostatistics. *Economic Geology,* 58:1246–1266, 1963.
6. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics,* 54:507–554, 2005.
7. A. Rigby and D. M. Stasinopoulos. Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. *Statistical Methods in Medical Research,* 23(4):318–332, 2013. doi: 10.1177/0962280212473302.
8. D. M. Stasinopoulos, R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani. *Flexible Regression and Smoothing: Using GAMLSS in R.* Chapman and Hall, Boca Raton, 2017.
9. van Buuren and M. Fredriks. Worm plot: A simple diagnostic device for modelling growth reference curves. *Statistics in Medicine,* 20:1259–1277, 2001.
10. N. Wood. *Generalized additive models. An introduction with R.* Chapman and Hall, 2nd edition, 2017.

Understanding greenhouse gas emission patterns in USA under a data science approach

Karina Gibert[1,2,4], Miquel Sànchez-Marrè[1,3,4], Andrea Galassi[4], Lovisa Persson[4], Natalia Sarmanto[4], Maria Jean Carla Prado[4], Gabriel Perelló[4], Nurhanim Kamarulzaman[4]

[1] Intelligent Data Science and Artificial Intelligence Research Center
[2] Statistics and Operations Research Department
[3] Computer Science Department
[4] Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona (Catalonia)

Abstract

Data science was performed on USEPA s GHG report of RY2017 data in order to understand trends in GHG emissions from United States facilities. The USEPA mandates that under the GHG reporting program all industrial facilities emitting more than 25 tons/year of GHG emissions must report emissions annually to the USEPA. Variables could have considered in three categories: geographical location (state, latitude and longitude), emissions, and type of industry (qualitative). Emission data was reported in metric tons of CO2eq per year (tpy) where the emissions of each chemical constituent was multiplied by the GWP for each constituent to obtain CO2eq. Basic descriptive analysis and data pre-processing was followed by clustering with heterogeneous data, profiling and principal component analysis to get a global overview of the phenomenon. The analysis showed that CO2-fossil is the main GHG emission contributing the total reported CO2-eq and hence to global warming. The clusters showed that the type of emmissions dominated the structure of the dataset and each cluster was associated with a particular combination of emmissions and highly associated by a certain industrial sector. From PCA it was confirmed than the greatest contribution to CO2-eq comes from burning fossil fuels and that this is directly related to the production of electricity from the power plant sector. In addition, observing the relationship of the gases with the geographic coordinates, it has been observed that the distribution of both type of emission and the type of industry is not homogeneous and certain geographical patterns of industrial sectors along the territory could be identified. This work shows how an integral data science view can provide interesting information about complex processes like industrial activities and emmissions in a certain territory and can be used as an input to develop sustainable policies in the near future.

Keywords

Greenhouse gases, Data Science, Sustainability

1. Introduction

Statistical analysis were performed on the annual greenhouse gas (GHG) emissions report for reporting year 2017 (RY2017) of the for the United States of America. This GHG emissions report was published by the United States Environmental Protectioon Agency (USEPA) and is considered an accurate representa_on of the na_onal industrial GHG emissions during RY2017 (January 1, 2017 - December 31- 2017) as federal law requires all facilities who emit over 25 tons per year (tpy) of GHG emissions submit their emissions report to the USEPA annually.

In this work, a data science approach is used to find the relationships between pollutants and the distribution of industrial sectors in USA. Greenhouse gas (GHG) data collected from the GreenhouseIntroduction Gas Reporting Program (GHGRP) through the USEPA was used. The USEPA mandates that all facilities in numerous industrious on American soil report the quantity of greenhouse gases emitted on an annual basis per Title 40 Code of Federal Register (CFR) part 98 titled Mandatory Greenhouse Gas reporting. Title 40 is the Federal regulations that manage all of the United States of America's environmental codes and regulations (CFR2019). An interesting subpart of the data collected includes greenhouse gas emissions from facilities which perform hydraulic fracturing a.k.a. fracking which is a relatively new technologies which employs drilling and high pressure water to access oil and gas reserves from onshore basins which have traditionally been considered harder to access oil and gas reserves. The USEPA defines a greenhouse gas as "carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), sulfur hexafluoride (SF6), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), and other fluorinated greenhouse gases" (40 CFR 98.6).

The global warming potential (GWPs) for each GHG were accounted for in the reported emissions as each emissions constituent was reported in carbon dioxide equivalent (CO2eq) according to the GRP calculation methodologies required. All emissions data is presented in of metric tons of carbon dioxide equivalent using GWP's from IPCC's AR4 (see FAQs tab).

2. Methodology

2.1 Data

The data used in this work is the dataset corresponding to the Greenhouse Gas Reporting Program for 2017. This data was reported to EPA by all applicabile facilities which emitted over the reporting threshold of 25 tons of CO2 equivalents combined for the facility for the relevant source categories. Data corresponds to the one used for the reporting year 2017 (RY2017) and is from 01/01/2017 to 12/31/2017 from the EPA's FLIGHT tool on 08/19/2018 and contains information about 6451 facilities with 84 columns. The data can be divided into three categories:

- location of facility, includes State and geolocalization
- typology of the industry, including Primary NAICS Industry (The North American Industry Classification System, see 3.1) and Sector of activity (Waste, Natural Gas and Natural Gas Liquids Suppliers, Power Plants, Minerals, Petroleum and Natural Gas Systems, Industrial Gas Suppliers, Pulp and Paper, Chemicals, Metals, Suppliers of CO2 , Petroleum Product Suppliers, Injection of CO2 , Import and of Equipment Containing Fluorinated GHGs, Refineries, and Coal-based Liquid Fuel Supply, Other).
- GHG emission reported in CO2-eq for each facilities' industry, including: Total reported direct emissions, CO2 emissions (non-biogenic), Methane (CH4) emissions, Nitrous Oxide (N2O) emissions, HFC emissions, PFC emissions, SF6 emissions, NF3 emissions, Other Fully Fluorinated GHG emissions, HFE emissions, Very Short-lived Compounds emissions, Other GHGs (metric tons CO2 -eq), and Biogenic CO2 emissions (metric tons).

2.2 Methodology

A data science approach has been used to find structural patterns in the GHG on USA:

1) Preprocessing. In this particular appliaction it includes missing data treatment, feature selection, recoding
2) Descriptive Analysis: visualizes the selected variables and interact with preprocessing to elicit specific data cleaning operations required in the data
3) Clustering and profiling: Hierarchical clustering with Wards method modified with squared Gower similarity coefficient to guarantee metrics properties is used, provided that no previous assumptions on the number of clusters exist and both qualitative and quantitative information is used simultaneously for the analysis. Calinski-Harabask index is optimized to find the resulting number of clusters. Profiling techniques by analyzing conditional distributions of variables against resulting clusters are used to interpret the clusters. Graphical representations and Kruskall-Wallis tests are used to identify which emmissions were signifficant for which clusters
4) Multivariate analysis. Principal component analysis is used to understand the relationships among pollutants and its global association with the type of industry
5) GIS representation of different parameters of the analysis is included, provided that geolocalization of the facilities is provided and territorial information is analyzed. This will allow to find geographical patterns on top of the results provided by clustering and PCA.
6) Integrated interpretation and visualization and knowledge production

3. Results

Preprocessing: Several issues are addressed through the preprocessing step, synthesized bellow:

- Variables selection: An expert-based variables selection process is followed, by eliminating all redundant variables from the analysis or columns corresponding to textual comments. The GHG emmissions were reported under two different measurement units in the dataset. For comparability reasons total emmissions by chemical constituent where retained. As a result, from the 84 original columns, 33 variables are finally targeted in this study. From then, 15 are numerical and 18 are qualitative.
- Recoding NAICS: The original 6451 facilities distribute along 247 different NAICS industry codes. Industry codes which were closely related were re-grouped into one type of NAICS industry code. For example, the NAICS industries "Automobile Manufacturing" and "Aircraft Manufacturing" were considered a similar industry segment and were classified as "Equipment Manufacturing". Additionally, "Industrial Gas Manufacturing" and "Petroleum Lubricating Oil and Grease Manufacturing" were considered a similar industry and were classified as "Oil and Gas Activities". These reduces the NAICS to a simplified version that contains 23 core NAICS Industry codes in this study: Solid waste landfill, Pipeline Transportation of Natural Gas, Chemical manufacturing, Metal and mineral activies, Food Industry, Paper and glass manufacturing, Crude Petroleum and Natural Gas Extraction, Oil and gas activies, Natural Gas Liquid Extraction, Ethyl Alcohol Manufacturing, Science and medical activies, Petroleum Refineries, Mining activities, Equipment manufacturing, Plastic Manufacturing, Waste treatment activities, Agriculture Industry, HVAC activities, Electrical power generation, Transportation activities, Other, Other manufacturing.
- Management of multivalued variables: The USA industries can be classified under several sectors simultaneously, this constituting a multivalued variable. In our study, the sectors have been converted to the complete disjunctive form and a set of 15 binary variables have been generated as dummy variables to be considered in the further steps.
- Missing data treatment: Only variables corresponding to GHG emmissions contain missing values, but they constitute a 53% of the dataset. According to the GHGRP Help Desk, all of them correspond to non-applicable values, that is, pollutants that are not produced in a particular type of industry. This supports the imputation of all these missing values by 0.

Descriptive Analysis: Classic descriptive analysis was used to elicit main characteristics of data and to guide part of the preprocessing. Fig (left) shows the main industrial activities of the dataset. Most of he GHG emmissions, distribute under very skewed distributions (Fig 1 right) and extreme values are not considered outliers.

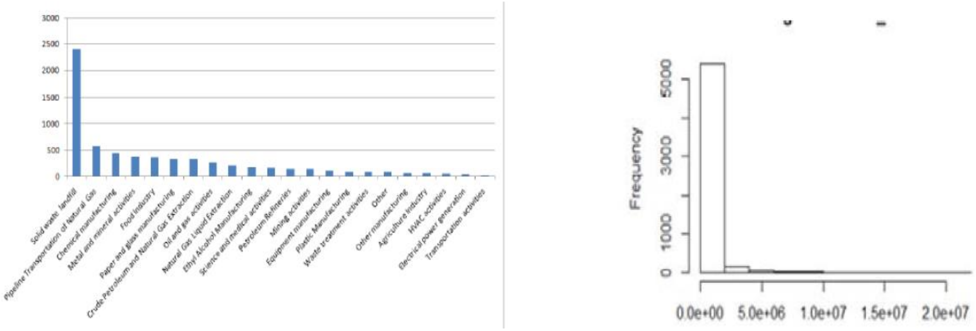


Fig 1. Simplified NAICS(left); CO2-fossil emmissions (right)

Clustering: The dendrogramm resulting from the clustering recommends 9 clusters. The profiling analysis shows a very clear association with the type of pollutant and some specific sectors, as partially shown in Fig2. Table 1 synthesizes the profiling analysis, and the main trends in each cluster

Multivariate analysis and GIS: The ACP is basically showing the relationship among pollutants and Sectors and NAICS of the facilities and verifies what was already found in the clustering. However, an interesting additional information is provided by the 4th principal component, which shows a totally negative association between longitude and latitude, this indicating a diagonal gradient existing among the distribution of the industrial sectors on the territory. Further geographical representation of both Sectors and Clusters helped to get a complete view of the situation.

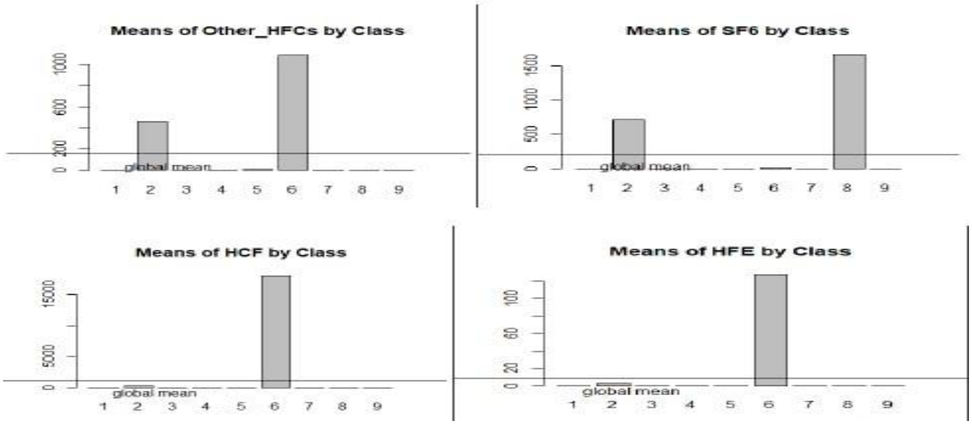


Fig 2. Conditional means of some pollutants versus clusters

4. Discussion and Conclusion

This work shows how a proper combination of several data science techniques can provide a global view of a complex phenomena such as how industrial activity distributes in a territory and how this is associated with the different kind of GHG emissions. The proposed methodology is applied to the USA data from EPA. From the descriptive analysis it is seen how Paretto Law fits and few facilites are responsible to a large ammount of emissions. CO2-fossil is the main GHG emission contributing to global warming. From cluster, it is seen that industries group according to sector and their specific combinations of pollutants. Class 4: Power plants (accounting for approximately ten times more total emissions than any other class) and Class 9: Refineries (small number of facilities with largest mean of CO2-eq per facility) were responsible for the largest share of total CO2-eq, which from a sustainability point of view makes these sectors highly problematic. The power plant sector is the largest and most polluting sector within the USA, while the refinery sector is a much smaller sector, but the worst pollutant. PCA shown that the greatest contribution to CO2-eq comes from burning fossil fuels and that this is directly related to the production of electricity from the power plant sector, confirming what was seen in the clustering. The lowest contribution on total emissions comes from fluorinated ethers (HFE) type gases, which is produced mainly by the aluminum industry. GIS shows a big concentration of industrial activity in the most populated areas of USA, such as the coastal regions, the east, coast, and New England. Also, for the Eastern part of the country, the existence of a gradient from Dakota and Minessota towards Georgia and South Caroline, where industries distribute from Alcohol production in the former to paper and glass manufacturing in the later in a certain geographical order that cannot be catched by classical statistical analysis.In fact Paper and glass industries tend to concentrate throughout the coast where the proximity to the water resources helps, as they are large consumers of water. The Gulf of Mexico also has a high concentration of petroleum and natural gas system facilities by sector as seen in the Class 3: Petroleum and Natural Gas distribution which is consistent with oil and gas activities performed in this region.

