62nd ISI World Statistics Congress

ISI 2019 Kuala Lumpur

18 - 23 August 2019

PROCEEDING

CONTRIBUTED PAPER SESSION

VOLUME 1

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 1)

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 1, 2019. 435 pages

Preface

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

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

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

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

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

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

Thank you.

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

Scientific Programme Committee of the 62nd ISI WSC 2019

Chair
Yves-Laurent Grize, Switzerland

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

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

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

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

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

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

Local Programme Committee of the 62nd ISI WSC 2019

Chairperson
Rozita Talha

Co-Chairperson
Prof. Dr. Ibrahim Mohamed

Vice Chairperson
Siti Haslinda Mohd Din
Daniel Chin Shen Li

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

TABLE OF CONTENTS

Contributed Paper Session (CPS): Volume 1

Preface i

Scientific Programme Committee ii

Local Programme Committee iii

CPS651: Ageing population in Morocco: Reality and challenges 1

CPS653: Stochastic search variable selection for definitive screening designs in split-plot and block structures 9

CPS658: The impact of mis-specification of the deterministic components in the Co-integration model: An application to the bivariate case on South African employment costs and gross earnings 15

CPS694: A critical image of statistical analyses in medicine between 2006 and 2018 23

CPS798: Annualization of the labor force survey in the Philippines 31

CPS863: Redesigning of the corn production survey in the Philippines 38

CPS877: The evolution of the OECD countries after the 2008 financial crisis 45

CPS1071: Redesigning of the Semi-Annual Survey of Dairy Enterprises in the Philippines 55

CPS1085: Inference on P(X Y) for Bivariate Normal Distribution based on Censored Data 63

CPS1110: Bayesian inversion into soil types with Kernel-Likelihood Models 70

CPS1119: Forecasting Agricultural Commodity Prices in Turkey Using Artificial Neural Network 78

CPS1158: Bayesian Inference of Multiple Structural Breaks in Multiple Regimes Threshold Autoregressive Model 84

CPS1167: Prediction for censored lifetimes from Weibull distribution in Khamis and Higgins Step- stress model 92

CPS1178: Statistics Education in Indian Agricultural Universities and Agro-Technical Institutes 101

CPS1196: Clustering Chinese cities' economic growth paths with dynamic time warping 108

CPS1198: Hedonic modeling and Brownian index: Case of Morocco 115

CPS1201: Combined criteria for dose optimisation in early phase clinical trials 124

CPS1216: Parameter estimation for misspecified diffusion processes with noisy, nonsynchronous observations 134

CPS1239: The Servicification Of Manufacturing In Asia: Redefining The Sources Of Labor Productivity Using Time 145

CPS1254: The measurement of inequality of opportunity in China 156

CPS1255: Predicting the number of newly found rare species 164

CPS1266: Improving the quality of official statistics in Iran's manufacturing establishments statistics 169

CPS1277: Use of GIS and statistics in measuring land consumption rate to population growth rate in Egypt 179

CPS1278: An application of the Generalized Lavallee-Hidiroglou Algorithm for Stratification in Business Surveys: A sampling methodology for the census of the Philippine Business and Industry 188

CPS1280: Spatial analysis of winter rainfall variability of a selected region in South Africa 194

CPS1282: Spatial distribution characteristics of lightning disaster risk and analysis of factors based on GIS---The Case of Hohhot 201

CPS1283: Application of Generalised Linear Models during the phased liberalisation of the Malaysian motor and fire tariffs 209

CPS1284: Refugees in labor market: Do they act as a marginalisation tool? A decomposition of the wage gap in Palestine 217

CPS1290: Probit model of decent living house ownership determinants in DKI Jakarta province 2017 226

CPS1297: High frequency value-at-risk analysis: An empirical study for IPC Mexican Index 236

CPS1304: Modeling seasonal epidemic data using Integer Autoregressive Model 241

CPS1307: Latin Square Designs: Causal inference in a potential outcomes framework 249

CPS1308: Undercoverage bias of web survey on smoking and heavy drinking 259

CPS1333: Spatial Fay-Herriot model for estimating expenditure in Bangka Belitung province, Indonesia 266

CPS1335: On modelling the frequency of antenatal care visits in Bangladesh 274

CPS1339: Identification of Disaggregated Hotspots of Child Morbidity in Bangladesh: An Application of Small Area Estimation Method 284

CPS1340: Mobility and state dependence in a labour market characterized by informality: The case of Morocco 292

CPS1341: Short-distance and long-distance elderly migration in Indonesia year 2016: Application of Multilevel Logistic Regression Analysis 301

CPS1342: Multidimensional poverty in east Nusa Tenggara: A structural equation modelling approach 310

CPS1343: Propensity score based adjustment for covariate effects on classification accuracy of biomarker using ROC curve 319

CPS1345: The construction of composite index to measure accessibility, quality and people behavior of drinking water in Indonesia 330

CPS1349: A solution to separation in Poisson regression for small or sparse count data 338

CPS1353: Inclusive economic growth in JAVA 345

CPS1356: Childhood under nutrition in Bangladesh: A policy - Suggestive empirical analysis 354

CPS1358: Bivariate Archimedean Copula model: An application to Customer Price Index (CPI) and Wholesale Price Index (WPI) in Indonesia 363

CPS1373: Evidence based indicators for local educational monitoring - detective work for the district Berlin-Mitte 368

CPS1376: Optimal sample size allocation: An investigation for evaluating the behavior of a dietetic supplement 377

CPS1379: A Bayesian quantile time series model for asset returns 386

CPS1381: Comparing record linkage methods for data integration on Brazilian agriculture 391

CPS1394: The impact of ageing population, unemployment, obesity, inflation, out-of-pocket expenses, and income per capita on Malaysia's health expenditures: A linear regression analysis 399

CPS1398: New Zealand crime and victims survey: Filling the gap 409

CPS1406: Clustering planar shapes combined with multidimensional scaling 418

Index 425

Ageing population in Morocco: Reality and challenges

El Mostafa TOGUI
High Commission for Planning, Rabat, Morocco

Abstract

The process and mechanisms of ageing are ongoing in Moroccan society. The extreme speed and extent of the phenomenon requires highlighting different aspects that characterize it. While in 2014 people aged 60 and over represent 9.6% of all the entire population of Morocco, this proportion is expected to increase by half between now and 2020. In 2050 one Moroccan in 4 will be over 60. Though individual situations differ greatly, the Moroccan elderly population is currently characterized by a feeble coverage of retirement schemes and health insurances. Several indicators emphasize the vulnerability of this age group especially women. Moreover, family solidarity modes are subjected to different influences and changes in parallel with all the societal changes. So, we note that this field, as in several countries, suffers from a very obvious lack of updated statistics. Indeed, an Integrated National Strategy based on several action plans with an adequate regulatory framework seems unavoidable to face the realities and challenges of demographic metamorphosis in the future, in particular, the social welfare system.

Keywords

Ageing population; Living Conditions; Family Support; Social Security; Retirement

1. Introduction

Many countries are currently experiencing, to varying degrees more or less advanced, the phenomenon of population ageing. The number of over 60 years in the world should at least double, from 900 million in 2015 to 2 billion in 2050. Demographically, Morocco is often presented as a "young" country to the extent that nearly one out of three is under 15 years. Over the decades, Morocco hasn't escaped this phenomenon since its elderly population represents a proportion increasingly important and is growing at an unprecedented pace. Thus, the share of people aged 60 and over represented 9.6% in 2014 versus 8.1% in 2004. This proportion should increase to half by 2020. In 2050, 1 Moroccan out of 4 will be over 60. As it was noted by the CERED[1] (2005, p. 58), "the ageing population represents, undoubtedly, the

[1] Center for Studies and Research in Demography

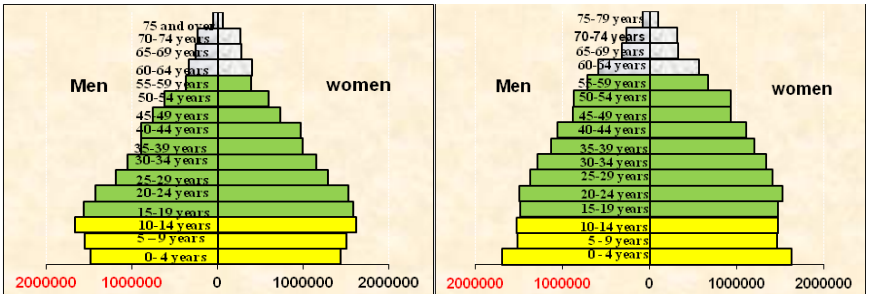
most spectacular demographic characteristic that will be a feature of Morocco in the next three or four decades ". After the challenge of controlling fertility, Morocco will have to take on a new one, that of economic and social support of the ageing population.

The demographic transition towards an upward demographic ageing

The continuous progress in raising life at birth summarizes the effects of the decline in mortality at all ages of life. During the period 1950-1955, life expectancy at birth in Morocco was only 43; it currently exceeds 72 years old. In the process of this demographic transition, lower risk of dying at each age was accompanied by a decline in fertility which also contributes to the ageing population. The average number of children per woman (fertility rate, or TFR) stood at 2.21 children per woman in 2014 against 2.47 in 2004, knowing it was 7 in 1960s. In urban areas, fertility has fallen below the replacement level of generations with a TFR of 2.01 versus 2.55 in 2004. In rural areas, it stood at 2.55 in 2014 against 3.10 in 2004. This trend is moving to a convergence in fertility between the two areas of residence.