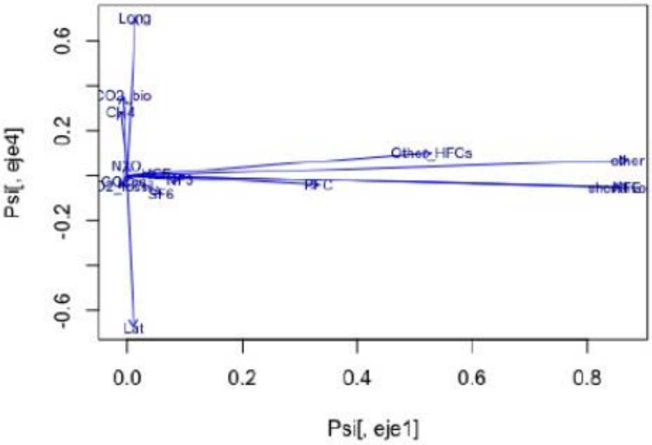


Fig 3. Fourth factorial plane of PCA

Acknowledgement

This work has been partially supported by the Research Institute for Sustainability Science and Technology of the UPC

References

1. CFR 2019: "Code of Federal Regula_ons." ECFR-Electronic Code of Federal Regula_ons: Part 98 TITLE 40—Protec_on of Environment, www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title40/40cfr98_main_02.tpl
2. European Commission. (2014, December 4). Semiconductor and aluminium industries underestimate greenhouse gas emissions. Science for Environment Policy
3. Interprofessional Technical Centre for Studies on Air Pollu_on. (2017b). Air and climate: The greenhouse effect
4. United States Environmental Protec_on Agency. (2013). Greenhouse Gas Reporting Program: Data Reported by Facilities Subject to the Direct Emitter Subparts C through II, RR, SS and TT. Publisher

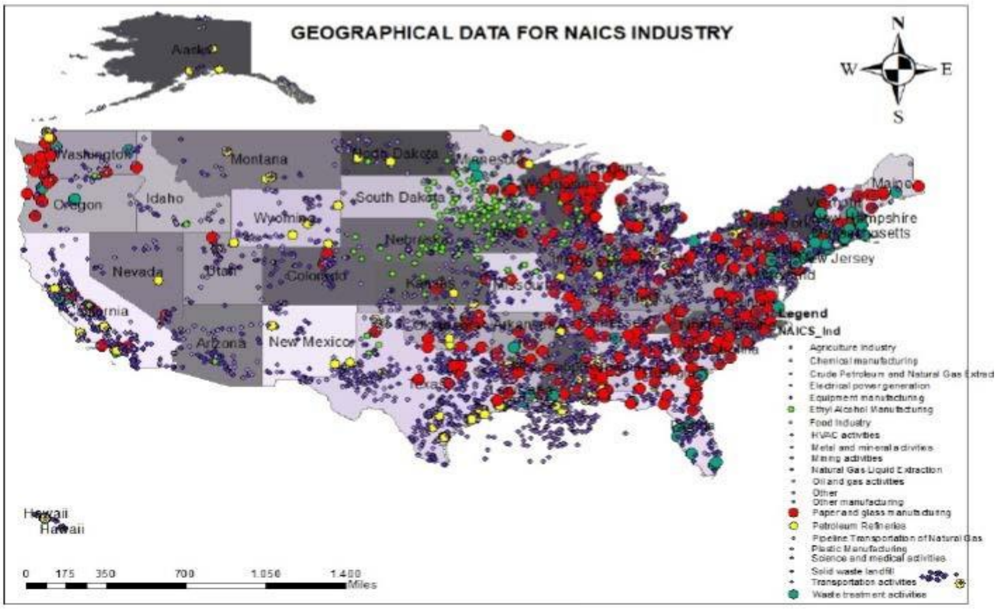


Fig 4. Geographical distribution of industries by sector

Table1. Profiling of clusters

Class	Dominating industry	Dominating sector	Main emission(s)	Profiling
1	Solid waste landfill (almost all *waste treatment activities* among classes but small share within class)	Waste	CH,	Dominated by the wane sector, consisting of approximately 60% of CH4 emissions
2	Mix of sectors (food industry, ethyl alcohol manufacturing. science and medical activities, mining activities, plastic	Other	CO, fossil & CH, [2nd biggest among classes for HFCs & SF,, PFC)	Dominated by the other sector (Sec_Other).
3	Pipeline Transportation of Natural Gas (all *Crude Petroleum and Natural Gos Extraction & Natural Gas liquid Extraction* but constitutes a smaller share within class)	Petroleum & Natural Gas Systems (all *Natural Gas & Natural Gas liquids Suppliers* but small share within class)	CO, fossil	Dominated by the petroleum and natural gas and natural gas sectors
4	Solid waste landfill (all *HVAC activities & Electrical power generation* among classes but small share within class)	Power Plant	CO2 fossil (2nd biggest among classes for N20 but small contribution as a whole)	Responsible for the largest amount of cumulative CO2eq emissions mainly consisting of CO2 fossil emissions. Dominated by the power plant sector
5	Chemical manufacturing (followed by *Metal and mineral activities & Paper and glass manufacturing)*	Minerals	CO2 fossil	Dominated by the mineral sector
6	Chemical manufacturing & Oil and gas activities (almost all *Agriculture Industry* but small share	Chemicals	CO2 fossil, N20 HFCs & SF6 HCF, HFE, other, PFC	Dominated by the chemical sector, responsible for the majority of of HCF
7	Paper and glass manufacturing	Pulp and Paper	CO2 bio & CO2 fossil	Dominated by the pulp and papersector, responsible for

8	Metal and mineral activities.	Metals	CO2 fossil PFC	Dominated by the metals sector, responsible for the
9	Petroleum Refineries	Refineries (all *Petroleum Product Suppliers* but small share within class)	COI fossil, N20	Dominated by the refinery sector, consisting of 144 facilities with the largest mean of CO2eq emissions, however due to low amount of

Estimation of the prevalence of depression via the latent class model using the national health survey sample, 2013

Rita de Cassia de Lima Idalino[1], Luzia Aparecida Trinca[2]
[1]Federal University of Piaui, Teresina, Brazil
[2]Paulista State University "Julio de Mesquita Filho", Brazil

Abstract

Depression is indicated as the non-transmissible chronic disease that will be more widespread until 2030, reaching directly or indirectly several sectors in which the population is inserted. In Brazil, the situation of depression is alarming, accounting for the highest rate in the Latin American continent. The National Health Survey (PNS) is a nationwide household survey. It is an initiative of the Ministry of Health in partnership with the Brazilian Institute of Geography and Statistics (IBGE) and aims to characterize the health situation and the lifestyles of the Brazilian population, and thus to know how health care happens in different groups of the population. In this study, depression will be the object of study because it is characterized as a complex disease of difficult measurement and observation due to its multifactorial causes. The PNS data collection used a complex sampling plan, which demands special attention in relation to the analysis of the information collected. Considering the magnitude of the research and taking into account the regional diversities, models were adjusted based on latent class theory. This approach identifies groups based on the patterns of responses observed in the categorical variables using a probabilistic model. Thus it is possible to classify each individual as belonging to a group, to estimate the prevalence and identify decisive characteristics for the emergence of groups. Based on items dealing with mental health, it was possible to propose classes with the profile of responses associated with sociodemographic issues of Brazilian adults. The results showed the influence of severity responses more markedly for age, sex and schooling. It was possible to investigate the behavior, in population terms, of depression and in spite of the complexity, to identify possible subtypes.

Keywords

depression, latent classes, complex sampling, mental health

1. Introduction

Several phenomena in the social, behavioral and health sciences can be represented by a model that considers the existence of subtypes or distinct categories in a population of individuals, although the specific category to which each individual belongs cannot be observed directly or objectively. In the health area in particular, many diseases are diagnosed or their severities accessed through manifest variables or indicators of the disease, variables that can be measured.

To aid in the diagnosis, for example, of a particular mental disorder, it is common to use a questionnaire involving several items, each with multiple categories of response, that receive a numerical score according to the individual's response. It is also common to use some cut-off point for punctuation in items or combinations of item sets as a criterion for identifying and/or classifying disease severity. Alternatively, from the standpoint of population knowledge, it is interesting to explore the relationships between items and associations of patterns of responses to the possible underlying subtypes that make up the population.

The use of a model that postulates the existence of subtypes characterizing a variable of latent classes and relates it to the manifested variables is of great potential. Such a model, known as latent class model, introduced by Lazarsfeld and Henry (1968) has shown itself to be widely applied in the social and also health areas.

With the emphasis given by health organizations to the problems related to depression, whose diagnosis does not obey an objectively measurable characteristic, combined with the availability of a national and public database, the National Health Survey, this work proposes to construct a model of latent classes with the objective of contributing to the knowledge of the profiles of the existing subgroups regarding mental disorders, amid the abundance of heterogeneities that compose the Brazilian population

2. Methodology

A latent variable is defined as one that is not observable or measurable, but evaluated indirectly through a set of two or more variables that are possible to observe. The observable variables are subject to errors or random variation and are called manifest or indicative variables because the model of latent variables postulates that these are manifestations of the latent characteristic. They can also be referred to as items, especially when they are questions entered in a questionnaire. In this context, the latent variable influences the answers that will be given to the questions, so that multiple indicators and / or scales are intended to measure reporting on the non-measured characteristic directly.

The database used in the development of this work comes from the National Health Survey is defined as a nationwide home-based survey, which arose from the need to improve and expand the Health Supplements of the National Household Sample Survey (PNAD). Surveys are essential to establish assessments under the various approaches to which they are associated. In relation to the health system, an inquiry aims, above all, to know and evaluate the service provided to a population from the perspective of who offers, as well as from the one who is a user. Surveys are efficient tools for understanding the real needs of the population and are essential for the design and improvement of public policies and programs. The PNS emerges as an essential tool to provide information that can guide social and health services to better understand the needs and expectations of the population in health promotion and prevention or in other social areas that affect the longevity and quality of life of the population.

The process for performing a latent class analysis on a data set would be the construction of contingency tables involving the items of interest in the search. For example, for m items each with R categories, the table that lists all m items contains R^m casela. If m = 9 and R = 4 we have 262.144 cells. Each box corresponds to a pattern of response to the item. The model of latent classes with the parameters, prevalences of classes and probabilities of response to the items, allows obtaining the expected frequencies in each of the boxes.

The latent class model considers that $\mathbf{Y}_i = (Y_{i1},...,Y_{iM})$ the response vector of the M categorical items associated with each individual i. It is considered that Y_{im} assumes one of the possible values in the set $1,...,r_m$, corresponding to the item response m. The observed value for $\mathbf{y}_i$ is denoted by $\mathbf{y}_i = (y_{i1},...,y_{iM})$.

In the usual model with C-classes, it is assumed that $y_{i1},...,y_{iM}$ are conditionally independent given a categorical latent variable C_i, which assumes values (classes) of $1,...,C$. The indication that Y_{im} assumes a specific category r is denoted by the indicator function such that $I(y_{im} = r) = 1$ if the individual i choose the category r with $r \in (1,...,r_m)$ of item m and 0 otherwise.

In the latent class model is possible to deal with multi-group populations so that each individual is classified a priori into one of G groups ($g = 1,...,G$). The values of parameters and or the probability of adhesion to each of the classes may or may not not very between groups. Consider g_i the group to which the individual i belongs, the measurement parameters, are defined by:

$$\rho_{m,r|c,g} = Pr(Y_{im} = r \mid L_i = c, g_i = g), \qquad (1)$$

That is, the probability of an individual i, belonging to the class c and the group g, choose the answer r of the item m. As these probabilities are, these parameters satisfy the constraint $\sum_{r=1}^{r_m} \rho_{m,r|c,g}$ = 1 for each combination of $m = 1,...M, c = 1,...,C$ and $g = 1,...,G$. Assuming that the values ρ's are equal between groups, a condition known as measurement invariance, results in the number of free measurement parameters given by $C\sum_{m=1}^{M}(r_m -$ 1). Otherwise, the number of free measurement parameters is $GC\sum_{m=1}^{M}(r_m - 1).$

The prevalence or probability of an individual *i* of the group *g* belonging to a certain class *c* is given by:

$$\gamma_{c|g} = Pr(L_i = c \mid g_i = g), \qquad (2)$$

subject to restrictions $\sum_{c=1}^{C}\gamma_{c|g}$ = 1 for *g* = 1,2*,...,G.*. If prevalences are considered equal between groups, then the number of free parameters of this type is (*C* −1). For *γ*'s heterogeneos between groups, the number of free parameters is *G* (*C* – 1).

3. Results

The use of latent class analysis (ACL) as a method of analysis leads to a characteristic of the severity of depression, linking probabilities to the answers contained in a given research instrument, such as PHQ-9. The problem in grouping variables or labeling a particular class that can be defined by the number of symptoms rather than the severity generates many misunderstandings regarding the efficiency of ACL in identifying patterns. When considering a population-based survey it is expected that there will be heterogeneity between and within the various segments that make up the sample space.

In order to examine the response patterns of individuals responding to survey for their own perception of depressive symptoms, latent class models will be adjusted and distinct groups (latent classes) will be identified based on the patterns of responses observed in categorical variables. All based on a probabilistic model, with the ability to identify characteristics that indicate the groups well, to estimate the prevalence of each group and to classify each individual within the groups.

Given the complexity and size of the NHP and taking into account regional diversities as regards economy, development, human development index, supply and access to health services among so many other follow-ups, it is necessary to consider possible differences that may exist between the regions regarding the problem addressed here. In this sense, separate analyzes were performed for each unit of the federation and, thus, a latent class model with covariates was adjusted for each one. However, for presentation the results were considered in the first stage for five states, one in each region of the country and two as a reference.

The dataset of the National Health Survey, composed of a sample of size n = 60; 202, was carried out preliminary descriptive analyzes in order to know general characteristics of the population as well as those that may be associated with the specific research problem, or the profile of response to issues associated with depression and their respective disorders.

In general, the sample presents a certain balance of the percentage relation between the sexes, and the

female category corresponds to 56:95% of the sample. With respect to age, the most expressive categories are related to the initial classes, where approximately 66% of the sample corresponds to the age group between 18 and 50 years. It is noteworthy that initially the variable age was distributed in 6 classes ([18; 30]; [30 : 40]; [40; 50); [60 : 70]; [70+]), however, evaluating the results of the preliminary modeling we verified the possibility of groupings in 4 classes. The schooling variable was subdivided into 4 classes: (Without Instruction or Fundamental Incomplete), (Full Core or Incomplete Medium), (Full or Incomplete Upper) and (Superior). More than half of respondents (57:34%) reported having a partner while only 14:02% responded living alone. The most prevalent breeds in the sample correspond to brown and white, which are respectively 49:03% and 40:04%.

The items related to the symptoms responsible for identifying the individual's profile for a possible depressive picture were measured in terms of the scores considered and the percentages of the responses. The scores from 0 to 3 correspond to the frequencies of symptoms observed at 0 (no day), 1 (less than half of the days), 2 (more than half of the days) and 3 (almost every day), respectively, in the last two study reference weeks. This scale is based on the The Patient Health Questionnaire (PHQ-9) which is a self-administered version of the PRIME-MD diagnostic tool for common mental disorders. The PHQ-9 is the depression module, which classifies each of the 9 criteria according to the symptom's itemity.

To fit a latent class model, the researcher encounters some great challenges such as the choice of latent classes, the decision on the constraints that must be imposed on the rho0s parameters as well as on the gamma's parameters, in case there are natural grouping in the population of interest, besides the selection of covariables.

Brazil is a country of great extension and with that it was considered the approach of the fit a global model with grouping given by the states of Brazil but incorporating the constraints that the parameters ρ's and α's do not present variability between the states.

Estimates of odds ratios and confidence intervals at 95% for the effect of each covariant included in the model. Except for the variable \Living alone", all covariates had significant effects on the prevalence of the classes. The variable \Sex" signalled that women are more likely to belong to classes that point to experiences of symptoms greater than men, especially in classes 3 and 4 \frequent or constant risk. of individuals without formal or incomplete basic education, in contrast to those of more advanced educational levels, belong to classes 3 and 4. The Age variable showed a higher probability of individuals belonging to the higher groups, above 30 years, in contrast to those from 18 to 29, belong to class 4. The fact of not having a mate also contributes to an increase in the chance of being classified in class 4.

4. Discussion and Conclusion

LCA provides a way to identify underlying subgroups characterized by multiple dimensions which in turn can be used to examine the effects of other covariates. Since depression is a complex and diffcult to measure disease, because of this configuration, the ACL represents a promising methodology in the discussion of this phenomenon that is increasingly present in the population in the most varied contexts. With the National Health Survey data, the LCA allowed the formation of four classes based on in nine items delineated to evaluate symptoms, making it possible to draw profiles for the Brazilian population avoiding the use of cut-off points in the commonly adopted scores.

Exploration of separate adjustments for each states led to the formulation of a global model that took into account different prevalences in the four latent classes, but similarities regarding the profile of responses to items and effects of covariates. In this global analysis it was possible to identify with more evidence the covariables associated to the classes.

The confidence intervals (95%) for the odds ratios for a variable of class over 1 signaling that women are more likely to be classified in some classes that show a risk of experiencing depressive symptoms when contrasted with men. This finding agrees with studies that show that women are more prone to depression.

The variable educational level was also shown to have a significant effect, in which individuals without formal education or incomplete elementary education, in contrast to the more advanced educational levels, are more likely to be classified in higher risk classes.

The age variable indicates that individuals over 30 years of age are more likely to be classified in class 4. Those over 60 years of age are likely to be classified in classes 3 or 4. They argue in their LCA application, that the symptomatology of depression in the elderly may be related to socioeconomic, cultural and biological aspects. It is noteworthy that in the mentioned study only a part of the population was considered, which suggests a pattern in population terms, since in the results found in the present study we also found similar results regarding the classification of groups of higher ages in the classes said to be more serious. Of the covariates investigated only the binary \living alone" was not significant.

In view of the statistics and forecasts provided by the World Health Organization regarding the number of people living and who may develop some type of mental disorder in the next 10 years, the subject is presented as a contribution in the scientific environment to explore the possible factors related to depression through statistical methods that deal with the characteristics of a population that cannot be measured directly.

The results showed the influence of responses in which specific groups of age, sex, and schooling present a probability of specific choices regarding items characteristic to the symptoms in question. Depression and health behavior in Brazilian adults observed in National Health Survey (PNS-2013) were studied and the results showed the importance of assessing the presence of depression and the frequency and severity of symptoms. There is no definition in terms of factors that carry more or less weight in maintaining the disease, but suggestions about implementing actions to promote healthy behaviors are highlighted.

This note also contemplates the process of access to health services, an issue linked to the fact that the distribution of health, in terms of service and access, is not a random phenomenon in society. Therefore, it is necessary to consider the social and economic structures in a given region and the health conditions of those who live there. It is noteworthy that in the present study, due to the complexity involved in collecting income information, it was not possible to access the relevance of the same, possibly important in the problem.

The results corroborate the discussion of the complexity around delineating a profile of response to the heterogeneity around the biological problem and the population components that guide this research. The Ministry of Health of Brazil and the Brazilian Institute of Geography and Statistics (IBGE) began planning the next National Health Survey, scheduled to have the data collection started in the second half of 2019, thus providing perspectives regarding the knowledge about changes that may have occurred in group formation in terms of responses to items associated with mental health.

References

1. Collins, L. M., & Lanza, S. T. (2010). Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences (Vol. 718). John Wiley & Sons.
2. Dratcu, L., da Costa Ribeiro, L., & Calil, H. M. (1987). Depression assessment in Brazil: the first application of the Montgomery-˚Asberg depression rating scale. The British Journal of Psychiatry, 150(6), 797-800.
3. Lazarsfeld, Paul Felix, and Neil W. Henry (1968). Latent structure analysis. Houghton Mifflin Co.
4. Lohr, S. L. (2009). Sampling: design and analysis. Nelson Education.
5. Lumley, T. (2011). Complex surveys: a guide to analysis using R (Vol. 565). John Wiley & Sons.
6. Malta, D. C., & Szwarcwald, C. L. (2017). Population-based surveys and monitoring of noncommunicable diseases. Revista de saude publica, 51, 2s.
7. Oberski, D. (2014). lavaan. survey: An R package for complex survey analysis of structural equation models. Journal of Statistical Software, 57(1), 1-27.
8. Paykel, E. S. (2008). Basic concepts of depression. Dialogues in clinical neuroscience, 10(3), 279.
9. Sartorius, N., Ustu¨n, T. B., Lecrubier, Y., & Wittchen, H. U. (1996). Depression comorbid with anxiety:¨ results from the WHO study on psychological disorders in primary health care. The British journal of psychiatry, 168(S30), 38-43.
10. Spitzer, A. L., Kroenke, K., Spitzer, A. R., Williams, E. J., & Williams, A. R. (2001). The PHQ-9: validity of a brief depression severity measure. Journal of General Internal Medicine.
11. Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. Journal of general internal medicine, 16(9), 606-613.

Effects of time-lagged meteorological variables on attributable risk of leishmaniasis in central region of Afghanistan

Majeed A. Adegboye[1], Jamiu Olumoh[1], Timor Saffary[2], Faiz Elfaki[3], Oyelola A. Adegboye[4]

[1]American University of Nigeria, 640001 Yola, Nigeria

[2]Independent researcher, MD, United States

[3] Qatar University, 2713 Doha, Qatar.

[4]Ton Duc Thang University, Ho Chi Minh City, Vietnam

Abstract

Leishmaniasis remains one of the world's most neglected vector-borne diseases, affecting predominantly poor communities mainly in developing countries. This study used data from 3 major leishmaniasis provinces of Afghanistan to provide an empirical analysis of change in heat-, cold- and other meteorological- leishmaniasis association between 2003 and 2009. The counts of leishmaniasis were modelled via quasi-Poisson regression adjusting for seasonality, a long-term trend with environmental variables, elevation, humidity, minimum, maximum and mean temperatures, rainfall separately for each province. In this study, the non-linear and delayed exposure-lag-response relationship between climatic variables and leishmaniasis was fitted with a distributed lag non-linear model applying a spline function describing the dependency along with the range of values with a lag of up to 12 months. We estimated the risk of leishmaniasis attributable to high and low temperature. The subgroup analysis showed an increased risk for males as well as young and middle-aged people at cold temperatures, however, the higher risk was observed for the elderly in heat. The overall leishmaniasis-temperature attributable fractions were estimated to be 7.6% (95% CI: 7.5% – 7.7%) and mostly due to temperature-related health burden.

Keywords

Lagged variables; Leishmaniasis; Generalized additive model; Afghanistan; Attributable risk

1. Introduction

Leishmaniasis is still one of the world's most neglected vector-borne diseases, largely affecting poor communities [1-3] in developing countries [4]. Approximately, 350 million people are at risk of contracting leishmaniasis, and some 2 million new cases occur every year [5]. Leshmaniasis infection causes significant health burdens and can have negative social and psychological stigma which could lead to significant economic losses [1,6-7]. Leishmaniasis is a parasitic disease caused by Leishmania parasite, which is transmitted to human and animal hosts by the bite of phlebotomine sandflies [2, 8]. The geographical distribution of leishmaniasis is tied in to the abundance of sandflies, their life cycle, and the presence of reservoirs [3, 8-9].

Clinical manifestations of can cause Visceral Leishmaniasis (VL), also called Kala-azar which is the most serious form of the disease; Cutaneous Leishmaniasis (CL), which is the most common; and Muco-Cutaneous Leishmaniasis (MCL) or Diffuse-Cutaneous Leishmaniasis (DCL), which rarely occurs [3, 10-13]. The coexistence of these clinical forms in the same patient is very rare [5] . VL is potentially fatal [2,14], with a case fatality rate of about 10%. Jaundice, wasting severe anaemia and HIV co-infection are associated with increased risk of VL lethality [7, 15] while malaria co-infection increases the risk of CL in the endemic region [10]. Afghanistan has been plagued by leishmaniasis since the ninth century when it was called the "Balkh Sore" named after the Afghan province north of the capital Kabul [16-18]. The disease continued to expand with number of infections growing across Afghanistan as endemics became more prevalent [18-20]. The occurrence of leishmaniasis in Afghanistan varies widely from year to year, showing an upward trend and repeating patterns related to the months of the years, thereby indicating of nonstationarity and nonlinearity characteristics [10, 21].

The close association of Leshmaniasis distribution with climate and meteorological conditions may be used to predict Leshmaniasis epidemics in Afghanistan [10, 22-24]. However, there is no study on the burden of leishmaniasis attributable to time-varying meteorological variables to inform a more understanding of the endemic disease. This study is aimed to fill this gap by investigating the meteorological variables-leishmaniasis association among vulnerable groups and assess the health burden of leishmaniasis attributable to meteorological variables in Central Afghanistan.

2. Methodology

Study area and data source

Leshmaniasis is endemic in Afghanistan. The data used in this study were retrospective records of clinically-diagnosed leishmaniasis cases in Afghanistan between 2003 and 2009 obtained from the Afghanistan Health Management Information System (HMIS) under the National Malaria and leishmaniasis Control Programme (NMLCP) of the Ministry of Public Health (MoPH). Three neighbouring provinces (Kabul, Kapisa and Logar) in the central region with high incidence of leishmaniasis were considered in this study (Figure 1). The meteorological variables used in this study were mean land surface temperature (LST) and rainfall. Satellite-derived environmental-LST was obtained from the Moderate Resolution Imaging Spectroradiometer (MOD11 L2 version 6, USGS/Earth Resources Observation and Science (EROS)

Center, Sioux Falls, South Dakota) at 1 km spatial resolution while the monthly accumulated rainfall data measured by the Tropical Rainfall Measuring Mission (TRMM: TMPA/3B43) jointly conducted by NASA and the Japan Aerospace Exploration Agency (JAXA).

Statistical analysis

To assess the effect of climatic variables, we used a distributed lag non-linear model (DLNM) [31-34] with random intercept [35] to modelled counts of leishmaniasis (Yit) at month t in province i via Poisson regression adjusting for population, seasonality, long-term trend with meteorological variables-temperature (ºC) and rainfall (inches),

$$Y_{it} \sim Poisson(\mu_{it})$$

$$log(\mu_{it}) = \alpha_{0i} + \log(Population_{it}) + \sum^{12}{}_{l=0}\, s(x_{it-l}\beta_l) + \sum{}^{J}{}_{j=1}\, f_j(u_{it,j}\gamma_j) \qquad (1)$$

$$\alpha 0i \sim N(\alpha 0,\ \sigma 20)$$

where α_{0i} is a random intercept to capture provincial level dependencies, α_0 overall average intercept and $\sigma^2{}_0$ is the province-level variability around α_0; $Population_{it}$ represents the population of province i at month t; The function, f_j is used to specify the functional relationship between variables u_{itj} and the nonlinear exposure-response curve, defined by the parameter vectors γ_j. Natural cubic spline with 3 degrees of freedom was used to define smooth function, $f_{j1}(u_{it,1}\gamma_j)$ for rainfall.

The function, $s(x_{it-l}\beta_l)$ described the dependency along the range of exposure values and lag dimension (up to 12 months). Thus, we modelled the non-linear and delayed exposure-lag-response relationship between the temperature and leishmaniasis with a spline function. The cross-basis parameterization for the exposure-lag-response is given by:

$$s(x,t) = \int_{l_0=0}^{12} f \cdot w(x_{t-l}, l)dl \approx \sum_{l_0=0}^{12} f \cdot w(x_{t-l}, l) = w_{x,t}^T \eta \qquad (2)$$

The bi-dimensional function $f \cdot w(x_{t-l}, l)$ represents the *exposure–lag–response function*, and model simultaneously the exposure–response $f(x)$ curve along temperature range and lag–response curve, $\omega(l)$ [33].

The predictions for the cumulative exposure–lag–response association derived from the parameter estimates from the Poisson regression model (1) for varying meteorological values and lags were then displayed as exposure-lag-response curve of relative risk.

Attributable risk measure

The attributable fraction, which is an indicator of exposure-related health burden, was calculated by using the effect summaries from the DLNM models and treating the associations with exposures at different lags as independent contributions to the risk [33]. We defined the optimum exposure as the value of meteorological variable at which leishmaniasis risk is the lowest in the estimated exposure-response curve. With the optimum exposure value as reference, for each lag of the series, in a province, we used the overall cumulative relative risk corresponding to each lag's exposure to calculate the attributable number and fraction of attributable number in the next l-lags.

Model selection and model assessment

To capture the flexibility of the exposure-lag-response relationship, $f \cdot w(xt-l,l)$, we explored constant, linear and quadratic B-splines for the temperature-lag-leishamniasis functional relationship with a lag of up to 12 months. Additionally, we explored different choices for the lag number (3, 6, 9 and 12), the number of knots and position as well as the dfs for seasonality and long-term trends (2-8) and rainfall (2-8). The models were assessed with Akaike Information Criteria (AIC). All analyses were done using the package dlnm [32] in R 3.4.2 statistical software [36]. The empirical confidence intervals (eCIs) were obtained via Monte Carlo simulations.