All these changes had the impact on the age structure in Morocco which has experienced more or less lucid transformation. Thus, between 2004 and 2014 the share of young people under 15 declined from 31.2% to 28.0%, the working age population (15-59 years) remains important in passing from 61.2% to 62.4% while the share of people aged 60 and over is steadily increasing (9.6% 2014).

Figure 1- Age structure (2004 and 2014)



Source: HCP RGPH 2014

1.1 The ageing of the Moroccan population: an ongoing process

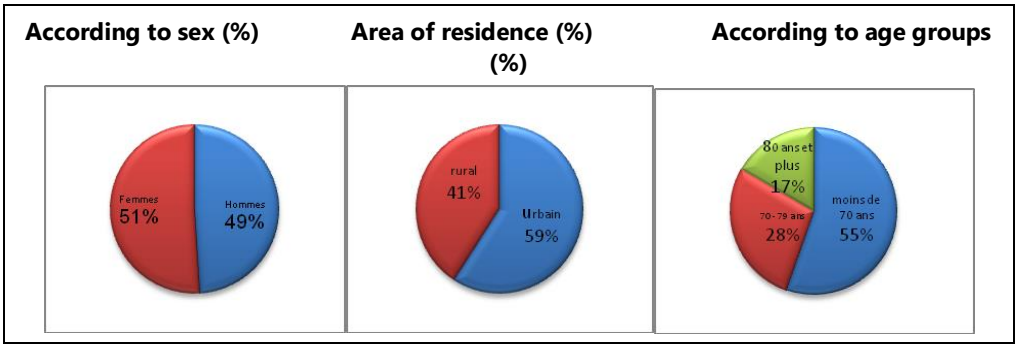
A demographic revolution is underway in the world. Today the number of people aged 60 and over is about 900 million; this number will double by 2025 and reach two billion by 2050, the majority in developing countries. In 2050, 80% of older people will be living in low- and middle-income countries.

Like many countries in the world, Morocco hasn't escaped this phenomenon because its elderly population represents a progressive gradual

increase in the total population and is growing at an unprecedented pace, and this is due to the changes that currently characterize the Moroccan society. The share of the people aged 60 years old and over is 9.6% now, but it was 8.1% in 2004. This corresponds to the numbers of 3,209,000 in 2014 and 2,376,000 in 2004, a relative increase of 35% during the census period.

The female proportion in the last census of the Moroccan population was 51% and 6 in 10 (59%) of seniors living in urban areas. By age, more than half (55.4%) are aged under 70 years old, 28.0% from 70 to 79 years and 16.6% aged 80 and over. Just over two-thirds of seniors (68.0%) are married (92.2% men and 44.8% women) and nearly 27.7% are widowed (4.8% men and 49.6% women).

Figure 2- Illustrated portrait of elderly According to sex, Area of residence and age groups



Source: HCP RGPH 2014

In the same context, the lengthening of the average life induces progressive enlargement of the top of the pyramid. But despite significant progress in longevity in the general population, large disparities persist, especially between urban and rural contexts.

Moreover, the spatial distribution of ageing in Morocco is illustrated as follows: it's more important in different regions Beni-Mellal-Khénifra (10.6%), Oriental (10.3%) and Fez-Meknes (10, 2%), probably because of high emigration. Yet, it's lower in other regions as Eddakhla-Oued Ed-Dahab (3.5%) and Laayoune-Sakia El Hamra (5.4%).

2. Illiteracy: A great vulnerability among the elderly

People aged 60 and over in 2014 were from the generations born before the independence, during which, access to education was still a privilege only for a minority, including the sons of notables and those of settlers (CERED 1995). Almost all people belonging to this age group could not access education, and this explains the high rate of illiteracy among them (70.5%). In rural areas, the rate is more prominent than in the urban areas with respectively 92.0% and 74.2%. More than 94.5% of women cannot read or write, against 69.6% of men.

However, despite their advanced age the RGPH 2014 revealed that 11.9% went to primary school, 9% are high school graduates and 2.4% studied in the university. The illiteracy rate increases with age from 3.7% among those aged less than 15 years to 61.1% among those aged 50 and more. It's relatively lower in all the age groups in 2014 than in 2004. It goes without saying that illiteracy is widespread among older Moroccans. Despite the efforts made by the competent bodies to eradicate illiteracy among adults, women remain the most vulnerable illiterate.

3. The activity of the elders: a closed circuit

Longer life opens possibilities not only for the elderly and their families but also for society as a whole. In Morocco, as in most developing countries, older people often continue to work. These extra years are an opportunity to engage in new activities such as additional training, a new career or a long-neglected passion. Older people also bring varied contributions to family and community. Nevertheless, the extent of these opportunities and contributions is largely dependent on the health factor. It should be noted in this context that the last census confirmed that 18.8% of people over 60 are employed with highly differentiated situations (by sex 34.8% for men and 3.4% among the women). This proportion is slightly higher in rural than in urban areas, it is 24.0% against 15.1% respectively.

Table 1- Persons aged 60 and over (%) by Activity Type, gender, Area of residence and age

	active Busy	Unemployed	Housewife	annuitant	Retirement	Sick or infirm	old	other inactive	Undeclared	Total
Male	34,8	1,6	0,0	0,6	29,7	8,1	24,5	0,6	0,1	100
Female	3,4	0,4	39,1	0,2	4,0	7,9	44,2	0,6	0,2	100
Total	18,8	1,0	20,0	0,4	16,6	8,0	34,5	0,6	0,2	100

Source: HCP RGPH 2014

We have to point out that the greater number of active seniors had an activity to meet their needs and those of their families. This majority (82.0%) worked for themselves, 71.1% as independents and 10.9% as employers, knowing that the participation rate of older people decreases progressively as people get older and older.

To better understand the conditions of life and activity of the elderly in both areas of residence, it would be important to consider the nature of the activity, the level of income received in return for those activities; time spent the exercise of the activity, the reasons for the exercise and the meaning given to it by individuals.

4. The elderly: environment and contrasting perspectives

Today we see that the population in these countries gets older at a much faster rate than has been the case for the currently developed world. This means they will have less time to prepare for the ageing population. Moreover, this process goes along over the years with many societal changes.

Thus, the review of existing data has highlighted that only 1/5 of the elderly has a social and medical insurance. Few have access to care, and their physical and financial dependence increases, in a context where the support of these persons within the family is threatened, particularly by having more and more nuclear households.

4.1. Family environment: an exceptional social haven

The examination of the distribution of the population aged 60 and over by family relationship helps shed light on the family environment in which they live. In general, it seems that almost all the elderly are surrounded by people in their families. Thus, in 2014, only 5.2% (170130) of older people live alone, 73% (124615) of women and, given the specificities of the Moroccan society and the Arab-Muslim culture in general that make the family a hard core of social solidarity for the elderly

The position occupied by the elderly in the society is very important and is seen through numerous symbols and images of respect. Facing weak role of specialized social welfare institutions, the majority of older Moroccans are supported by the family "generous refuge". 77.5% of this population receive material assistance from their families, in particular their children. This shelter also refers to the issue of cohabitation of several generations in the same housing and lodging for the elderly.

Thus, referring to the relationship of elders with their family / social environment raises the notion of dependency that often refers to the representation health / morbidity. Ageing is the product of the accumulation of a wide range of molecular and cellular damage over the years. This one leads to progressive degradation of physical and mental abilities.

4.2. Ageing and Health: a close correlation

Generally, the ageing process is characterized by number of living conditions (physical, emotional, psychological etc.). Also Old age is characterized more by diseases known to others and are quite complex and usually occure late in life and don’t constitute separate disease categories. This is what is commonly called geriatric syndromes also knowing that according RGPH 2014, 50.6% of disabled people are aged 60 and over.

This observation leads us to put the emphasis on vulnerability afflicting this population which particularly faces the shortcomings of health services.

Thus, this finding also leads us to meditate on this vulnerability in its various expressions.

Many studies reported that this population suffers from a glaring lack security, only 20% of seniors have a social and medical insurance. "Few of them have access to care in the presence of physical and financial dependence that increases with time, in a context where the support of these persons within the family is threatened, including many societal changes which the increasing number of nuclear households illustrates, "the EESC in 2015.

Furthermore, the results of (ENPA) revealed that almost two thirds (59.1%) of these people said they did not have the means. This proportion is higher among relatively Disadvantaged populations, it represents in rural areas (62.1%) and women (62.8%) while in urban areas (55.2%) and men (55.1 %).

The non-use of health care, once the person is ill, is higher in rural (62.1%) than the urban areas (55.2%), particularly among women (62.8%). This difference between the two areas is due, among others, to the uneven spatial distribution of healthcare services. Moreover, the hospitals suffer from a lack of specialized geriatric structures that take into account the peculiarities of the elderly.

Indeed, the current situation reveals failures in the health system in Morocco, among others, the weakness of medical coverage. The elderly are the first to suffer.

4.3. Retirement a major challenge to overcome

The transition from working life to retirement affects an entire system of life of older people whose financial resources (pensions) are an important component especially through some representations based on the existential security and survival of this fringe in conditions worthy of a human.

Like many countries, to varying degrees more or less advanced, Morocco is experiencing a significant increase in its population aged 60 and over, while it represented 9.6% in 2014 of the total population, its share should increase by half by 2020 to get to a Moroccan out of four over 60 years in 2050.

Therefore, several questions rise through the rate of the social coverage (retirement) for our elders and the quality value and the persistence of pensions and their collection modes. Furthermore, 16% of the Moroccans aged 60 and older reported receiving a retirement pension with a clear dichotomy in the two sexes because women are the most penalized by this system. Just 3% of women benefit versus 30.4% of men. Otherwise, the physical and financial dependence increases, in a context in which taking care of these persons within the family is threatened, particularly by the increasing number of nuclear households. The emergence of some specific centers for

the elderly reflects the various changes that the Moroccan society is undergoing.