3. Result

Summary of disease and meteorological variables

A total 67,942 cases of leishmaniasis were reported in the study region between 2003 and 2009. The baseline characteristics of infection are presented in Table 1. Kabul province accounted for 67.2% of the total cases, primarily in middle aged (15-59 years) population. The median monthly mean temperature was 16.1oC with an interquartile range between 7.1oC and 25.0oC. The relative risks (RRs) from the best fitted DLNM model described by a quadratic B-splines for a temperature-leishmaniasis relationship, linear function for lag-leishmaniasis, natural cubic splines for rainfall with 6 degrees of freedom (dfs) to capture seasonality and long-term trend were presented in Figure 3 and Table 3. The 3-D plot shows that the temperature-leishmaniasis association was nonlinear, immediate and persisted throughout the 12-month lag period. The effect was significantly higher at low temperatures and maximum at lag 0. Figure 3 and Table 3 show the overall effect, log Relative Risk (logRR) of temperature on the risk of leishmaniasis for up to 12 months' lag and at specific lags and temperature values. The result indicates that the disease was associated with temperature, was highest at lag 0 and declined over the entire lag periods but remained significant. The increased risk of leishmaniasis during the cold temperature was highest at moderate cold temperature. For example, at 2.16oC the cumulative risk (logRR) was 6.16 (95% CI: 5.74 – 6.58).

The leishmaniasis-temperature curve revealed (not shown) different optimum temperatures for the subgroup analysis, ranging from 21.4 oC for the elderly group (> 59 years) to 26.0oC for the middle age group (15-59 years). Similarly, the magnitude of the association between temperature and incidence of leishmaniasis varied slightly among different subgroups (Table 3). Generally increased risk of leishmaniasis during extreme cold and extreme heat among different subgroups was discovered. The increased risk of leishmaniasis during the cold period was higher for males than females at temperatures lower than 5oC. However, the temperature-leishmaniasis effect was protective during the heat period at lower lags but increased at longer lags (Table 3). There was an increased risk of cold weather among young and middle-aged people. In contrast, the cumulative risk for heat (6.99, 95% CI: 6.40 – 7.58) was higher for the elderly (60+ years) than for cold (1.97, 95% CI: 1.19 – 2.74)

Attributable risk of leishmaniasis

Table 4 presents the leishmaniasis fraction (%) attributable to non-optimum temperatures for the entire group as well as its subgroups. The attributable fractions (AF) of the disease are much higher for cold temperatures, especially for moderately cold temperatures (temperatures below the optimum temperature but above 2.5th percentile). The overall estimated AF was 7.6% (95% CI: 7.5% – 7.7%) and mostly due to cold. In the sub-group analysis, the total AF was lowest due to temperature in elderly people (3.7%, 95% CI: 1.4% – 5.0%) while the younger age group had the highest AF (7.5%, 95% CI: 7.1% – 7.7%). Generally, most of leishmaniasis-temperature AF can be attributed to moderate temperatures except in the case of the older group (2.5% (95% CI: 2.3% – 2.9%).

4. Discussion and Conclusion

Several previous studies have shed light on the role of environmental factors such as temperature, rainfall and altitude on Leishmaniasis and vector, sandflies [37- 43]. In fact, recently introduced forecasting models heavily depend on weather and climate data as predictors, an approach which has led to higher precision over longer time periods [8, 42, 44-45]. Because of this dependency on environmental factors, global warming is expected to continue shifting the geographical distribution of sandflies and leishmaniasis northward with the first phlebotomine sandflies already being sighted in previously leishmaniasis-free regions such as Germany and Belgium [46]. One previous study focuses on other risk factors for Anthropological CL (ACL) in Kabul at the household level including things such as sex, age, number of households, household construction materials, etc.[27].

Previous studies revealed that temperature had a significant effect on the development and spread of leishmaniasis[10,46-47]. In the present study, we examined the delay and nonlinear impact of temperature on clinically-diagnosed cases of leishmaniasis in three major affected Afghan provinces, Kabul, Kapisa and Logar, between 2003 and 2009, with Kabul having the highest CL cases world-wide. The almost four-decade-long civil war in Afghanistan has increased the number of leishmaniasis cases significantly to an estimated 67,000 annual cases nationwide [48]. Unlike most previous studies, this study applied distributed lag nonlinear model with a random effect to capture the spatial proximity of the three provinces.

Whereas temperature and rain fall were the environmental variables, the temperature was the main dependable variable in this study. Following the first visualization of seasonal patterns, Mann-Kendall trend and chi-square tests were conducted to analyze the monthly trends and potential differences between monthly Leishmaniasis cases of different subgroups. The impact of climatic factors on leishmaniasis was analyzed with a distributed lag non-linear model and Poisson regression. The exposure-lag-relationship was examined with constant, linear and quadratic B-splinesfor the temperature-lag-leishmaniasis relationship. Our study findings and key implications are summarized below.

The first outcome of this study is that the impact of temperature on leishmaniasis is observed immediately and could persist throughout the entire year,indicating a crucial temperature dependency of the sandfly. Thus, the aforementioned studies were correct to incorporate climate dependencies in their models for forecasting leishmaniasis. The relatively high accuracy of twelve-month CL prediction of 72%-77%, which is significantly higher than models with no climate predictors, justifies this approach as well [44].

The leishmaniasis-temperature association in this study has different optimums for the subgroups, with the relative risk being lowest for the older subgroup (above 59 years) at 21.4oC and highest for the younger group (between 15 and 59 years) at26.0oC. Our results also show that the relative risk during the cold period was higher for males than females which is consistent with previous findings [48].

Finally, in addition to the parasite favourable ecological and environmental parameters in parts of Afghanistan, other factors such as armed conflict, movement of large parts of the population, destructed infrastructure, very low living standards, and an insufficient health-care system have contributed to the fast spread of the disease[48-50].

This study on documented leishmaniasis cases in three Afghan provinces has confirmed the importance of environmental and climatic factors for the spread of the disease, in particular, its dependenceon temperature. The differentiation between different subgroups of affected people led to a deeper insight into the infection process and could contribute to new and more precise forecasting models as well as assist in formulating new preventive initiatives.

References

1. Kassi M, Afghan AK, Rehman R, Kasi PM. 2008. Marring leishmaniasis: The stigmatization and the impact of cutaneous leishmaniasis in pakistan and afghanistan. PLoS neglected tropical diseases 2:e259.
2. Leslie T, Saleheen S, Sami M, Mayan I, Mahboob N, Fiekert K, et al. 2006. Visceral leishmaniasis in afghanistan. CMAJ 175:245-246.
3. Mubayi A, Castillo-Chavez C, Chowell G, Kribs-Zaleta C, Ali Siddiqui N, Kumar N, et al. 2010. Transmission dynamics and underreporting of kala-azar in the indian state of bihar. Journal of Theoretical Biology 262:177-185.
4. Bailey F, Mondragon-Shem K, Hotez P, Ruiz-Postigo JA, Al-Salem W, Acosta-Serrano A, et al. 2017. A new perspective on cutaneous leishmaniasis-implications for global prevalence and burden of disease estimates. PLoS neglected tropical diseases 11:e0005739.
5. Gradoni L, López-Vélez R, Mokni M. 2017. Manual on case management and surveillance of the leishmaniases in the who European region. (Copenhagen: World Health Organization Regional Office for Europe).
6. Adegboye O, Kotze D. 2012. Disease mapping of leishmaniasis outbreak in Afghanistan: Spatial hierarchical bayesian analysis. Asian Pacific Journal of Tropical Disease 2:253-259.
7. Bern C, Maguire JH, Alvar J. 2008. Complexities of assessing the disease burden attributable to leishmaniasis. PLoS neglected tropical diseases 2:e313.
8. Erguler K, Pontiki I, Zittis G, Proestos Y, Christodoulou V, Tsirigotakis N, et al. 2019. A climate-driven and field data-assimilated population dynamics model of sand flies. Scientific reports 9:2469.
9. Koch LK, Kochmann J, Klimpel S, Cunze S. 2017. Modeling the climatic suitability of leishmaniasis vector species in europe. Scientific reports 7:13325.
10. Adegboye O, Adegboye M. 2017. Spatially correlated time series and ecological niche analysis of cutaneous leishmaniasis in Afghanistan. International Journal of Environmental Research and Public Health 14:309.
11. Bates PA. 2007. Transmission of Leishmania metacyclic promastigotes by phlebotomine sand flies. International Journal for Parasitology 37:1097-1106.
12. Casolari C, Guaraldi G, Pecorari M, Tamassia G, Cappi C, Fabio G, et al. 2005. A rare case of localized mucosal leishmaniasis due to leishmania

infantum in an immunocompetent italian host. European journal of epidemiology 20:559-561.

13. Salah AB, Kamarianakis Y, Chlif S, Alaya NB, Prastacos P. 2007. Zoonotic cutaneous leishmaniasis in central tunisia: Spatio-temporal dynamics. International Journal of Epidemiology 36:991-1000.
14. Ready PD. 2014. Epidemiology of visceral leishmaniasis. Clinical Epidemiology 6:147-154.
15. Assunção RM, Reis IA, Oliveira CDL. 2001. Diffusion and prediction of leishmaniasis in a large metropolitan area in brazil with a bayesian space–time model. Statistics in Medicine 20:2319-2335.
16. Reyburn H, Rowland M, Mohsen M, Khan B, Davies C. 2003. The prolonged epidemic of anthroponotic cutaneous leishmaniasis in kabul, afghanistan:'Bringing down the neighbourhood'. Transactions of the Royal Society of Tropical Medicine and Hygiene 97:170-176
17. Hepburn NC. 2003. Cutaneous leishmaniasis: An overview. Journal of postgraduate medicine 49:50.
18. Stewart CC, Brieger WR. 2009. Community views on cutaneous leishmaniasis in istalif, afghanistan: Implications for treatment and prevention. International quarterly of community health education 29:123-142.
19. Ashford R, Kohestany K, Karimzad M. 1992. Cutaneous leishmaniasis in Kabul: Observations on a 'prolonged epidemic'. Annals of Tropical Medicine & Parasitology 86:361-371.
20. Omar A, Saboor A, Amin F, Sery V. 1969. Preliminary study on the foci of cutaneous leishmaniasis in kabul city. Zeitschrift fur Tropenmedizin und Parasitologie 20.
21. Elnaiem D-EA, Schorscher J, Bendall A, Obsomer V, Osman ME, Mekkawi AM, et al. 2003. Risk mapping of visceral leishmaniasis: The role of local variation in rainfall and altitude on the presence and incidence of kala-azar in eastern Sudan. American Journal of Tropical Medicine and Hygiene 68:10-17.
22. Galgamuwa LS, Dharmaratne SD, Iddawela D. 2018. Leishmaniasis in sri lanka: Spatial distribution and seasonal variations from 2009 to 2016. Parasite & vectors 11:60.
23. Plourde M, Coelho A, Keynan Y, Larios OE, Ndao M, Ruest A, et al. 2012. Genetic polymorphisms and drug susceptibility in four isolates of leishmania tropica obtained from canadian soldiers returning from afghanistan. PLoS neglected tropical diseases 6:e1463.
24. Reithinger R, Coleman PG. 2007. Treating cutaneous leishmaniasis patients in kabul, afghanistan: Cost-effectiveness of an operational program in a complex emergency setting. BMC Infect Dis 7:3-3.

25. Gasparrini A, Armstrong B, Kenward MG. 2010. Distributed lag non-linear models. Statistics in medicine 29:2224-2234.
26. Gasparrini A. 2011. Distributed lag linear and non-linear models in r: The package dlnm. Journal of statistical software 43:1-20.
27. Gasparrini A. 2014. Modeling exposure–lag–response associations with distributed lag non-linear models. Statistics in medicine 33:881-899.
28. Wu Y, Qiao Z, Wang N, Yu H, Feng Z, Li X, et al. 2017. Describing interaction effect between lagged rainfalls on malaria: An epidemiological study in south–west china. Malaria journal 16:53.
29. R Core Team. 2017. R: A language and environment for statistical computing. R foundation for statistical computing, vienna, austria.
30. Adegboye O, Al-Saghir M, Leung D. 2017. Joint spatial time-series epidemiological analysis of malaria and cutaneous leishmaniasis infection. Epidemiology and Infection 145:685-700.
31. Bhunia GS, Kesari S, Jeyaram A, Kumar V, Das P. 2010. Influence of topography on the endemicity of kala-azar: A study based on remote sensing and geographical information system. Geospatial health 4:155-165.
32. Cardenas R, Sandoval CM, Rodriguez-Morales AJ, Franco-Paredes C. 2006. Impact of climate variability in the occurrence of leishmaniasis in northeastern Colombia. The American journal of tropical medicine and hygiene 75:273-277.
33. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. 2016. Extracting information from the text of electronic medical records to improve case detection: A systematic review. J Am Med Inform Assoc 23:1007-1015.
34. Kassem HA, Siri J, Kamal HA, Wilson ML. 2012. Environmental factors underlying spatial patterns of sand flies (diptera: Psychodidae) associated with leishmaniasis in southern sinai, egypt. Acta tropica 123:8-15
35. Lewnard JA, Jirmanus L, Júnior NN, Machado PR, Glesby MJ, Ko AI, et al. 2014. Forecasting temporal dynamics of cutaneous leishmaniasis in northeast brazil. PLoS neglected tropical diseases 8:e3283.
36. Toumi A, Chlif S, Bettaieb J, Alaya NB, Boukthir A, Ahmadi ZE, et al. 2012. Temporal dynamics and impact of climate factors on the incidence of zoonotic cutaneous leishmaniasis in central tunisia. PLoS neglected tropical diseases 6:e1633.
37. Chaves LF, Pascual M. 2006. Climate cycles and forecasts of cutaneous leishmaniasis, a nonstationary vector-borne disease. PLoS medicine 3:e295.
38. Talmoudi K, Bellali H, Ben-Alaya N, Saez M, Malouche D, Chahed MK. 2017. Modeling zoonotic cutaneous leishmaniasis incidence in central

tunisia from 2009-2015: Forecasting models using climate variables as predictors. PLoS neglected tropical diseases 11:e0005844.

39. Chalghaf B, Chemkhi J, Mayala B, Harrabi M, Benie GB, Michael E, et al. 2018. Ecological niche modelling predicting the potential distribution of Leishmania vectors in the Mediterranean basin: Impact of climate change. Parasites & vectors 11:461.

The evolution of environment statistics in Department of Statistics Malaysia

Zaitun Mohd Taha @ Abd. Rahman, Siti Zakiah Muhamad Isa
Department of Statistics (DOSM), Malaysia

Abstract

Department of Statistics Malaysia (DOSM) is the national institution which is responsible in collecting, compiling and disseminating social, economic and demographic statistics. Environment statistics in DOSM is an emerging statistical field in official statistics. Environment is multi-disciplinary and multi-dimensional topics, thus the scope and coverage of the environment statistics is also wide. The environment statistics have become even more prominent in the national statistical system after the introduction of the concept of sustainable development. This paper presents an overview of the environment statistics in DOSM and its evolution in support of sustainable development.

Keywords

Environment statistics; environmental concerns; sustainable development

1. Introduction

The wake of public awareness on environment was in 1972, when the United Nations Conference on the Human Environment took place in Stockholm to discuss the state of the global environment. Stockholm Declaration includes 26 principles concerning the environment and development; among others are natural resources must be safeguarded; non-renewable resources must be shared and not exhausted; and damaging oceanic pollution must be prevented. In relation to this, the concept of sustainability as the basis for an integrated approach to economic and environmental policies was introduced in Brundtland Commission Report, Our Common Future by the United Nations in 1987.

The concept of sustainable development has made society conscientious and, in the same time, recognize, the role and importance of environmental factors as well as of the functions and the services the environment provides (Bortelmus, 1986). Furthermore, the adoption of Sustainable Development Goals (SDG) in 2015, significantly increase the demand on environment statistics to inform and monitor on green economy issues and policies to reflect the current situation.

Malaysia as many other developing countries faces conflict between economic growth and conservation of environment. Hence, the concern on environmental issue and the principle of sustainable development has been embedded in its five years plan starting from the Fifth Malaysia Plan 1985–1990. In the Eleventh Malaysia Plan 2016-2020, one of the key thrusts is pursuing green growth for sustainability and resilience. It is Malaysia's commitment to pursue development in a more sustainable manner from the start, rather than a more conventional and costly model of 'grow first, clean up later'. A reinforced commitment to green growth will ensure that Malaysia's precious environment and natural endowment are conserved and protected for present and future generations.

In the light of this, statistics and indicators on the environment are required to meet the increasing demand for high quality information at the national, regional and international levels.

2. Methodology

Environment is multi-disciplinary and multi-dimensional topics, thus the scope and coverage of the environment statistics is also wide. Given the cross-cutting nature of environment statistics, the production of environmental data and statistics involves national statistical office, environmental ministries and sectoral authorities, which means that different agencies collect the information they need to specifically inform their priorities and goals.

The compilation of environment statistics and indicators in DOSM are in line with the concepts and guidelines outlined by the United Nations and other relevant international organizations. Data used for the production of environment statistics in DOSM are based on primary and secondary data. DOSM has also incorporated environment-related questions (modules) into economic census primarily intended to fulfill the needs to produce integrated environment and economic statistics.

3. Result

Over the last decades, DOSM has compiled Compendium of Environment Statistics and Environmental Protection Expenditure Statistics. Currently, the production of new statistical products on environment statistics have been published including Statistics on Water Supply, Sewerage, Waste Management and Remediation Activities; Green Economy Indicators and System of Environmental-Economic Accounting. This is as a response to the growing needs to develop and combine statistics and indicators such as green economy that are more inclusive of environmental and social aspects in order to cover the full realm of sustainable development.

3.1 Compendium of Environment Statistics (CES)

The pioneering work on environment statistics in DOSM began when Malaysia was selected as one of developing countries which receive regional technical assistance (RETA 5555), project on the Institutional Strengthening and Collection of Environment Statistics initiated by The Asian Development Bank in 1995. DOSM was appointed as the implementing agency for this activity. Malaysia embarked on this project to develop the Compendium of Environment Statistics (CES) which was based on Framework for the Development of Environment Statistics (FDES) 1984 published by the United Nations. This framework was modified according to Malaysia's situation known as Framework for The Development of Environment Statistics (FDES), Malaysia 1998. Prior to this, inter-agency committee which consist of representatives from various agencies related to environment statistics was established in 1997 to ensure the use and sustainability of CES.

DOSM has compiled the CES annually since 1998. This publication provides information on the interrelationships between human activities, environmental ecosystems and their environmental, social and economic impacts and the social responses to mitigate these impacts. The compilation of statistics are according to four environmental media classifications namely Air/ Atmosphere; Water/ Aquatic Environment (inland and marine); Land/ Terrestrial Environment and Urban Environment/ Human Settlements. The analyses are based on the Pressure-State-Response (PSR) model developed by the Organisation for Economic Co-operation and Development (OECD).

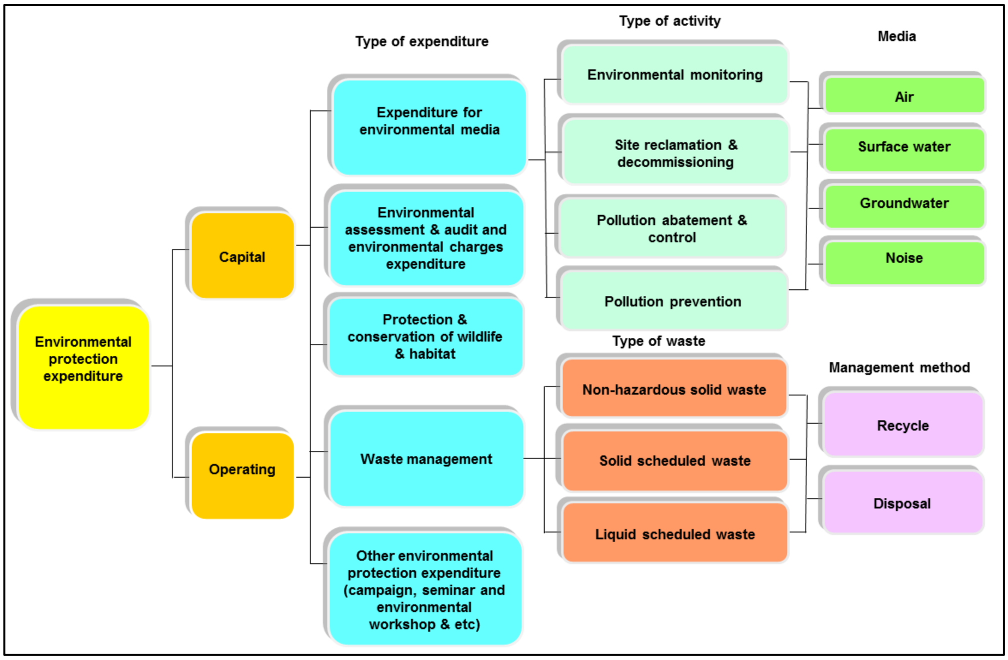
Following the endorsement of the FDES 2013 by the United Nations Statistical Commission at its 44th session (2013), DOSM has compiled environment statistics which apply the FDES 2013 in CES 2018. Thus, environment domain is expanded on statistics informing about extreme event & disasters and environmental protection expenditure in which the analyses are based on the Driving Force-Pressure-State-Impact- Response (DPSIR) model. This model provides a very effective representation of the environmental/economic interaction circuit in a sustainability perspective.

3.2 Statistics on Environmental Protection Expenditure

A pilot study on Environmental Protection Expenditure was carried out in 2004. Prior to this survey, there was no known information that exists on the cost to industry of pollution prevention and abatement programmes, environment management systems and environmental assessment. The study was carried out as a result of the decision and proposal in the workshop in finalizing the FDES Malaysia and CES in 1998. Initially, the study covers expenditures made by local authority and primary industries (minerals, petroleum refineries and natural gas), transport equipment and manufacturing industries.

Subsequently, statistics on the environmental protection expenditure was generated from the Survey of Environmental Protection Expenditure which was conducted annually since 2008. However, no survey was conducted in 2011 and 2016 due to Economic Census. Nevertheless, information on environmental protection expenditure was incorporated in Economic Census under Environmental Protection Compliance Module. Economic Census was conducted once in every 5 years. The coverage of the sector has been extended from time to time and at present, it covers five main sectors namely agriculture, mining & quarrying, manufacturing, construction and services. Statistics published includes environmental protection expenditure covering capital and operating expenditure by activity, media and type of expenditure as depicted in Figure 1.

Figure 1 : Environmental Protection Expenditure Statistics

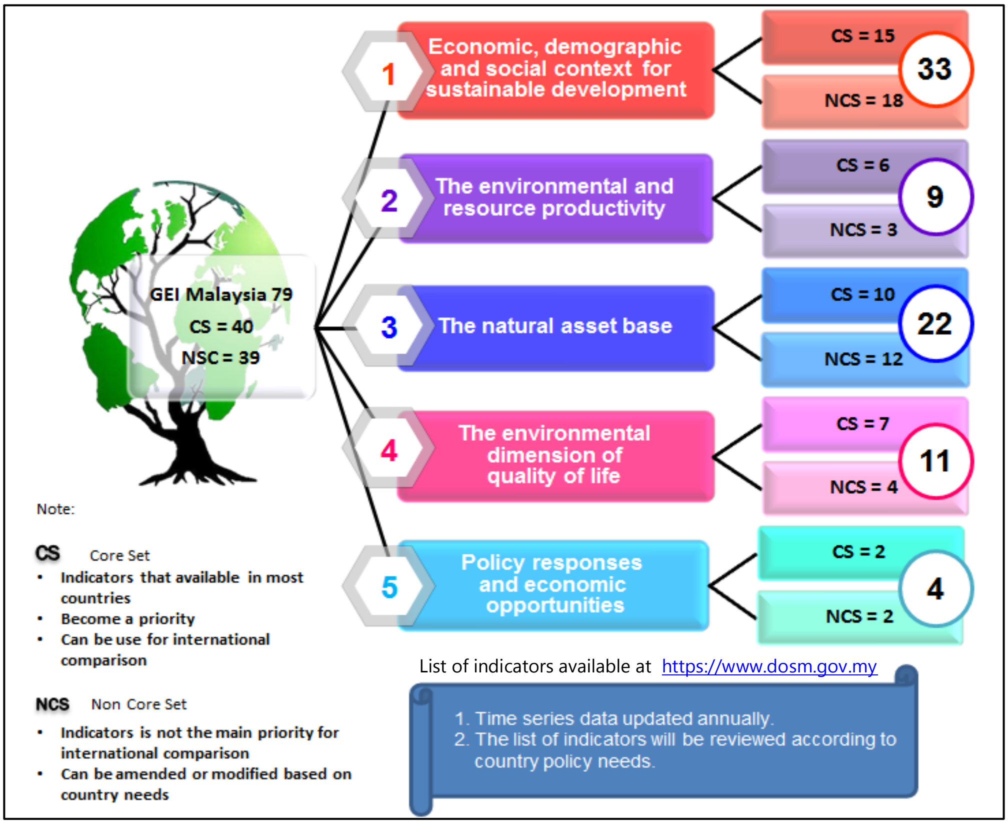

Environmental protection expenditure
Capital
Operating
Type of expenditure
Expenditure for environmental media
Environmental assessment & audit and environmental charges expenditure
Protection & conservation of wildlife & habitat
Waste management
Other environmental protection expenditure (campaign, seminar and environmental workshop & etc)
Type of activity
Environmental monitoring
Site reclamation & decommissioning
Pollution abatement & control
Pollution prevention
Media
Air
Surface water
Groundwater
Noise
Type of waste
Non-hazardous solid waste
Solid scheduled waste
Liquid scheduled waste
Management method
Recycle
Disposal

3.3 Green economy indicators

Malaysia has benefited through the development account project "Supporting Developing Countries Measure Progress towards Achieving a Green Economy" by United Nations Statistics Division (UNSD) for the period of 2014-2015. A first set of Malaysia Green Economy Indicators (GEI) was identified and published in 2015 as a result from this project.

A set of 79 indicators structured along five themes and categorised into Core Set and Non Core Set as listed in Figure 2. The set of indicators is mainly presented in the form of time series. The criteria used for selecting indicators are an indicator should be informative and relevant in terms of sustainability. The data should be readily available in official statistical datasets and, if possible and appropriate, be annual data covering a long time period. These indicators will be reviewed from time to time according to country policy needs and new indicators will also be developed in the coming years.

Figure 2 : Theme and Category of Malaysia Green Economy Indicators


GEI Malaysia 79
CS = 40
NSC = 39
1
Economic, demographic and social context for sustainable development
CS = 15
NCS = 18
33
2
The environmental and resource productivity
CS = 6
NCS = 3
9
3
The natural asset base
CS = 10
NCS = 12
22
4
The environmental dimension of quality of life
CS = 7
NCS = 4
11
5
Policy responses and economic opportunities
CS = 2
NCS = 2
4

Note:

CS Core Set

- Indicators that available in most countries
- Become a priority
- Can be use for international comparison

NCS Non Core Set

- Indicators is not the main priority for international comparison
- Can be amended or modified based on country needs

List of indicators available at https://www.dosm.gov.my

1. Time series data updated annually.
2. The list of indicators will be reviewed according to country policy needs.

The report of green economy indicators will be published biennially taken into account the elements on measurement of green development performance in Malaysia. On that note, the governance was set up consists of steering committee and two technical working groups (TWG) i.e. TWG for data compilation led by DOSM and TWG to determine the methods of measurement and performance of Malaysia's green development led by Ministry of Economic Affairs. Hence, the indicators of Green Economy and its underlying basic statistics can serve as a useful tool to monitor the achievement of targets under Sustainable Development Goals.

3.4 System of Environmental-Economic Accounting

Environmental accounts have been in the air for decades. As early as 1970s, European countries like Canada, Denmark, Norway and Netherlands have extensive experience in the compilation of the accounts.

System of Environmental-Economic Accounting (SEEA) are useful planning tool because it is able to cater two major defects of conventional national accounting which neglect of scarcities of natural resources and the degradation of environmental quality. SEEA is a mechanism to examine the interaction between economy and environment to support environmental and resource policy consistent with economic growth. As such, SEEA facilitates data integration within environment statistics and with economic statistics.

The development of SEEA in Malaysia is at an infancy stage. Work on SEEA has been developed at DOSM in 2011, in response to parliament question on green GDP. Immediately after the endeavor of project GEI, DOSM gets the opportunity to cooperate with the UNSD on "Supporting Member States in Developing and Strengthening Environment Statistics and Integrated Environmental-Economic Accounting for Improved Monitoring of Sustainable Development" project in 2016. This project is a stepping stone towards strengthening national statistical capacities in compilation of environmental-economic accounts and supporting statistics by integrating environment and economic statistics and linking it to policy demand. The milestones from this project are Roadmap for SEEA Malaysia 2016-2020 and SEEA Water Account.