Thus, the reduced number of members raises many questions of multiple dimensions concerning the pension system be it on social equity, structural dysfunction in the employment sector or on the suspicions of the financial situation of the system pension currently inflames social polemic and major policy on the national stage.

In the presence of this finding, numerous state institutions and civil and political society discussed this problem by proposing responses to different challenges facing Morocco to overcome the crisis threatening the sustainability of that system.

5. Conclusion

Morocco is among the countries whose development started later, but it seems that it has to accelerate at a significantly faster rhythm. On the other hand, the social status of elderly people has degraded and the situation is alarming: low education, precarious socio-economic situation for many seniors, low coverage by the pension system and discriminating access to care and a great vulnerability of these persons against the diseases etc.

This of course may be much less controllable, hence the need to take the necessary measures in time.

In this context, the economic, social and environmental council (EESC) established in 2015 a report on the elderly with the aim to develop a diagnosis and to make recommendations and proposals for future public policies. Here are the eight strategic domains for the protection of this category of the population:

- Improving social protection for the elderly, and primarily those in situation of dependence and vulnerability;
- Improving the legal and institutional framework;
- Enhancing the accessibility;
- Improving the care of the elderly
- Improving the welfare and health of the elderly;
- Promoting social participation of older people;
- Promoting knowledge of the situation of the elderly;
- Supporting and accompanying the elderly residing abroad.

Thus, it’s time to develop "an integrated public policy of protection for the elderly, with accompanying resources and assessment, which takes into account their rights in terms of dignity, participation and social inclusion ".

References

1. Legaré J. F., 1995. « Qualité de vie ou quantité de vie : un défi pour les sociétés industrielles », La sociologie des populations, pp. 463-480.
2. Rovira Eduardo Rodriguez, 2002. "Social Discrimination of the Aged", in Conferences 1998-2001. Madrid: Abumar Grandparents in Motion/Ministerio de Trabajo y asuntos socials, pp. 210-215.
3. Enquête Nationale sur les Personnes Agées, HCP, (CERED), 2006.
4. Conseil économique, social et environnemental (CESE) «les personnes âgées au Maroc » 2015.
5. Ghizlan Loumrhari, « Ageing, Longevity and Savings: The Case of Morocco» International Journal of Economics and Financial Issues Vol. 4, No. 2, 2014, pp.344-352.
6. United Nations New York, 2015, «World Population Ageing».
7. High Commission for Planning, 2014, general census of population -Morocco-.
8. High Commission for Planning 2005 « Personnes âgées au Maroc » CERED- Morocco-.

Stochastic search variable selection for definitive screening designs in split-plot and block structures

Chang-Yun Lin
Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung, Taiwan

Abstract

Split-plot definitive screening (SPDS) and block definitive screening (BDS) designs have been developed for detecting active second-order effects in screening experiments when split-plot and block structures exist. In the literature, the multistage regression (MSR) and forward stepwise regression (FSR) methods were proposed for analyzing data for the two types of designs. However, there are some limitations and potential problems with the regression approaches. First, the degrees of freedom may not be large enough to estimate all active effects. Second, the restricted maximum likelihood (REML) estimate for the variances of whole-plot and block errors can be zero. To overcome these problems and to enhance the detection capability, we propose a stochastic search variable selection (SSVS) method based on the Bayesian theory. Different from the existing Bayesian approaches for split-plot and block designs, the proposed SSVS method can perform variable selections and choose more reasonable models which follow the effect heredity principle. The Markov chain Monte Carlo and Gibbs sampling are applied and a general WinBUGS code that can be used for any SPDS and BDS designs is provided. Simulation studies are conducted and results show that the proposed SSVS method well controls the false discovery rate and has higher detection capability than the regression methods.

1. Introduction

Split-plot definitive screening (SPDS) and block definitive screening (BDS) designs have been developed for detecting active second-order effects in screening experiments when split-plot and block structures exist. In the literature, the multistage regression (MSR) and forward stepwise regression (FSR) methods were proposed for analyzing data for the two types of designs. However, there are some limitations and potential problems with the regression approaches. First, the degrees of freedom may not be large enough to estimate all active effects. Second, the restricted maximum likelihood (REML) estimate for the variances of whole-plot and block errors can be zero. To overcome these problems and to enhance the detection capability, we propose a stochastic search variable selection (SSVS) method based on the Bayesian theory. Different from the existing Bayesian approaches for split-plot and block designs, the proposed SSVS method can perform variable selections

and choose more reasonable models which follow the effect heredity principle. The Markov chain Monte Carlo and Gibbs sampling are applied and a general WinBUGS code that can be used for any SPDS and BDS designs is provided. Simulation studies are conducted and results show that the proposed SSVS method well controls the false discovery rate and has higher detection capability than the regression methods.

2. Models and estimations

For an m-factor SPDS (or BDS) design with w whole plots (or blocks), let y_{ij} denote the response from the jth run in the ith whole plot (block), where $i = 1, \dots, w, j = 1, \dots, n_i$, and n_i is the size of the ith whole plot (block). Then the full second-order model for the SPDS (BDS) design can be written as (1)

$$y_{ij} = \beta_0 + \sum_{r=1}^{m} \beta_r x_{r,ij} + \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} \beta_{rs} x_{r,ij} x_{s,ij} + \sum_{r=1}^{m} \beta_{rr} x_{r,ij}^2 + \gamma_i + \epsilon_{ij},$$

where β_0 is the intercept, $x_{r,ij}$ is the level of factor X_r for the jth run in the ith whole plot (block), β_r and β_{rr} are the main effect and quadratic effect of factor X_r, respectively, β_{rs} is the interaction of factors X_r and X_s, γ_i is the random effect for the ith whole-plot (block), and ϵ_{ij} is the random error for the jth run in the ith whole plot (block). It is assumed that γ_i and ϵ_{ij} have zero means with variances σ_γ^2 and σ_ϵ^2, respectively, and are mutually independent. Equation (1) can be further expressed in the form of matrices as follows. Let $\boldsymbol{y} = \left(y_{11} y_{12,\dots,} y_{w(n_w-1)}, y_{wn_w}\right)'$ be the $n \times 1$ vector of responses, where $n = \left(\sum_{i=1}^{w} n_i\right.$, $\beta = (\beta_0, \beta_1, \dots, \beta_m, \beta_{12}, \dots, \beta_{(m-1)m}, \beta_{11}, \dots, \beta_{mm})'$ be the $(1+q) \times 1$ vector of the intercept and fixed effects, where $q = 2m + m(m-1)/2$, $\boldsymbol{X}$ be the $n \times (1+q)$ model matrix corresponding to $\beta, \gamma = (\gamma_1, \cdots, \gamma_w)'$ be the $w \times 1$ vector of the random whole-plot (block) effects, **U** be the $n \times w$ indicator matrix corresponding to γ and $\epsilon = \left(\epsilon_{11}, \epsilon_{12}, \dots, \epsilon_{w(n_w-1)}, \epsilon_{wn_w}\right)'$ be the $n \times 1$ vector of the random errors. Then Equation (1) can be expressed as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{U_\gamma} + \boldsymbol{\epsilon}.$$

If σ_γ and σ_ϵ are known, the GLS estimate for β is

$$\hat{\beta} = (X'\Sigma^{-1}\mathrm{X})^{-1}\mathrm{X}'\Sigma^{-1}y \tag{2}$$

and the covariance matrix of $\hat{\beta}$ is

$$V = (X'\Sigma^{-1}\mathrm{X})^{-1}, \tag{3}$$

Where $\sum = \sigma_\gamma^2 \mathrm{UU}' + \sigma_\epsilon^2 \mathrm{I}_n$ (I_n is an n x n identity matrix) is the covariance matrix of $\boldsymbol{y}$. In practice, σ_γ and σ_ϵ are usually unknown and can be estimated using the REML method. By replacing Σ in (2) and (3) with $\hat{\Sigma} = \hat{\sigma}_\gamma^2 \mathrm{UU}' + \hat{\sigma}_\epsilon^2 \mathrm{I}_n$, we obtain $\hat{\beta} = (\mathrm{X}'\hat{\sum}^{-1}\mathrm{X})^{-1}\mathrm{X}'\hat{\sum}^{-1}\mathrm{y}$ and $\hat{\mathrm{V}} = (\mathrm{X}'\hat{\sum}^{-1}\mathrm{X})^{-1}$. Although it is known

that $\hat{\Sigma}$ is a biased estimator of Σ (Kenward and Roger, 1997), $\tilde{\beta}$ and $\widetilde{\nabla}$ are commonly used for hypothesis tests in the regression methods.

For SPDS and BDS designs, the full second-order model is inestimable due to insufficient degrees of freedom. Hence, the process for model selections must be implemented. Two regression methods, MSR and FSR, have been proposed in the literature for performing model selections for SPDS and BDS designs. The two methods have the following limitations: (i) the active effects cannot be estimated if the number of them is greater than the rank of the model matrix and (ii) $\hat{\sigma}_\gamma = 0$ may be obtained by using the REML method. To circumvent these problems, we propose a Bayesian variable selection method as described in the next section.

3. Stochastic search variable selection

Let $\boldsymbol{\Theta}' = (\boldsymbol{\beta}', \boldsymbol{\delta}', \boldsymbol{\gamma}', \sigma_\epsilon, \sigma_\gamma)$ denote the vector of the model parameters, where $\delta = (\delta_1, \cdots, \delta_m, \delta_{12}, \cdots, \delta_{m-1,m}, \delta_{11}, \cdots, \delta_{mm})'$. By the Baye's Theorem (see DeGroot, 1970, p28, and Robinson et al., 2012), the joint posterior density of the model parameters is proportional to the product of the likelihood $f(\mathrm{y}|\Theta)$ and the joint prior density $\pi(\Theta)$, which is

$$\pi(\Theta|y) \propto f(y|\Theta)\pi(\Theta). \tag{4}$$

The responses conditioned on the model parameters are assumed independent and hence the likelihood in (4) has the form

$$f(\mathrm{y}|\Theta) = \prod_{i-1}^{w}\prod_{j=1}^{n_i} f(y_{ij}|\mu_{ij}, \sigma_\epsilon),$$

where $\mu_{ij} = \beta_0 + \sum_{r=1}^{m}\beta_r x_{r,ij} + \sum_{r=1}^{m-1}\sum_{s=r+1}^{m}\beta_{rs}x_{r,ij}x_{s,ij} + \sum_{r=1}^{m}\beta_{rr}x_{r,ij}^2 + \gamma_i$. The joint prior density of the model parameters can be obtained by

$$\pi(\Theta) \propto \prod_{v\in\Theta} f_{v|pa[v]}(v)$$

$$\propto f_{\beta_0}(\beta_0)\prod_{r=1}^{m} f_{\beta_{r|\delta_r}}(\beta_r)\prod_{r=1}^{m-1}\prod_{s=r+1}^{m} f_{\beta_{rs}|\delta_{rs}}(\beta_{rs})\prod_{r=1}^{m} f_{\beta_{rr}|\delta_{rr}}(\beta_{rr})\prod_{i=1}^{w} f_{\gamma_i|\sigma_\gamma}(\gamma_i)$$

$$\times\prod_{i=1}^{m} f_{\delta_r}(\delta_r)\prod_{r-1}^{m-1}\prod_{s=r+1}^{m} f_{\delta_{rs}|\delta_r,\delta_s}(\delta_{rs})\prod_{i=1}^{m} f_{\delta_{rr}|\delta_r}(\delta_{rr}) f_{\sigma_\gamma}(\sigma_\gamma) f_{\sigma_\epsilon}(\sigma_\epsilon).$$

To obtain an approximated distribution, we apply the Markov chain Monte Carlo method (Gelfand and Smith, 1990, Casella and George, 1992, and Chib and Greenberg, 1995). The Gibbs sampling introduced in the next section is a widely used algorithm for simulating Markov chains. The algorithm sequentially generates samples from the full conditional posterior distribution for each parameter conditioned on the current values of all other parameters.

The resulting values of the latent variables which have higher marginal or joint posterior probability are then used to identify active factors and select promising models for further consideration. This procedure is called the stochastic search variable selection (SSVS) by George and McCulloch (1993).

4. Examples: the SSVS method for SPDS designs

We apply the proposed SSVS method to analyze the data for the eight-factor SPSD design in Lin and Yang (2015). The SPDS design and the responses are listed in Table 1. This design was constructed by arranging a 17×8 definitive screening design into nine unbalanced whole plots ($n = 17, m = 8, \text{and } w = 9$), in which X_1 and X_2 are whole-plot factors and $X_3, \cdots, X_8$ are subplot factors. The responses were generated from the true model $E(Y) = 50 - 4X_1 + 3.5X_3 + X_1X_3 + 3.5X_1^2 - 4.5X_3^2$ and the covariance matrix $\mathbf{\Sigma} = \mathbf{UU}' + \mathbf{I}_n$. We choose the tuning parameters $c = 10$ and $\tau = .5$. To ensure that the selected models follow the effect heredity principle, we set $\rho_0 = \rho_{00} = .01$ and $\rho = \rho_1 = \rho_{11} = .75$ To reduce the uncertainty, we use $p = p_1 = p_{11} = .75$ (lower probability increases the model uncertainty). The Gibbs sampling with 11000 iterations are conducted by using the WinBUGS code. The algorithm starts with $\beta_0 = 0$ for the intercept, $\beta_i = 0$ and $\delta_i = 0$ for the fixed effects, $\gamma_i = 1$ for the random effects, and $\sigma_\gamma = \sigma_\epsilon = 1$ for the variances of the whole-plot and random errors. A burn-in of the first 1000 iterations is taken for eliminating the non-stationary portion of the chain at beginning and then only every 5th value from the Gibbs sampling

Table 1: The SPDS design and data in Lin and Yang (2015).

Whole plot	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	-1	-1	1	-1	1	0	-1	1	55.073
1	-1	-1	1	1	-1	1	0	-1	56.359
1	-1	-1	-1	1	1	-1	1	0	50.529
2	-1	0	-1	-1	-1	1	1	1	50.349
3	-1	1	0	1	-1	-1	-1	1	58.019
3	-1	1	-1	0	1	1	-1	-1	49.619
3	-1	1	1	-1	0	-1	1	-1	55.049
4	0	1	1	1	1	1	1	1	48.806
5	0	0	0	0	0	0	0	0	48.955
6	0	-1	-1	-1	-1	-1	-1	-1	42.708
7	1	-1	0	-1	1	1	1	-1	50.650
7	1	-1	1	0	-1	-1	1	1	51.752
7	1	-1	-1	1	0	1	-1	1	41.401
8	1	0	1	1	1	-1	-1	-1	48.219
9	1	1	-1	1	-1	0	1	-1	41.676
9	1	1	-1	-1	1	-1	0	1	42.943
9	1	1	1	-1	-1	1	-1	0	52.089

is retained for inference (this process is known as thinning) to decrease the autocorrelation between the iterations. Total 2000 samples for each parameter are saved and the distribution of the latent variables is calculated. In Table 2, we list the top five models which have highest joint posterior probabilities of the latent variables. It shows that the model with highest posterior probability has the same model terms with the true model and the other four models are all submodels of the true model. Figure 1 is a plot for the marginal posterior probability of each effect being active. It shows that the probabilities for effects $\beta_1, \beta_3, \beta_{13}, \beta_{11}, \text{and } \beta_{33}$ being active are significantly higher than the others. In Table 2, we also list the models selected by the two regression methods, MSR and FSR, with significance level α = .05. Results show that the MSR method correctly selects the active effects but the FSR method fails to identify the interaction of X_1 and X_3. The FSR method also falsely selects $X_5 \text{ and } X_2X_7$ into the model. With the FSR method, we obtain $\hat{\sigma}_\gamma = 0$, which implies that the FSR method cannot detect the split-plot structure for this example. In addition, the model selected by the FSR method disobeys the effect heredity principle.

5. Concluding remarks

In this paper, we propose the SSVS method to analyze data for SPDS and BDS designs. This method overcomes the problem of σˆγ = 0 with the REML method and the problem of being unable to estimate effects with the GLS method when the degrees of freedom are not

Table 2: Models selected by the SSVS, MSR, and FSR methods for the SPDS design.

Method	Model	Probability
SSVS	$X_1, X_3, X_1X_3, X_1^2, X_3^2$	.097
	X_1, X_3, X_1^2, X_3^2	.086
	X_3, X_3^2	.041
	X_1, X_3, X_1^2	.039
	X_1, X_3, X_1X_3, X_3^2	.038
MSR	$X_1, X_3, X_1X_3, X_1^2, X_3^2$	-
FSR	$X_1, X_3, X_5, X_2X_7, X_1^2, X_3^2$	-

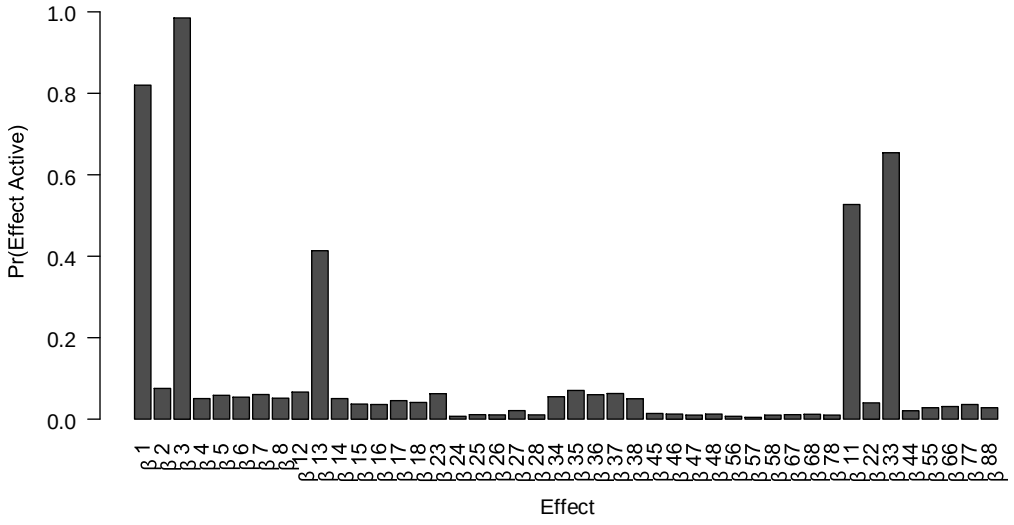


Figure 1: Marginal posterior probabilities of the effects being active for the SPDS design.

large enough. Different from the existing Bayesian methods for multi-stratum designs, the proposed method selects more reasonable models that follow the effect heredity principle. To make the proposed method easy to use, we conduct the MCMC and Gibbs sampling with the WinBUGS software and provide a general code that can be applied to analyze data for any SPDS and BDS designs. Applying the SSVS method, we reanalyze two datasets from the 17 x 8 SPDS design in Lin and Yang (2015) and the 15 x 4 BDS design in Jones and Nachtsheim (2016). To compare the performance of the SSVS method and the regression methods MSR and FSR, we conduct simulation studies for SPDS designs with strong effect heredity models and BDS designs with weak effect heredity models. Results show that the proposed SSVS method well controls the false discovery rate and has higher power than the two regression methods on identifying active effects. Although in this paper the SSVS method and the WinBUGS code are developed under the framework of the SPDS and BDS designs, they can be easily extended and applied to analyze data for any other multi-stratum deigns.