DOSM and UNSD had jointly launched the Roadmap for SEEA in 2017. This roadmap is a dynamic document that outlined policies/programmes/initiatives related to environment in Malaysia, governance structure, data requirements, proposed accounts and implementation strategies as well as critical success factors for SEEA implementation. The roadmap lays a foundation for the development of integrated environment-economic statistics which identified four potential SEEA accounts that could be developed in Malaysia namely energy, water, air emission (for energy use) and land accounts (agriculture). This roadmap will be revised and updated periodically as and when it is required to fulfil the need of policy makers and stakeholders.

At present, DOSM has developed SEEA Physical Supply and Use Table (PSUT) Energy, SEEA PSUT Water and SEEA PSUT Air Emissions (for energy use). Among statistics published include supply and use of energy products by sector, energy intensity, sources of water supply, water consumption by sector and greenhouse gas by economic units. Currently, DOSM is in the midst of developing SEEA land account and is expected to be completed in 2021. Committee on Planning and Development of Environment Statistics has been set up in 2017 to strengthen the network and coordination among agencies as well as to ensure continuous production of SEEA.

3.5 Statistics on water supply, sewerage, waste management and remediation activities

DOSM conducted economic census once in 5 years. The first census was initially conducted in the year 2001 covering the manufacturing and services sectors. The coverage of the sectors was expanded to agriculture, construction and mining & quarrying sectors since economic census 2005. A sub-sector in the services sector that is related to environment i.e. Water Supply, Sewerage, Waste Management & Remediation Activities Sector was first covered in economic census 2011 and annual economic survey in 2018.

DOSM took the initiative to publish the statistics on water supply, sewerage, waste management and remediation activities to meet the growing demand and the need for the compilation of SEEA and SDG. Among key statistics published are value of gross output, intermediate input, value added, number of persons engaged, salaries & wages and value of fixed assets.

4. Discussion and Conclusion

Environment statistics in DOSM is an emerging statistical field in official statistics. The role of DOSM is significantly crucial to adequately respond to the increasing demand for environmental information in the adoption of SDG and Agenda 21 primarily to inform and monitor on green economy issues and policies.

Given that environment currently is an important issues, the availability of the data is crucial. Furthermore, the environment terminologies and jargon used is totally different compared to other definition of statistics. Without the proper understanding of technical terms, it becomes very difficult to move ahead in this area of work. Some data may not be complete in terms of coverage and some may only available for a special study. Statisticians need to select and aggregate data, with a view to fitting them into the environment statistics framework. Formation of technical groups would certainly be very helpful to the statisticians.

Furthermore, the selection of Malaysia as one of pilot countries in compilation of CES, GEI and SEEA has helped DOSM to evolve environment statistics in the national statistical system. Since the importance of environment statistics is now widely recognized with an increasing demand from the policy makers, DOSM has included the production of environment statistics in its five years plan to produce and sustain this field of official statistics. Moving forward, DOSM will embark on the development of ocean accounts and natural capital accounting.

Thus, DOSM will continue with its ongoing effort to improve data collection/compilation on environment statistics to publish a comprehensive environment statistics. The usage of Geographical Information System (GIS) and remote sensing in environment statistics will be expanded since it provides high quality of data and very useful primarily in the production of disaster statistics and land use information.

There is a need to establish a one-stop centre that can provide access and linkage to all data in the country, including data that are housed in relevant agencies. Moving forward, DOSM will take the lead to establish an appropriate mechanism and coordinate with relevant ministries and agencies as stated in the Eleventh Malaysia Plan 2016-2020.

DOSM will continually provide a coherent set of environment statistics and indicators to support government with evidence-based policies and decision making to support sustainable development.

References

1. Asian Development Bank, (1999). Development of Environment Statistics in Developing Asian and Pacific Countries
2. United Nations, (1972). Report of the United Nations Conference on the Human Environment and Development
3. United Nations Statistics Division, (1987). Report of the World Commission on Environment and Development: Our Common Future
4. Organisation for Economic Co-operation and Development, (2004). Measuring sustainable development – integrated economic, environmental and social frameworks
5. Economic Planning Unit, (2015). Eleventh Malaysia Plan 2016-2020
6. Lucretia Dogaru (2013). The importance of environmental protection and sustainable development. Procedia - Social and Behavioral Sciences, 93, 1344 – 1348
7. Department of Statistics Malaysia, (2004). Report on Pilot Study Environmental Protection Expenditure 2003
8. Department of Statistics Malaysia, (2017). Laporan Projek Rintis Supporting Developing Countries Measure Progress Towards Achieving a Green Economy 2014-2015
9. Department of Statistics Malaysia, (2017). Economic Census 2016 – Water Supply; Sewerage, Waste Management and Remediation Activities
10. Department of Statistics Malaysia, (2017). Roadmap for System of Environmental-Economic Accounting 2016-2020
11. Department of Statistics Malaysia, (2018). Compendium of Environment Statistics 2018
12. Department of Statistics Malaysia, (2019). Report on the Survey of Environmental Protection Expenditure 2018

A comparison between annual maximum and partial duration series of high-flow at Langat River

Firdaus Mohamad Hamzah[1*], Hazrina Tajudin[1], Hafizan Juahir[2]
[1]Universiti Kebangsaan Malaysia
[2] Universiti Sultan Zainal Abidin (UniSZA)

Abstract

Most flood frequency analysis (FFA) conducted in Malaysia involves only a single peak for each year. However, the small and medium floods which happen more frequently is neglected from the analysis. Large database can ensure a more precise result in order to determine the best fit distribution for each study area. Despite the complexities in the implementation of partial duration series, it has the ability to provide a better flood estimation. This study employs a streamflow data recorded at Kajang station, Sungai Langat, Malaysia over a 36-year period spanning from 1978 to 2013. The optimal threshold value selected is 48.7 m^3/s, in which the dispersion index is stabilize at around 1, $DI = 1$. Generalized Extreme Value (GEV) distribution describes the annual maximum, whilst the Lognormal (LN3) and Generalized Pareto (GPA) distribution describes the partial duration series. Parameter estimation is made using L-moment method since it is the best method usually applied in hydrological phenomena. There is a slight difference between estimated streamflow magnitude when using GPA and LN3 for selected return period, while a considerable difference was observed when using annual maximum at a higher return period. The estimated magnitude is crucial since it provides a measurement parameter to analyse damage which correspond to specific flow during flooding event.

Keywords

Annual Maximum Series; Partial Duration Series; Statistical Distribution; Flood Frequency Analysis; Kajang Station.

1. Introduction

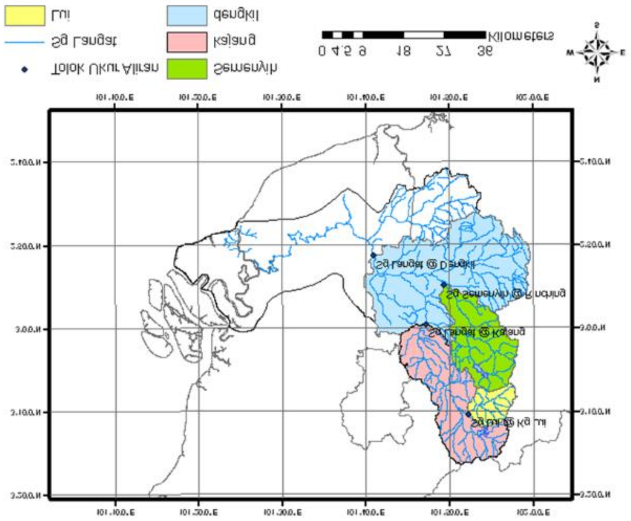
Most flood events in Malaysia are caused primarily by heavy rainfall that results in excess runoff which eventually exceeds river capacity. The short period of time series data available for the presents study is the main challenge in making the computations for hydrological analysis (Gado, Nguyen, & Asce, 2016). Most research consider extreme flood events instead of medium and frequent floods (Karim, Hasan, & Marvanek, 2017); (Keast & Ellison, 2013). This problem can be solved by taking into consideration the partial duration series in preference to the annual maximum series data. From a hydrological perspective flood frequency analysis is an important tool used to estimate future flood events based on the historical data of streamflow events. Result of the analysis is presented in terms of frequency and magnitude of flood events (Keast & Ellison, 2013). The performance of PDS and annual maximum series is dependent on the shape parameter. PDS is preferable for negative shape parameter while the AMS is more effective for positive shape parameters (Madsen, Rasmussen, & Rosbjerg, 1997).

The two types of daily streamflow data are daily mean and daily instantaneous maximum streamflow data (Karim et al., 2017). Most studies prefer to employ the daily mean data since it is able to give a clear description of the magnitude-frequency relationship for most type of river which operate based on a daily time step. When considering daily instantaneous maximum to generate a flood series data in preference to daily mean data, the extreme cases recorded are typically higher compared to a certain period of the day. For example, it is quite unrealistic to consider multiple data in a day when extracting a POT data series given that the area is flooded the whole day. Probability distribution is a statistical tool for describing the characteristics of a data structure. It is frequently employed to predict and estimate flood events (Garba, Ismail, & Tsoho, 2013). Prediction of most hydrological events, such as rainfall, streamflow, temperature, wind speed, etc., can be made based on occurrence probability since it involves statistical quantities. PDS offers more peaks to be considered in an analysis in comparison to when using annual maximum series (AMS) (Claps & Laio, 2003).

High intensity rainfall often triggers flooding in most areas (Franchini, Galeati, & Lolli, 2005). Rapid industrialisation and urbanization has led to deforestation and uncontrolled land use and, in many areas, this has altered the relationship between rainfall and flooding events (Tekolla, 2010). Flood events can cause loss of life and severe properties damage (Syed Hussain & Ismail, 2013). Thus, it is important to determine flood frequency in an effort to minimize the effects of flood events in Malaysia. The objective of this paper is to explore the effect of using different threshold values in the peak over threshold method in an attempt to determine the proper distribution which best fit the data for Kajang station and calculate the return period for the best fit distribution for 5, 10, 20, 50 and 100 years.

2. Methodology

The site chosen for this study is the Kajang station in the Sungai Langat sub basin. The station’s ID is 2917401 and it is located at latitude 020 59’ 34” and longitude 1010 47’ 13” (Figure 1). The duration of the data series is approximately 40 years of daily data spanning from 1978 to 2017. The data is measured in cubic meter per second (m3/s). There is no missing data in the dataset provided by the Department of Irrigation and Drainage (DID) Malaysia.



PDS does not exclude any significant high flow events, even if it is not the highest event of the year. As such, this method ensures a better representation of the sampling procedure of extreme values. There are two factors which cause difficulty in implementing this series: peak independence and determination of threshold level. Selecting a higher threshold value in a series would results in a smaller number of events being included in the series and this leads to a loss of valuable information. It often increases the likelihood of independence of each peak.

Then, it is important to determine the proper statistical distribution function that is able to describe the time series data. GEV distribution is frequently used for AMS data (Jiang & Kang, 2019). Generalized Pareto Distribution, also known as GPD, was introduced by Pikands in 1975. It is often implemented in peak over threshold data set (Jiang & Kang, 2019); (Gharib, Davies, Goss, & Faramarzi, 2017). Table 1 shows the probability density function and cumulative distribution functions.

Table 2: Probability Density Function and Cumulative Distribution Function for Each Distribution

Distribution	**Probability density function (pdf)**	**Cumulative distribution function (cdf)**
Gumbel	$f(x) = \frac{1}{\sigma} e^{(-z-e^{-z})}$ where $z = \frac{x-\mu}{\sigma}$	$F(x) = 1 - e^{-\left(\frac{x-\zeta}{\beta}\right)^{\delta}}$
Generalized Extreme Value	$f(x) = \frac{1}{\sigma} e^{(-(1+kz)^{-\frac{1}{k}})(1+kz)^{1-\frac{1}{k}}}$	$F(x) = e^{(-(1+kz)^{-\frac{1}{k}})}$ where $y = -k^{-1} log(1 - \frac{k(x-\xi)}{\alpha})$

Lognormal (3P)	$f(x)=\frac{e^{(-\frac{1}{2}(\frac{\ln(x-\gamma)-\mu}{\sigma})^2)}}{(x-\gamma)\sigma\sqrt{2\pi}}$	$F(x)=\phi(y)$ where $y=\frac{(\log(x-\zeta)-\mu)}{\sigma}$
Generalized Logistics	$f(x)=\frac{(1+kz)^{-1-1/k}}{\sigma(1+(1+kz)^{-\frac{1}{k}})^2}$	$F(x)=\frac{1}{1+(1+kz)^{-1/k}}$
Generalized Pareto	$f(x)=\frac{1}{\sigma}(1+k(\frac{x-\mu}{\sigma})^{-1-\frac{1}{k}}$	$F(x)=1-(1+k(\frac{x-\mu}{\sigma})^{-\frac{1}{k}})$

Evaluation method performance is dependent on sample size and skewness of the data. MLE is able to give the best parameter value compared to other methods. It maximizes the likelihood or joint probability of occurrence of the observed sample. However, MLE is not suitable for implementation in a small sample size. MLE estimator does not exist for a shape parameter of -1. L-moment has theoretical advantages over conventional moments in that it is able to characterize a broader range of distribution and is less affected by bias (Bílková, 2014),(Schlögl & Laaha, 2017). Hydrological parameters usually contain outliers. ML is said to be robust and does not affected much by sampling variability(Murthy, Jyothy, & Mallikarjuna, 2017). ML moments can be defined as follows:

$$l_1=\beta_0$$
$$l_2=2\beta_1-\beta_0$$
$$l_3=6\beta_2-6\beta_1+\beta_0$$
$$l_4=20\beta_3-30\beta_2+12\beta_1-\beta_0$$

where $\beta_r(r=0,1,2,3)$represents probability weighted moments such that:

$$\beta_r=n^{-1}\sum_{i=r+1}^{n}\binom{j-1}{r}\binom{n-1}{r}^{-1}*X(j,n)\,,r=0,n-1$$

Each moment represents location, dispersion, symmetry and peakedness of a data series. Based on the calculated moments, ML ratios can be established using the following calculation for coefficient of variation (CoV), skewness and kurtosis.