The impact of mis-specification of the deterministic components in the Co-integration model: An application to the bivariate case on South African employment costs and gross earnings

Sagaren Pillay
Statistics Finland

Abstract

This paper investigates the impact of different specifications of deterministic compontens in the vector error correction model (VECM) form estimated with Johansen's multivariate maximum likelihood approach. Using time series for employment costs and gross earnings data we show the impact of the misspecification of the deterministic components of the estimated Co-integration model. The study suggests that great care must be exercised in model specification. The inclusion or exclusion of the deterministic trend should be clearly justified to avoid misleading results.

Keywords

Deterministic; cointegration; trends

1. Introduction

Johansen's (1988) multivariate maximum likelihood approach to Co-integration is arguably the most popular approach in estimating long-run economic relationships. The main objective of Co-integration analysis is to determine the Co-integration rank of the model. Many studies have focused considerable attention on the modelling of economic relationship between variables, and variables to include in the model. Johansen (1995) has emphasized that the choice of the deterministic components of a model has important implications for the asymptotic distribution of the test statistics. There are five different ways that the deterministic components could be included in a Johansen Cointegration model. The choice of deterministic components included in the model has a very significant influence on the empirical results. It is therefore very important that, in actual application, the modelling of the deterministic components of Co-integration models needs to be carefully considered. Consider the Var(p) process without determinant terms. Suppose all individual variables are I(1) or I(0)

$$y = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \mu_t \quad (1)$$

The corresponding VECM may be written as

$$\Delta y_t = \Pi y_{t-1} + \phi \Delta y_{t-1} + \cdots \phi_{p-1} \Delta y_{t-p+1} + \mu_t \quad (2)$$

where $\Pi = -(I_k - A_1 - \ldots - A_p)$and $\phi = -(A_{k+1} + \ldots + A_p)$
Since all the variables are assumed I(1) or I(0), Πy_{t-1} must be I(0) and stationary.
Therefore det $(I_k - A_p Z^p) = 0$ which implies that $z = 1$ and Π is singular. Suppose Π has rank r :Let α and β be two k × r matrices of rank r Since ΠY_t is stationary, The linear transformation $(\alpha'\alpha)^{-1}\alpha'\Pi Y_t$ is also stationary

$$i.e(\alpha'\alpha)^{-1}\alpha'(\alpha\beta')y_t = (\alpha'\alpha)^{-1}(\alpha'\alpha)\beta' y_t = \beta' y_t$$

Hence the rows of β' are co-integrated vectors
Thus $\Delta y_t = \alpha\beta' y_{t-1} + \phi_1 \Delta y_{t-1} + \ldots + \phi_{t-p+1} + \mu_t$ (3)
After substituting $\pi = \alpha\beta'$in (2)

2. The deterministic components on a co-integration model

Following Hansen and Juselius (1995) the deterministic components can be modelled in five different ways. Case 1 does not allow for any deterministic components in the data. This is unlikely since for stochastic co-integrated variables in the Var model, deterministic terms should appear in the VECM via the error correction term or as independent terms. In Case 2 the model does not allow for a separate drift (constant) but allows for a constant in the Co-integration space. In Case 3, there are no trends in the Cointegration space; the linear trends enter the VECM as drift terms. In Case 4 there is a separate drift in the VECM and a linear trend in the Co-integration space. Finally, in Case 5, there is a separate linear trend in the VECM.

For the constants or trends in Johansen's Co-integration model to be meaningful, they must be related to the Co-integration space or excluded from it. The consequences of mis-specifying the deterministic components in empirical studies is not well known (Ahking, 2002).

3. Data: South African Gross Earnings and Employment costs (2009-2017)

The Quarterly Financial Statistics (QFS) survey produces employment costs estimates. These estimates are based on information as defined by the International Accounting Standards from an accounting perspective. The Quarterly employment statistics (QES) survey produces estimates for employment and earnings from a payroll perspective. Conceptually both gross earnings are closely related and follow similar trends with level differences. The similarities in the QFS employment costs and the QES gross earnings provide a good basis to investigate both the Co-integration as well as the mis-specification of the deterministic components in the Co-integration model.

When Constructing a VECM(p) from a Var(p) models, the deterministic terms in the VECM(p) model can differ from those in the Var(p) model. (1) When there are deterministic co-integrated relationships among variables-deterministic terms in the Var(p) model are not present in the VECM(p) form. (2) If there are stochastic co-integrated relationships in the Var(p) model, deterministic terms appear as the VECM(p) form in the EC term or as an independent term in the VECM(p) form.

4. Data Analysis

Based on the results of the of the Augmented Dickey Fuller (ADF) unit root test both series have a unit root and are first order difference stationary. The results show that both series are I (1) processes. To test for co-integration, Johansen's test was used. The maximum lag length was set to 7 quarters and an autoregressive order of p=2 was selected based on the minimum information criterion.Both series were found to be co-integrated with rank=1. The next step was to investigate the model specification for the 5 cases below.
Case 1: If there is no separate drift in the VECM(p) form then the model is given by:

$\Delta y_t = \alpha\beta' y_{t-1} + \phi_1 \Delta y_{t-1} + \ldots + \phi_{p-1}\Delta y_{t-p+1} + \mu_t$ from (2)

Case 2:

Suppose there is no separate drift in the VECM(p) form, but a constant V_0 enters only via the error correction term.
Consider the K-dimensional Var(p) process where $\mu_t$is K×1 and $y_t = \mu_t + x_t$
Suppose $\mu_t = \mu_0$ are fixed K-dimensional parameter vectors. Then

$$\begin{aligned}\Delta y_t &= \alpha\beta' y_{t-1} + \phi_1 \Delta y_{t-1} + \ldots + \phi_p \Delta y_{t-p+1} + \mu_t \\ &= \alpha\beta'(y_{t-1} - \mu_0) + \phi_1 \Delta y_{t-1} + \ldots + \phi_p \Delta y_{t-p+1} + \mu_t \\ &= \alpha\beta' y_{t-1} - \alpha\beta'\mu_0 + \phi_1 \Delta y_{t-1} + \ldots + \phi_p \Delta y_{t-p+1} \\ &= \alpha\beta' y_{t-1} + V_{0+}\phi_1 \Delta y_{t-1} + \ldots + \phi_p \Delta y_{t-p+1} + \mu_t\end{aligned}$$

where $V_0 = -\alpha\beta'\mu_0$ is the restriction of the intercept.

Case 3:

There is a separate drift δ_0 and no separate linear trend in the VECM(p) form the following model is used.

$$\Delta y_t = \alpha\beta' y_{t-1} + \phi_1 \Delta y_{t-1} + \ldots + \phi_p \Delta y_{t-p+1} + \mu_t$$

Case 4:

There is a separate drift and no separate linear trend in the VECM(p) form, but a linear trend enters only via the error correction term.
If $\mu_t = \mu_0 + \mu_1 t$ is a linear trend, we have $x_t = y_t - \mu_1$

$x_t = y_t - \mu_0 - \mu_1 t$ and $\Delta x_t = \Delta y_t - \Delta(\mu_0 - \mu_{t-1})$

Therefore from (3) $\Delta x_t = \alpha\beta' x_{t-1} + \phi_1 \Delta x_{t-1} + \ldots + \phi_{p-1}\Delta x_{t-p+1} + \mu_t$

Substitute $x_t = y_t - \mu_1; \Delta x_{t-1} = \Delta y_{t-1} - \mu_1$; and $x_t = y_t - \mu_0 - \mu_1 t$

Yields $\Delta y_t = \mu_1 + \alpha\beta' y_{t-1} - \alpha\beta'\mu_0 - \alpha\beta'\mu_1 t + \alpha\beta'\mu_1 + \phi_1 \Delta y_{t-1}$
$- \phi_1\mu_1 + \ldots + \phi_{p-1}\Delta y_{t-p+1} - \phi_{p-1}\Delta\mu_1 + \mu_t$
$= V + \Pi^* y^* + \ldots + \phi_{P-1}\Delta y_{t-p+1} + \mu_t$

with $V = -\Pi\mu_0 + \left(I_k - \phi_1 \ldots - \phi_{p-1}\right)$ and $\Pi^* = \alpha[\beta'; \beta'\mu_1]$ is K×(K+1) and $Y^* = \begin{bmatrix} y_{t-1} \\ t-1 \end{bmatrix}$

Note that in this model V is unristricted whilst the linear term can be absorbed in the error correction mechanism.

Case 5:

There is a separate linear trend in the VECM(p) form the model is:

$$\Delta y_t = \alpha\beta' y_{t-1} + \delta_0 + \delta_1 t + \phi_1 \Delta y_{t-1} + \ldots + \phi_{p-1}\Delta y_{t-p+1} + \mu_t$$

5. Estimations

Estimations for each of the five cases follows below.

Case 1: No deterministic components

Table 1

Co-integration Rank Test Using Trace						
H0: Rank=r	H1: Rank>r	Eigenvalue	Trace	Pr > Trace	Drift in ECM	Drift in Process
0	0	0,4360	19,175	0,0032	NOINT	Constant
1	1	0,0847	2,5655	0,1291		

Table 1 shows the results of the "Co-integration Rank Test Using Trace" where no intercept term is assumed,for a 5% significance level H0: Rank=0 can be rejected because p< 0.05 and H0:Rank=1 cannot be rejected because p> 0.05 therefore both series are Co-intergrated.

The model excludes all deterministic components in the data, implying no growth and zero intercepts in every co-integrating relation. From the given data it is evident that an intercept is needed to account for the initial level of measurements Y_0 . Thus, this option is clearly not justified.

$$\Delta y_t = \begin{bmatrix} -0.21684 & 0.22084 \\ -0.24048 & 0.24492 \end{bmatrix} y_{t-1} \begin{bmatrix} -0.62327 & 0.34321 \\ -0.39214 & -0.2925 \end{bmatrix} \Delta y_{t-1} + \varepsilon_t$$

$$\beta' = [1 \quad -0.101846] \quad \alpha = \begin{bmatrix} -0.21648 \\ -24048 \end{bmatrix}$$

$$\text{Therefore } \alpha\beta' = \begin{bmatrix} -0.21684 \\ -0.24048 \end{bmatrix} [1 \quad -0.101846] \text{ and } \beta' y_t$$
$$= [1 \quad -0.101846] \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$
$$= y_t - 0.101846\, y_{1t} = 1.101548\, y_{2t}$$

Case 2: No separate drift in the VECM but a constant enters the Co-integration space.

Table 2

Co-integration Rank Test Using Trace Under Restriction						
H0: Rank=r	H1: Rank>r	Eigenvalu	Trace	Pr >	Drift in ECM	Drift in Process
0	0	0,4369	24,032	0,0142	Constant	Constant
1	1	0,2246	7,3754	0,1079		

Table 3

Hypothesis Test of the Restriction					
Rank	Eigenvalue	Restricted Eigenvalue	DF	Chi-Square	Pr > ChiSq
0	0,2259	0,4369	2	16,00	0,0003
1	0,0206	0,2246	1	6,77	0,0093

Here H_0 :Rank=0 can be rejected and H0:Rank=1 cannot be rejected implying that the series are Cointegrated. Further the p-values (Table 3) are less than 0.05 showing that the model with restriction on the intercept term is appropriate.

The model does not allow for any linear deterministic trends in the data. The only deterministic components in the model are the intercepts in any co-integrating relations, implying that some equilibrium means are non-zero. This model has a restricted intercept term.

$$\Delta y_t = \begin{bmatrix} -0.07434 & 0.05796 & 0.24894 \\ -0.0835 & 0.0651 & 0.2796 \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \\ 1 \end{bmatrix} + \begin{bmatrix} -0.65912 & 0.41718 \\ -0.43642 & -0.21037 \end{bmatrix} \Delta y_{t-1} + \varepsilon_t$$

$$\beta' = [1 \quad -0.77968 \quad -3.34869] \quad \alpha = \begin{bmatrix} -0.07434 \\ -0.0835 \end{bmatrix}$$

$$\text{Therefore } \alpha\beta' = \begin{bmatrix} -0.07434 \\ -0.0835 \end{bmatrix} [1 \quad -0.77968 \quad -3.34869]$$

$$\Pi = \begin{bmatrix} -0.07434 & 0.05796 & 0.24894 \\ -0.0835 & 0.0651 & 0.2796 \end{bmatrix}$$

$$y_t = [1 \quad -0.77968 \quad -3.34869] \begin{bmatrix} y_1 \\ y_2 \\ 1 \end{bmatrix}$$

$$= y_{1t} - 0.77968 y_{2t} - 3.34869 \text{ and } y_{1t} = 0.779868 y_{2t} + 3.34869$$

Case 3: There are no trends in the Co-integration space; the linear terms enter the VECM as drift terms.

Table 4

Co-integration Rank Test Using Trace						
H0: Rank=r	H1: Rank>r	Eigenvalue	Trace	Pr >	Drift in ECM	Drift in Process
0	0	0,2259	8,028	0,4625	Constant	Linear
1	1	0,0206	0,603	0,4370		

Table 4 shows the results of the "Co-integration Rank test using trace" where there is no trends in the Co-integration space and VECM has drift terms .H_0: Rank=0 cannot be rejected

Since the drift term is unrestricted, there are still linear trends in the data, but no deterministic trends in any Co-integration relations. In this application a non-zero unrestricted intercept and no trends in the co-integrating relationship may be plausible.

Case 4: There is a separate drift in the VECM and a linear trend in the Co-integration space.

The trend is restricted to lie in the Co-integration space, but the drift term is unrestricted in the VECM.

This model may be used to test for trend stationarity.

Table 5

Co-integration Rank Test Using Trace Under Restriction						
H0: Rank=r	H1: Rank>r	Eigenvalu	Trace	Pr >	Drift in ECM	Drift in Process
0	0	0,4219	22,3309	0,1286	Linear	Linear
1	1	0,1992	6,4410	0,4060		

Table 6

Hypothesis Test of the Restriction					
Rank	Eigenvalue	Restricted Eigenvalue	DF	Chi-Square	Pr > ChiSq
0	0,4135	0,4219	2	0,43	0,8058
1	0,1987	0,1992	1	0,02	0,9016

Clearly (from Table 5) H0: Rank=0 can be rejected at the 15% significance level whilst,H1 cannot be rejected, there is evidence of Co-integration with restriction on the trend.The model may not be appropriate i.e given the p-values in Table 6.

Case 5: No restrictions on the trend and intercept in the VAR model.

With unrestricted parameters the model is consistent with linear trends in the differenced series Δy_t and thus quadratic trends in y_t. In this case, estimation and inference may be unreliable (Doornik, Hendry, and Nielson, 1998).

Table 7

Co-integration Rank Test Using Trace						
H0: Rank=r	H1: Rank>r	Eigenvalue	Trace	Pr > Trace	Drift in ECM	Drift in Process
0	0	0,4135	21,899	0,0156	Linear	Quadratic
1	1	0,1987	6,4257	0,0114		

From Table 7 above, it is clear that for a linear drift in the VECM there is a quadratic drift in the Co-integration space.Further the series is not co-integrated implying that the model is inappropriate.

6. Conclusions

It is important to give considerable attention to the deterministic components of the models when modelling long run relationships. The estimation results for each of the five cases show theimportance of the choice of the deterministic components in the cointegration model. For each of the five cases, based on the choice of the deterministic components, the results are significantly different.

When the model excludes all deterministic components in the data (case 1), implying no growth and zero intercepts in every cointegrating relation, favourable cointegration results are obtained. However,from the given data it is evident that an intercept is needed to account for the initial level of measurements Y_0. Consequently, this option with the long run relationship given by $y1_t = 1.101548\, y2_t$ cannot be justified. On the other hand, when the model does not allow for a separate drift (constant) but allows for a constant in the cointegration space (case 2) we obtain favourable cointegration results with a restriction on the intercept term. This option is with the long relationship, $y1_t = 0.779868\, y2_t + 3.34869$ is clearly approriate. For cases 3 ,4 and 5, the eviedence in favour of cointegration is extremely weak.

Given the different results for each of the cases for this empirical application, we recommend that extreme care should be taken when specifying deteministic components during cointegration modelling.

References

1. Johansen, S., 1995. Likelihood-Based Inference in Cointegrated Vector Auto-regressive Models. Oxford University Press, Inc., New York.
2. Hansen, H., Juselius, K., 1995. Cats in rats: co-integration analysis of time series. Distributed by Estima,Evanston, IL
3. Doornik, J.A., Nielsen, B. and Hendry, D.F. (1998). Inference in Cointegrating Models: UK M1 Revisited, Journal of Economic Surveys, 12, 533-572.
4. Ahking,F.W., 2002. Model mis-specification and Johansen's co-integration analysis: an application to the US money demand. Journal of Macroeconomics, 24, 51-56.

A critical image of statistical analyses in medicine between 2006 and 2018

Ordak Michal
Department of Pharmacodynamics, Centre for Preclinical Research and Technology (CePT), Medical University of Warsaw, Poland

Abstract

Statistics play an extremely important role in medicine. Over a period of 13 years, i.e. between 2006 and 2018, I observed a negative trend in medical statistics. At that time, I examined a group of 14,000 people working in the medical environment, including doctors and researchers. The paper presents the basic mistakes made by the medical community in the performance of statistical analysis. For this purpose, I analyzed additional 30 publications from 50 medical journals (n = 1500). During the last thirteen years, i.e. between 2006 and 2018, I carried out 25,300 statistical consultations in a group of students of various medical majors. Only 40% of the articles analyzed contained correct statistical analyses. The most common errors I have noticed include: no description or the use of wrong statistical tests (assumptions of parametric equivalents are not met) and an incorrect record of results obtained. During the last thirteen years, I carried out 25,300 statistical consultations in a group of students of various medical majors. Almost 95% of respondents said that their insufficient knowledge of statistical analysis results from four factors. What education path should be chosen to increase the quality of statistical analysis performed by the future medical community? This paper answers these questions.