$$\tau_2=\frac{l_2}{l_1}$$
$$\tau_3=\frac{l_3}{l_2}$$
$$\tau_4=\frac{l_4}{l_2}$$

3. Results

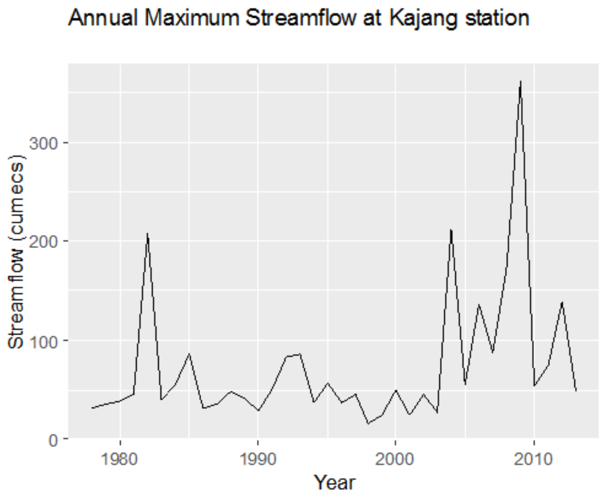


Figure 7: Annual Maximum Streamflow at Kajang Station

Table 3: Characteristics of PDS samples at Kajang station

Percentile (%)	90	91.5	93	94.5	96	97.5	98	98.5
Threshold (m^3/s)	18.6	20.2	22.6	25.6	30.6	39.2	43.6	48.7
Sample size	380	337	291	249	187	116	91	66
Rejected Peaks	935	781	629	474	339	213	172	131
λ	10.56	9.36	8.08	6.92	5.19	3.22	2.53	1.83
Mean (m^3/s)	36.2	38.93	42.02	45.94	52.27	63.79	70.03	79.12
Standard Deviation (m^3/s)	28.55	29.72	30.97	32.35	35.18	40.84	44.14	48.91
Skewness	5.87	5.65	5.51	5.33	5.02	4.37	4.03	3.58
Kurtosis	51.13	46.90	43.73	40.2	34.28	24.65	20.4	15.53

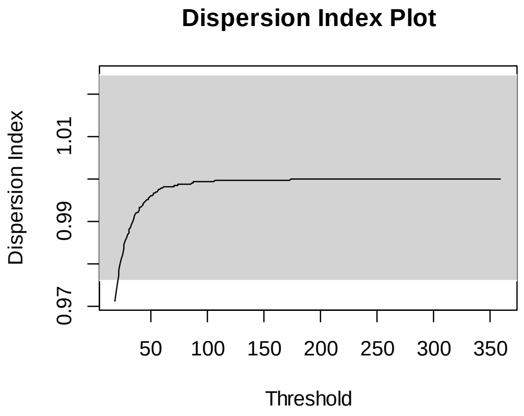


Figure 8: Dispersion index plot

Table 4: Goodness of fit testing for each series

Test	Annual Maximum		Partial Duration Series	
Lognormal (3)	0.6724	<0.05	**0.0086**	**0.3371**
Generalized Logistic	<0.05	<0.05	0.0002	<0.05
Generalized Extreme Value	**0.4549**	**0.5027**	0.0003	0.0205
Generalized Pareto	0.6136	<0.05	0.0060	0.2444
Gumbel	0.0385	0.0076	0.0095	<0.05

Table 5: Streamflow magnitude at selected return period

Return	**Streamflow Magnitude (m^3/s)**		
Period	**Annual Maximum**	**Partial Duration Series**	
(Years)	**Series**	**Lognormal**	**Generalized Pareto**
		(3 Parameters)	
5	136.1	89.0	89.0
10	197.1	115.9	116.9
20	280.4	150.6	154.5
50	441.1	212.2	224.7
100	618.4	274.5	299.5

4. Discussion and Conclusion

The larger the number of recorded data, the higher probability of observing infrequent events of high magnitude; therefore, the data will be more skewed and the analysis more accurate (Westen & Jetten, 2015); (Schlögl & Laaha, 2017). The threshold value selected in this study is based on percentile, where data above 90% in flow duration curves (FDC) is selected for the analysis. Table

2 presents the characteristics of the PDS samples for the period from 1978-2013.

Sample size in Table 2 represents the number of events which exceeds the selected threshold after considering the independence of the events. The rejected peaks are the number of events excluded from the study based on the threshold value. As the threshold value gets larger the number of samples being considered decreases. The magnitude of discharge selected is the highest peak in each cluster. Based on the dispersion index plot, the optimal threshold should be selected when the plot stabilizes at around 1. Figure 2 shows the threshold value against the dispersion index based on the assumption of the Poisson process. It begins to stabilize as the threshold approaches 50m^3/s. Hence, a threshold of 48.7m^3/s is selected for this study.

The calculated return period which considers annual maximum series gives a higher estimated magnitude compared to that of partial duration series. There is slight difference between the estimated streamflow value for the 5- and 10-year return period, and the difference increases as longer periods of between 20 years to 100 years return period are used. However, the estimated flow determined using the partial duration series for both distributions shows a small difference for a short period although the difference is rather significant at about 25m^3/s for a 100-year return period.

Rainfall has a strong influence on water flow rates, which in turn directly influence the occurrence of flood events in Malaysia. Previous research has shown that the intensity and frequency of extreme rainfall events are on the increase, thus creating a non-stationary component. This is essentially the consequence of climate change. Thus, it is important to make a precise estimation of flooding events that are triggered by rainfall in an effort to minimize property damage and environmental impact. It is also important to consider the change point detection brought forth by climate change and other human activities.

Acknowledgements

The authors would like to acknowledge the Earth Observation Center, Universiti Kebangsaan Malaysia for providing the climate change data for this research. Besides that, the authors would like to thank the Ministry of Higher Education (MOHE), Malaysia for supporting this research via research grant, FRGS/1/2016/STG06/UKM/02/1.

References

1. Alahmadi, F., Abd Rahman, N., & Abdulrazzak, M. (2014). Evaluation of the best fit distribution for partial duration series of daily rainfall in Madinah, western Saudi Arabia. *Proceedings of the International Association of Hydrological Sciences*, *364*(June), 159–163. https://doi.org/10.5194/piahs-364-159-2014
2. Bílková, D. (2014). Alternative Tools of Statistical Analysis: L-moments and TL-moments of Probability Distributions. *Pure and Applied Mathematics Journal*, *3*(2), 14. https://doi.org/10.11648/j.pamj.20140302.11
3. Chang, K. B., Lai, S. H., & Othman, F. (2016). Comparison of Annual Maximum and Partial Duration Series for Derivation of Rainfall Intensity-Duration-Frequency Relationships in Peninsular Malaysia. *Journal of Hydrologic Engineering*, *21*(1), 05015013. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001262
4. Claps, P., & Laio, F. (2003). Can continuous streamflow data support flood frequency analysis? An alternative to the partial duration series approach. *Water Resources Research*, *39*(8), 1–11. https://doi.org/10.1029/2002WR001868
5. Franchini, M., Galeati, G., & Lolli, M. (2005). Analytical derivation of the flood frequency curve through partial duration series analysis and a probabilistic representation of the runoff coefficient. *Journal of Hydrology*, *303*(1–4), 1–15. https://doi.org/10.1016/j.jhydrol.2004.07.008
6. Gado, T. A., Nguyen, V., & Asce, M. (2016). Regional Estimation of Floods for Ungauged Sites Using Partial Duration Series and Scaling Approach. *Journal of Hydrologic Engineering*, *21*(12), 1–12. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001439.
7. Garba, H., Ismail, A., & Tsoho, U. (2013). Fitting Probability Distribution Functions To Discharge Variability Of Kaduna River. *International Journal of Modern Engineering Research*, *3*(5), 2848–2852.
8. Gharib, A., Davies, E. G. R., Goss, G. G., & Faramarzi, M. (2017). Assessment of the combined effects of threshold selection and parameter estimation of generalized Pareto distribution with applications to flood frequency analysis. *Water (Switzerland)*, *9*(9). https://doi.org/10.3390/w9090692
9. Jiang, S., & Kang, L. (2019). Flood frequency analysis for annual maximum streamflow using a non-stationary GEV model. In *ARFEE 2018*. Wuhan. https://doi.org/10.1051/e3sconf/20197903022
10. Karim, F., Hasan, M., & Marvanek, S. (2017). Evaluating Annual Maximum and Partial Duration Series for Estimating Frequency of Small Magnitude Floods. *Water*, *9*(7), 481. https://doi.org/10.3390/w9070481
11. Keast, D., & Ellison, J. (2013). Magnitude frequency analysis of small

floods using the annual and partial series. *Water (Switzerland)*, *5*(4), 1816–1829. https://doi.org/10.3390/w5041816

12. Madsen, H., Rasmussen, P. F., & Rosbjerg, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologi events, *33*(4), 747–757.
13. Makkonen, L. (2006). NOTES AND CORRESPONDENCE: Plotting Positions in Extreme Value Analysis. *Journal of Applied Meteorology and Climatology*, *45*(February), 334–340. Retrieved from https://journals.ametsoc.org/doi/pdf/10.1175/JAM2349.1
14. Malamud, B. D., & Turcotte, D. L. (2006). The applicability of power-law frequency statistics to floods. *Journal of Hydrology*, *322*(1–4), 168–180. https://doi.org/10.1016/j.jhydrol.2005.02.032
15. Murthy, D. S., Jyothy, S. A., & Mallikarjuna, P. (2017). Probability Distributions of Annual Maximum Daily Streamflows using L-Moments-A Case Study. *International Journal of Civil Engineering and Technology (IJCIET)*, *8*(6), 290–302. Retrieved from http://www.iaeme.com/IJCIET/issues.http://www.iaeme.com/IJCIET/issues.IJCIET/index.asp290http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=6http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=5
16. Schlögl, M., & Laaha, G. (2017). Extreme weather exposure identification for road networks – a comparative assessment of statistical methods. *Natural Hazards and Earth System Sciences*, *17*(4), 515–531. https://doi.org/10.5194/nhess-17-515-2017
17. Syed Hussain, T. P. R., & Ismail, H. (2013). Flood frequency analysis of Kelantan River Basin, Malaysia. *World Applied Sciences Journal*, *28*(12), 1989–1995. https://doi.org/10.5829/idosi.wasj.2013.28.12.1559
18. Tallaksen, L., & Hewa, G. A. (2008). *Extreme value analysis*. Retrieved from http://www.wmo.int/pages/prog/hwrp/publications/low-flow_estimation_prediction/WMO 1029 en.pdf
19. Tekolla, A. W. (2010). *Rainfall and Flood Frequency Analysis in Pahang River Basin, Malaysia*. *Master of Science Thesis in Water Resources Engineering*.
20. Ummi Nadiah, A., Ani, S., & Zahrahtul Amani, Z. (2013). An analysis of annual maximum streamflows in Terengganu, Malaysia using TL-moments approach. *Theoretical and Applied Climatology*, *111*(3–4), 649–663. https://doi.org/10.1007/s00704-012-0679-x
21. Westen, C. V., & Jetten, V. (2015). 2.3 Magnitude-frequency analysis. *Carribean Handbook of Risk Information Management*, 1–9.

Trend analysis and return period of sea level on West Coast of Peninsular Malaysia

Firdaus Mohamad Hamzah[1], Christer Loh Chai Jia[2], Hazrina Tajudin[3], Hafizan Juahir[4]

[1 2 3] Universiti Kebangsaan Malaysia

[4] Universiti Sultan Zainal Abidin (UniSZA)

Abstract

In recent years, rising sea levels will affect about three million location. Besides that, the global average sea level increased in the 20th century and continued to do so. The large tide phenomenon occurring in several western coastal states of Peninsular Malaysia since September 19, 2016 was due to the gravitational attraction of the moon with the earth when the moon was parallel to the earth in orbit or when the moon was full floating. This phenomenon is characterized as normal and happens every month, but with strong winds and heavy rains that occur simultaneously, it can cause floods in certain areas. The purpose of this study was to determine the trend and changing pattern of sea level and predict the return period of sea level in three study areas located on the West Coast of Peninsular Malaysia. In this study, Mann-Kendall test and Theil-Sen Trend Line Test were used to study the rising trend of sea level. Extreme General Value Distribution (GEV) is used to obtain a return period of sea level. In this study, Port Klang did not show any significant monotonic trend as well as the Permatang Sedepa and Bagan Datuk stations. Port Klang shows a downward trend due to the negative gradient value. However, Permatang Sedepa and Bagan Datuk shows an upward trend due to positive value of gradient. The elevation of sea level for Port Klang, Permatang Sedepa and Bagan Datuk for a return period of 100 years are 3.3821, 2.9912 and 2.1441 m respectively. This study able to raise public awareness on sea level rise issues and take reasonable measures to prevent this problem from getting serious.

Keywords

Sea level; Mann-Kendall test; Theil-Sen Test; Generalized Extreme Value; L-Moment.

1. Introduction

Water level specifically refer to sea level elevations above some benchmark (Pugh & Woodworth, 2014). The sea surface is always tilted by waves, gravities, tsunamis, internal and long-term wave effects and ocean currents, density and dynamic meteorological effects (Church et al., 2013). Increased sea level (SLR) due to climate change is a serious global threat. Scientific evidence is very encouraging. Global sea-level rates are faster than 1993 to 2003, about 3.1 [2.4

to 3.8] mm per year, compared with an average rate of 1.8 [1.3 to 2.3] mm per year from 1961 to 2003 (Poh Poh Wong et al., 2014); and far higher than the average rate of 0.1 to 0.2 mm / year recorded by geological data over the last 3,000 years. There are many of factors that cause sea level rise.

One of the factors that caused the phenomenon of high tide is global warming. The increase in population, and subsequently industrial development and agricultural activity increase the increase in greenhouse gas emissions (McLean et al., 2001). Study by (Khasnis & Nettleman, 2005) indicates that the global temperature rise recorded in the 20th century is between 0.3˚C to 0.6˚C and continue to increase rapidly. According to (CDM-Executive Board, 2007), weather changes occur due to activities involving greenhouse gases and industries. Apart from industrial activity, logging and land use activities also contribute to global weather changes. In 1994 alone, Malaysia produced 144 million tons of greenhouse gases (Ministry of Science, Technology and Environment Malaysia 2000).

The effects of this natural system have many potential socioeconomic effects, including the following identified by (McLean et al., 2001). It will increase the loss of property and coastal habitat. Not only that, rising sea levels will cause increased flood risk and loss of life. Additionally, the increasingly serious sea level increase will cause damage to coastal protection and coastal infrastructure. Ultimately, this will also cause the country to have problems losing its tourism, recreation, and transportation functions.