Keywords

Biostatistics; Medical statistics; Statistics education

1. Introduction

In this age of modern technologies and development, medicine is one of the most dynamic areas of life. Statistics play an extremely important role in medicine. They are a source of knowledge about people and their lives. They should be a 'signpost' for all medical employees. The thousands of statistical analyses and reviews that I carried out between 2006 and 2018 have inspired me to write this abstract. Nowadays, due to many factors, people performing statistical analyses in medicine do not really realise how many significant errors they make. This is connected with the pressure to publish results in high-ranked journals. Consequently, many medical researchers break statistical rules in order to obtain statistically significant results. I would, therefore, like to provide

the medical research community with my research results that are based on 13 years of experience in conducting statistical analyses and reviews in everyday practice. My intention is to make the global medical community aware of why there are reasons for concern when it comes to the direction in which medical statistical analysis is going.

"Some people hate the very name of statistics but I find them full of beauty and interest. Whenever they are not brutalised, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man" (Galton, 1986). I often say these words to students of various medical fields to introduce them to the interesting world of medical statistics. Statistics plays a very important role at every stage of scientific biomedical research. Data analysis is used not only by scientists and physicians directly involved in clinical trials, but also by medical employees who keep up to date with the results of new research and want to interpret their results correctly (Zelen, 2006). Unfortunately, nowadays knowledge of basic statistical methods is negligible in the medical environment. Medical students unfortunately do not have full awareness of the importance and necessity of applying appropriate statistical methods in their research. Consequently, students are often unable to perform statistical analysis for publications, dissertations, etc. (Altman et al., 2002). Medical reviewers therefore assess submitted publications primarily in terms of innovation and very often do not evaluate the correctness of statistical analysis (Ozonoff, 2006; Petrovečki, 2009). As a long-time expert in statistics, I must unfortunately state that a large number of medical reviewers accept submitted works even in very high-ranked magazines despite significant statistical errors.

2. Methodology

I examined a group of 14,000 people working in the medical environment, including doctors and researchers. The article presents the basic mistakes made by the medical community in the performance of statistical analysis. For this purpose, I analysed additional 30 publications from 50 medical journals. During the last thirteen years, i.e. between 2006 and 2018, I carried out 25,300 statistical consultations in a group of students of various medical majors.

3. Results

Repeated testing of means is a growing problem today. Increasingly, more medical researchers use t-tests for research schemes with one independent variable on more than two levels. Making multiple comparisons by testing the difference between two means within one dependent variable and a multilevel independent variable leads to false conclusions. This phenomenon is called type

I cumulative probability. The more independent comparisons of pairs of levels of a given independent variable are made, the greater the risk that the statistically significant result obtained is a matter of chance. If many tests are carried out, each of which is affected by an error (e.g. p=0.05), the type I cumulative error can be very large. To avoid this, an analysis of variance should be performed. When a statistically significant result of the F test is obtained from the analysis of variance, explanatory analyses are performed, i.e. a posteriori (post-hoc) or a priori (contrasts). This makes it possible to determine which pairs of levels of a given factor have statistically significant differences (McHugh, 2011). A large number of people performing statistical analyses in medicine do not understand why they often do not obtain statistically significant differences when they perform explanatory analyses, but the situation is just the opposite when they conduct a series of t-tests. The great number of statistical analysis and reviews that I carried out between 2006 and 2018 led to one conclusion. Namely, there was a tendency to publish results that could be type 1 cumulative errors. Researchers admitted that they performed a series of over a dozen t-tests in order to obtain statistically significant results instead of performing explanatory analyses. This was to increase the chances of their articles being accepted for publication in medical journals. My long term research in a group of 14,000 researchers (including physicians, graduate students, PhD students, PhDs and professors) in various fields of medicine has allowed me to state that we have reasons for concern. Namely, as many as 76% of respondents stated that they did not know what type 1 cumulative error is; 46% of people admitted that they often performed several or over a dozen t-tests instead of conducting an analysis of variance. While 10% of them did so due to a lack of knowledge, others wanted to increase the chance of obtaining a statistically significant result. Moreover, 52% of respondents chose a wrong post-hoc test. This is one of the reasons why I have written this article. Its aim is to make medical researchers aware of the very important problem, repeated testing of means, and thus obtaining results that are a matter of chance.

Another problem that I observed between 2006 and 2018 is the growing lack of knowledge among people performing statistical analyses in medicine regarding the use of parametric or non-parametric equivalents of statistical tests used. The high relevance of parametric tests is related to numerous assumptions that should be met (Fagerland, 2012). One of the most common questions researchers ask is: 'Can I use a parametric equivalent if the assumption about the normality of distribution is not met?' Unfortunately, my research indicates that many of the assumptions are poorly known and rarely used. Although assumptions are not met, numerous authors use parametric equivalents of statistical tests in order to increase the chances for their articles to be accepted in peer-reviewed journals. Unfortunately, this is a big mistake. One such assumption concerns the equinumerosity of groups studied and the

homogeneity of variance. Almost 62% of respondents do not use statistical tests to examine the equinumerosity of groups compared, but guesstimate them instead. Another example is the heterogeneity of variance that may be related to the occurrence of outliers, zero variance from one of the groups, or it may be the result of an impact of an independent variable not considered in the analysis. As many as 49% of respondents do not take account of the assumption about the homogeneity of variance. As many as 69% people surveyed do not know the names of non-parametric equivalents of the statistical tests used, such as the Friedman test.

During the many-year review of thousands of various scientific works, such as publications, projects, grants, etc., I observed many very significant statistical errors. They have a significant impact on the scientific value of published research results. This is the reason for frequent discrepancies in the results obtained in specific areas. When assessing submitted publications, reviewers evaluate them primarily in terms of innovation. As an experienced expert in statistics, I must state that numerous reviewers, even in very high-ranked journals, accept submitted works despite significant statistical errors. The problem is the lack of professionals who can review submitted works in terms of statistics. A significant number of journals included in the ISI Master Journal List do not ask statistical reviewers for help (Petrovečki, 2009). In order to illustrate this problem, I assessed the correctness of statistical analyses carried out in journals included in the ISI Master Journal List. I analysed 30 publications from 50 medical journals (n=1500). Table 1, below, contains basic statistical errors made by the medical community between 2006 and 2018. Only 40% of the articles analysed contained correct statistical analyses. The most common errors I have noticed include: no description or the use of wrong statistical tests (assumptions of parametric equivalents are not met) and an incorrect record of results obtained. Type I cumulative error can also be included here.

Comments on the statistical analysis of results	**Example**
Results are not recorded according to standards.	r=0.36 (significance is not provided)
Results are not separated by spaces.	t(56)=5.69;p<0.05
No statistical tests are used.	Only numbers are described and appropriate statistical tests are not used.
The statistical tests used are either roughly described or not described at all.	Only the statistical package is described without the tests used.

	Sentences like: 'The study is conducted using a t-test and an analysis of variance'.
Wrong statistical tests are used.	T-test vs explanatory analyses. The type of measuring scale is not taken into account when choosing a test.
Parametric equivalents of tests are performed even though assumptions are not met.	Non-equinumerous groups, i.e. the experimental group and the control group Assumptions of regression analysis, such as the correlation of predictors, are not tested.
There is no detailed description of the results obtained.	A chi-square relationship is generally described without detailed examination and description.
No validation of instruments and questionnaires	No calculated Cronbach's alpha coefficient
No valid explanation of outliers	No including outliers in regression
Repeating the results of the analysis several times	The same data in the form of tables, graphs
No explanation of the changes in the number of subjects	Missing data
Wrong conclusions drawn on the basis of the analyzes	Drawing conclusions, despite the use of incorrect statistical Tests
Others	No data on the recruitment of participants, inadequate sample size, unreadable list of variables, no clear baseline demographic and clinical parameters

Table 1 : Basic statistical errors made by authors in medical journals included in the ISI Master Journals List between 2006 and 2018 (n=1500 publications).

During the last thirteen years, i.e. between 2006 and 2018, I carried out 25,300 statistical consultations in a group of students of various medical majors. Almost 95% of respondents said that their insufficient knowledge of statistical analysis result from four factors, namely:

- Exercises in medical statistics are conducted too quickly

Almost 95% of respondents stated that their insufficient knowledge about the use of statistics in medical sciences resulted from the rapid conduct of exercises during studies. Individual statistical tests are performed too quickly.

Instead of focusing on the use of specific statistical tests, students of medical faculties focus primarily on how to perform tests step-by-step. The time that should be devoted to becoming familiar with the application of a specific statistical test is spent on writing down in stress long procedures of performing individual statistical tests, checking assumptions, etc. This results, unfortunately, in the lack of knowledge of the subject of the course.

- Practical application of individual statistical tests is not sufficiently demonstrated

Respondents also stated that too little emphasis is put on the practical application of individual statistical tests, e.g. in the form of results published in medical journals. Greater pressure is put on details related to conducting individual tests using statistical packages. After mentioning in a few sentences what a specific statistical test is used for, teachers go straight to a series of activities that are difficult to note and remember in a short time.

- A long break between classes in medical statistics and the performance of statistical analysis for the purposes of M.A. theses, doctoral dissertations, publications, etc.

A significant number of respondents also noticed that the break between classes in medical statistics and the necessity to perform statistical analysis of the obtained results for the purposes of M.A. theses, doctoral dissertations, publications, etc. was too long. Very often, classes in medical statistics are conducted in the first year of various medical faculties, which only exacerbates problems related to statistical analysis in subsequent years. Students do not remember statistical analyses performed during classes as they are conducted too quickly. A significantly larger number of students ask a third person for help in performing the statistical analysis of the results obtained. Usually, this happens a few years after classes in medical statistics, which are conducted during the first or second year of study.

- Increased stress during the exam in medical statistics associated with memorising hundreds of activities related to statistical testing.