Malaysia has experienced serious erosion problems from which 4800 km, 1400 km (30%) of the coastal zone are subject to various erosion. Floods often occur in the southern states of Malaysia, including Selangor, Negeri Sembilan, Melaka, Johor and Pahang. Approximately 9% of the land area in Malaysia is exposed to floods that affect 3.5 million people. The floods of these states can be increased due to SLR. Three SLR scenarios 20 cm, 50 cm and 90 cm were checked for measuring biophysical effects in Malaysia. Biophysical effects are assessed based on existing studies but updated where relevant and aggregated to the national level (*MALAYSIA THIRD NATIONAL COMMUNICATION AND SECOND BIENNIAL UPDATE REPORT TO THE UNFCCC*, 2018). As a result of 3190 cm SLR, a large proportion of polder land beaches about 1200 km in Peninsular Malaysia alone will sink after the failure of the bund, if the deck is not met. There is also a problem of salt invasions in Malaysia due to SLR effects. This study determined presence of trend and changing pattern of sea level using Mann-Kendall and Theil-Sen test. Magnitude at selected return period is also computed using GEV distribution.

2. Methodology

The study area is located at Selangor and Perak, namely Bagan Datuk, Port Klang and Permatang Sedepa station. Bagan Datuk is the southern tip of Perak, as well as an end to the downstream Perak River. Port Klang is a city and a main shipping gateway in Malaysia. It is known as the Port of Swettenham and is the largest and busy port in Malaysia. Permatang Sedepa is an offshore lighthouse in the Straits of Malacca on the Malaysian waters, called One Fathom Bank (Permatang Sedepa), near the coast of Selangor.

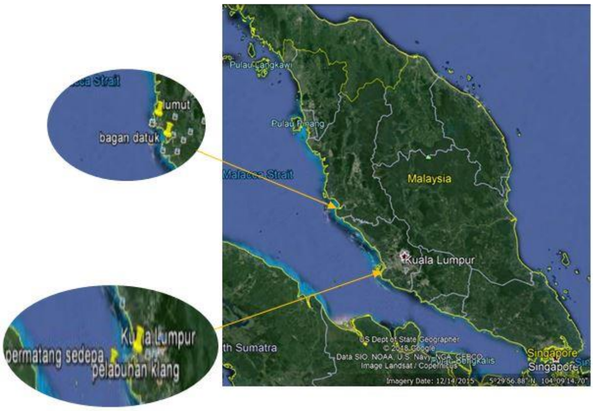


Figure 9: Maps of Tidal Stations

Mann Kendall test statistics is calculated using equation (1).

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(x_j - x_i)$$

where x_1, x_2, x_3, ..., x_n are the observation data, and x_j is data point and time *j*, where $(x_j - x_i) = \theta$

$$sgn(\theta) = \begin{cases} 1\ if\ (\theta) > 0 \\ 0\ if\ \ (\theta) = 0 \\ -1\ if\ (\theta) < 0 \end{cases} \qquad (2)$$

The data should be independent and identically distributed (i.i.d). S has an approximately normal distribution when $n \geq 8$ $w$ith the following mean and statistics:

$$E(S) = 0 \qquad (3)$$

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{m} t_i(t_i - 1)(2t_i + 5)}{18} \qquad (4)$$

where m is the number of groups with tied ranks and t_i is the tied observations. The standardized test statistics Z is calculated using the following formula:

$$Z = \begin{cases} \frac{S-1}{\sigma} \; if\; S > 0 \\ 0\; if\; S = 0 \\ \frac{S+1}{\sigma} \; if\; S < 0 \end{cases} \tag{5}$$

Theil-Sen is used to calculate trend slopes. It takes median of a slope between two points in a time series as the true slope. The approximate slope of each pairing for Theil-Sen estimator is calculated using equation (6).

$$m_{ij} = \frac{\left(x_j - x_i\right)}{j - i} \tag{6}$$

Q represents trend of steepness calculated using equation (7).

$$Q = \begin{cases} m\left(\frac{N+1}{2}\right) \quad if\; N\; is\; odd \\ \left(\frac{m\left(\frac{N}{2}\right) + m\left(\frac{N+2}{2}\right)}{2}\right) \quad if\; N\; is\; even \end{cases} \tag{7}$$

Generalized Extreme Value (GEV) probability density function and cumulative distribution function is shown in equation 8 and 9.

$$(x) = \frac{1}{\alpha}\left[\left(1 + \frac{\kappa(x-\xi)}{\alpha}\right)^{-\frac{1}{\kappa}}\right]^{\kappa+1} exp\left[-\left(1 + \frac{\kappa(x-\xi)}{\alpha}\right)^{-\frac{1}{\kappa}}\right] \tag{8}$$

$$F(x) = exp\left\{-\left(1 - \frac{\kappa(x-\xi)}{\alpha}\right)^{\frac{1}{\kappa}}\right\} \tag{9}$$

3. Results

Table 6: Descriptive Analysis

Stations	Minimum	Maximum	Mean	Standard Deviation
	(m)	(m)	(m)	
Pelabuhan Klang	2.76	3.25	2.96	0.10693
Permatang Sedepa	2.19	2.87	2.61	0.12272
Bagan Datuk	1.56	1.83	1.68	0.06854

Table 7: Mann-Kendall Trend test

Station	τ	p-value
Pelabuhan Klang	-0.0438	0.1235
Permatang Sedepa	0.093	0.0064
Bagan Datuk	0.121	0.292

Table 8: Theil-Sen test

Station	Slope	Confidence interval	
Pelabuhan Klang	-0.0030	-0.0019	-0.0014
Permatang Sedepa	0.0012	0.0005	0.0008
Bagan Datuk	0.0007	0.0001	0.0004

Table 9: Magnitude at selected return period

Station	Magnitude (m)				
	5	10	20	50	100
Pelabuhan Klang	2.90	2.93	2.96	3.17	3.38
Permatang Sedepa	2.52	2.54	2.57	2.78	2.99
Bagan Datuk	1.67	1.69	1.73	1.93	2.14

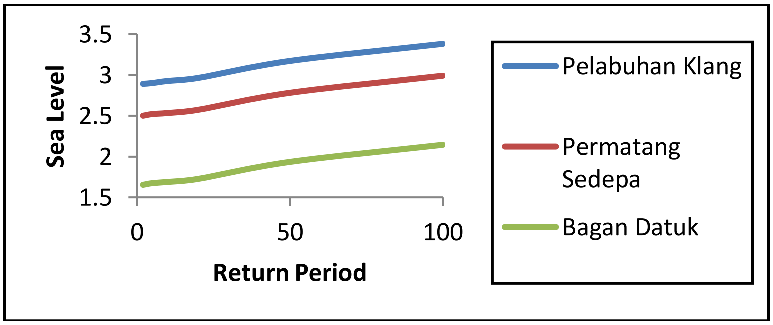


Figure 10: Return period plot

4. Discussion and Conclusion

Based on the results of the analysis, the study involved trends for sea level rise is important for early detection to predict sea levels in the future. This can give the engineering perspective to pay attention when planning a structure close to the sea, to minimize the rise of sea level on the building. Sea level rise will result in coastal inundation in low-lying coastal areas (Umi Amira Jamaluddin, Choun-Sian Lim, & Joy Jacqueline Pereira, 2017). Based on the results of the analysis, the sea level rise phenomenon for all three study areas is due to the impact of greenhouse gases as human daily activity releases carbon dioxide gas such as combustion of coal, oil and natural gas produced by plantation activity and changes in land use and some other industrial gas that have long been and will not happen naturally.

Since the impact of sea level increases can be a disaster for developing countries. The World Bank's estimates in 2007 suggest that although the rise of one meter at sea level in coastal countries in developing countries will drown 194,000 square kilometers of land and affect at least 56 million people (Umi Amira Jamaluddin et al., 2017). Malaysia has experienced an average loss of RM100 million a year (conservative estimates) due to the flood (31). Loss of agricultural production due to flood (or erosion) during floods in Johor lost RM 46 million for Johor West Agricultural Development Project. This is because SLRs with 1 m in Malaysia in 1999 will cause flooding and flooding about 100,000 hectares of land cultivated with palm oil and 80,000 million land under rubber production (L. C. Report 3). However, the real effect depends on the country's ability to adapt to the potential for sea level rise.

The elevation of sea level for Port Klang for a return period of 100 years is 3.38 m. For the same period of return, Permatang Sedepa and Bagan Datuk stations are 2.99 m and 2.14 m. It shows the maximum height of sea level for a return period of 100 years in Pelabuhan Klang and Permatang Sedepa higher than in Bagan Datuk station. There is an increasing trend of sea level so some of the actions need to take to prevent or decrease the increasing of elevation of sea level in Malaysia.

In 1997, the erosion had destroyed some of the estuary of the river and the coast of Kuala Kemaman. As can be seen on the estuary and coastline, the area is near the fish landing jetty. In addition, the erosion caused the coastline to retire as far as 20.5m from its original position. This caused the beach front to disappear due to being eroded. In this area, three houses, 0.52 hectares of cemeteries and part of the football field have been eradicated (Toriman, 2006). Based on the research conducted, Malaysia cannot avoid the imbalance of natural phenomena such as erosion of rivers and coastlines. Of the total 4,809 km of beaches that surround Malaysia, it is estimated that almost 1,400 km of coastlines are detected by critical erosion problems. The total number is 65 coastal areas with most of the prime locations located on the east coast of Peninsular Malaysia.

Through the National Coastal Era Study conducted since 1987, the Malaysian government has implemented a two-way strategy (short and long term) for coastal erosion control. A short-term strategy that focuses on the construction of coastal erosion control structures in critical areas, aimed at preventing the loss of many other valuable facilities, property and land due to coastal erosion. Long-term strategy, focusing on management aspects and avoiding any future coastal protection needs, with due consideration to the effects of coastal erosion because of development, during the planning and implementation of new projects in coastal zone areas. This can be achieved through non-structural and enforcement measures.

Adaptation measures are needed to minimize the adverse effects of SLR and to protect coastal resources and livelihoods in Malaysia. This in turn requires cost to achieve adjustment measures but this is much less than the cost of SLRs without adjustments. According to (Smith, Cialone, Wamsley, & McAlpin, 2010) the cost of SLRs for Malaysia with and without adjustments is 160.92 and 655.09 million US $ / year in 2100. But adjustments must be made to consider greater uncertainty about the future climate (and many other factors), so there is a need for risk-based and uncertain ways of finding solutions. Thus, all levels of the government have an important role in developing planned adjustments (Nicholls et al., 2011). There are several adaptation approaches that can be used to minimize the negative impact of SLR. In conclusion, this study is important for investigating flood estimates for them to design coastal zone structures to reduce the risk of failure and minimize the impact of environmental damage caused by rising sea levels.

Acknowledgements

The authors would like to acknowledge the Earth Observation Center, Universiti Kebangsaan Malaysia for providing the data for this research. The authors would also like to thank Yayasan Sime Darby (YSD) and Ministry of Higher Education (MOHE) Malaysia for supporting this research via research grants ZF-2017-008 and FRGS/1/2016/STG06/UKM/02/1 respectively.

References

1. CDM-Executive Board. (2007). *Procedure to determine when accounting of the soil organic carbon pool may be conservatively neglected in CDM A/R project activities*. Retrieved from https://cdm.unfccc.int/methodologies/ARmethodologies/tools/ar-am-tool-06-v1.pdf
2. Church, J. A., P. U. Clark, A. Cazenave, J. M. Gregory, S. Jevrejeva, A. Levermann, … A. S. Unnikrishnan. (2013). *Sea Level Change*. (T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, … P. M. Midgley, Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group 1 to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, New York, USA. Retrieved from https://www.ipcc.ch/site/assets/uploads/2018/02/WG1AR5_Chapter13_FINAL.pdf
3. Khasnis, A. A., & Nettleman, M. D. (2005). Global Warming and Infectious Disease. *Archives of Medical Research*, *36*(6), 689–696. https://doi.org/10.1016/j.arcmed.2005.03.041

4. *MALAYSIA THIRD NATIONAL COMMUNICATION AND SECOND BIENNIAL UPDATE REPORT TO THE UNFCCC*. (2018). Putrajaya. Retrieved from https://unfccc.int/sites/default/files/resource/Malaysia NC3 BUR2_final high res.pdf
5. McLean, B. R., Tsyban, A., Burkett, V., Codignott, J., Forbes, D., Mimura, N., … White, K. S. (2001). Coastal Zones and Marine Ecosystems. In James J. McCarthy Osvaldo F. Canziani, Neil A. Leary, David J. Dokken, & Kasey S. White (Eds.), *Climate Change 2001: Impacts, Adaptation and Vulnerability* (pp. 347–379). Cambridge: Cambridge University Press. Retrieved from http://papers.risingsea.net/IPCC.html
6. Nicholls, R. J., Marinova, N., Lowe, J. A., Brown, S., Vellinga, P., de Gusmão, D., … Tol, R. S. J. (2011). Sea-level rise and its possible impacts given a 'beyond 4°C world' in the twenty-first century. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1934), 161–181. https://doi.org/10.1098/rsta.2010.0291
7. Poh Poh Wong, Inigo J. Losada, Jean-Pierre Gattuso, Jochen Hinkel, Abdellatif Khattabi, Kathleen L. McInnes, … Athanasios Vafeidis. (2014). Coastal Systems and Low-Lying Areas. In Robert J. Nicholls & Filipe Santos (Eds.), *Climate Change 2014: Impacts,Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 361–409). New York: Cambridge University Press, Cambridge, United Kingdom and New York, USA. Retrieved from https://www.ipcc.ch/site/assets/uploads/2018/02/WGIIAR5-Chap5_FINAL.pdf
8. Pugh, D., & Woodworth, P. (2014). *Sea-level science : understanding tides, surges, tsunamis and mean sea-level changes* (Second). Cambridge: Cambridge University Press.
9. Smith, J. M., Cialone, M. A., Wamsley, T. V., & McAlpin, T. O. (2010). Potential impact of sea level rise on coastal surges in southeast Louisiana. *Ocean Engineering*, *37*(1), 37–47. https://doi.org/10.1016/J.OCEANENG.2009.07.008
10. Toriman, E. (2006). Hakisan Muara dan Pantai Kuala Kemaman, Terengganu: Permasalahan Dimensi Fizikal dan Sosial. *Akademika*, (Julai), 37–55.
11. Umi Amira Jamaluddin, Choun-Sian Lim, & Joy Jacqueline Pereira. (2017). Climate change and its implications on the coastal zone of Kuala Selangor, Malaysia. *Warta Geologi*, *43*(July-September), 338. Retrieved from http://archives.datapages.com/data/meta/geological-society-of-malaysia/warta-geologi-newsletter/043/043003/pdfs/338_firstpage.

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 978-967-2000-74-7

#ISIWSC2019