A significantly larger number of respondents stated that instead of studying the application of individual statistical tests to a greater extent, they have problems remembering the long procedures of performing the analysis. The practical aspect of the use of statistical tests is of secondary importance. After the exam in medical statistics, stress is mainly evoked by the uncertainty of whether a statistical package has been successfully used. Despite the fact that there are more and more scripts describing particular statistical tests, according to the surveyed students, the majority of them are too detailed. It is necessary to develop simple scripts showing step by step methods of statistical analysis, along with the interpretation of the results obtained.

What education path should be chosen to increase the quality of statistical analysis performed by the future medical community?

It is necessary to implement an appropriate method allowing students to obtain practical knowledge necessary to conduct statistical analysis. Effective education requires taking account of several factors. I have drawn the following conclusions after conducting several thousand statistical consultations. Thanks to this type of education, the knowledge of medical staff regarding medical statistics has grown rapidly.

Let us suppose that a teacher deals with a specific statistical test, e.g. a Student's t-test for independent samples. He or she explains when to apply this test by referring to students' interests. Then, he or she demonstrates the practical aspect of this statistical test based on a specific publication in a medical journal. An example here can be a work published in The Nature Medicine in which the authors compared two independent groups: one was continued on a doxycycline diet and the other one was fed a normal-chow diet (Beckerman et al., 2017). The next step is to perform this statistical test using a specific scheme. It should include a step-by-step description of the methods of conducting the test (and meeting the assumptions) and interpreting the results obtained. If classes are carried out in this way, the knowledge of medical students will increase by almost 100%. There is no stress associated with the rapid implementation of individual steps necessary to conduct a test.

Another recommendation that would reduce problems associated with the performance of statistical analyses by students for thesis, publications, etc., is to conduct classes in medical statistics in later years of study. Due to a few-year break in contact with statistics, a significantly larger number of students do not even know how to get down to it. Conducting classes in medical statistics in the last year of study would reduce the percentage of people having problems with analysing statistical results. Students would have an additional motivation that would encourage them to participate even more in the classes. Thanks to this, the quality of conducted research necessary, for example, to write a diploma thesis would significantly increase. Freshly acquired knowledge could be used in a practical way in the further course of study.

My last advice concerns the method of conducting a medical statistics exam. The best way would be to use proper schemes during the exam containing a step-by-step description of how individual statistical tests are carried out. The exam should involve reading several research problems, formulating hypotheses and selecting the appropriate statistical test. Then, based on the scheme, the selected statistical test would be carried out and the data would be interpreted. This would save the main problem of remembering the great number of procedures necessary for conducting statistical analysis.

4. Discussion and Conclusion

Summing up the research carried out between 2006 and 2018, I can say that there are significant reasons for concern. Every year, more and more errors are observed in medical statistics. This is often connected with the lack of statistical knowledge among medical researchers and the tendency to break statistical rules in order to obtain a specific result. Taking account of the factors described in this article will significantly increase the value of published medical research results in the future. For example, the percentage of publications presenting ambiguous results on particular topics will decrease. The purpose of this article is to make the community around the world aware of the most common errors made in medical statistics. These errors can often affect the most precious gift, which is our life. The statistical consultations carried out between 2006 and 2018 allowed me to conclude why the statistical knowledge of medical students is insufficient. This translates into many negative aspects in the future, such as the lack of statistical knowledge among medical reviewers, problems encountered by e.g. doctoral students and physicians when performing statistical analyses, as well as many other problems. It seems advisable to implement the recommendations described here in the field of medical statistics education. Consequently, the percentage of medical employees for whom statistical analysis will not be something completely new and who will be able to use their own knowledge in practice will increase in the future.

References

1. Altman, D. G., Goodman, S.N. & Schroter, S. (2002). How statistical expertise is used in medical research. JAMA. 287(21):2817-20.
2. Beckerman, P., Bi-Karchin, J., Park, A. S., Qiu, C., Dummer P. D., Soomro, I. et al. (2017). Transgenic expression of human APOL1 risk variants in podocytes induces kidney disease in mice. Nat. Med. 23(4): 429-38.
3. Fagerland, M. W. (2012). T-tests, non-parametric tests, and large studies—a paradox of statistical practice? BMC. Med. Res. Methodol.12:78.
4. Galton, F. (1986). Natural inheritance; London:Macmillan.
5. McHugh. M. L. (2011). Multiple comparison analysis testing in ANOVA. Biochem. Med. 21:203-9.
6. Ozonoff, D. (2006). Statistics in peer review. Nature's peer review debate. Nature.
7. Petrovečki, M. (2009). The role of statistical reviewer in biomedical scientific journal. Biochem. Med. 19(3):223-30.heir insufficient knowledge of statistical analysis results from four factors, namely:
8. Zelen, M. (2006). Biostatisticians, Biostatistical Science and the Future. Stat. Med. 25:3409-14.

Annualization of the Labor Force Survey in the Philippines

Quindale E. Caraos[1], Sarah B. Balagbis[1], Abubakar S. Asaad[1,2], Divina Gracia L. del Prado[3], Erniel B. Barrios[4]
[1]Statistical Methodology Unit, Philippine Statistics Authority
[2]College of Public Health, University of the Philippines Manila
[3]Economic Sectoral Statistics Section, Philippine Statistics Authority
[4]School of Statistics, University of the Philippines Diliman

Abstract
The 2016 Labor Force Survey (LFS) starting April round used the 2013 Master Sample (MS) design. The 2013 MS was developed to produce reliable estimates at the provincial/highly urbanized city (HUC) levels and rates at the end of the year. With this new MS, there is a need to develop a new estimation procedure for annualized provincial estimates. Annualized provincial estimates of the LFS using the simple averaging method and bootstrap method were explored. The bootstrap method (i.e., the weighted average of the four bootstrap quarter estimates) is better compared to the simple averaging method since it generally yields more reliable estimates.

Keywords
annualized, bootstrap, simple averaging, indicators, master sample

1. Introduction

The LFS is a nationwide quarterly survey conducted by the Philippine Statistics Authority (PSA) during the months of January, April, July and October. The survey was designed to provide statistics on levels and trends of the key labor and employment indicators in the country and for each of the administrative regions using 2003 Master Sample (MS)1. A stratified, three-stage sampling design was employed in the 2003 MS: the selection of primary sampling units (PSUs) for the first stage, sample enumeration areas (EAs) for the second stage, and sampling units for the third stage. The regions were used as domains2. The clamor of the local government in all provinces in the country is to come up with provincial estimates but cannot be answered by LFS which uses the 2003 MS. With the results of the 2010 Census of Population and Housing (CPH 2010)3, the former 2003 MS was redesigned resulting to the 2013 MS. A total of 117 major domains were considered consisting of 81 provinces (including the newly created province of Davao Occidental), 33 HUCs (including 16 cities in the National Capital Region), and 3 other areas (Pateros, Isabela City and Cotabato City)4. This 2013 MS was developed to provide reliable estimates not only at the national and regional levels but also

at the provincial/highly urbanized city (HUC) levels at the end of the year. Along with this new MS, a new estimation procedure for annualized provincial estimates was explored. This paper aims to develop an estimation procedure for the annualized provincial estimates of the LFS.

2. Methodology

The data sets were checked and inspected for completeness and lonely PSUs. Lonely PSUs were collapsed since at least two responding sample housing units per PSU were required to have a valid variance. Recomputation of SSU weights were also performed for these collapsed PSUs. Estimation of annualized totals and rates at the provincial level was simplified by merging the provincial data from all replicates for each quarters into one set of data for the province/HUC to get rid of the individual estimation of replicate levels and variance. This modifications did not affect the derived estimators for totals and rates but there is a slight change in variance of totals. The estimation procedure takes into account the design of the master sample and the selection of sample housing units for the LFS. The procedure entails two major steps: computation of survey weights and estimation of population parameters.

Survey weights are used as raising factor for sample data to produce estimates for population parameters such as population total, mean and rate. These weights compensate for the unequal selection probabilities in the survey design, for nonresponse and for noncoverage. The weights for the LFS using the 2013 MS design were developed in three stages: computation of base weights, adjustment of the base weights to take into account unit nonresponse and population weighting adjustment to make population counts conform to the population projection.

2.1 Simple Averaging Method

Estimate and variance for the annualized provincial total,

$$\hat{Y}_p = \frac{1}{4}\sum_{t=1}^{4} \hat{Y}_{pt}, t = 1,2,3,4\ (quarter), \hat{V}\left(\hat{Y}_p\right) = \frac{1}{16}\sum_{t=1}^{4} \hat{V}\hat{Y}_{pt},$$

where: $\hat{Y}_p \equiv$ annualized estimate of Y for province $p, \hat{Y}_{pt} \equiv$ estimate of Y for quarter t in province p. The **rate** and its corresponding variance is:

$$\hat{R}_p = \frac{\hat{Y}_P}{\hat{X}_P}, \qquad \hat{V}(\hat{R}_P) \approx \frac{1}{\hat{X}_P^2}\{\hat{(}\hat{Y}_P) + \hat{R}_P^2\hat{V}\left(\hat{X}_P\right) - 2\hat{R}_{Pc}(\hat{Y}_P, \hat{X}_P)\}$$

where: $\hat{X}_{P \equiv}$ estimated total of X for province p.

2.2 Bootstrap Method

For each indicator in each province and in each quarter, bootstrap estimate and corresponding variance were computed by following the steps below:

1. Draw a simple random sample of size n with replacement (n =equal to the total number of households in the province).

2. Compute the mean $\hat{R}^{(b)} = \frac{1}{n}\sum_{k=1}^{n} \hat{R}_k$, where $\hat{R}^{(b)}$ is the estimate of an indicator in resample b, n is the total number of households in a province and $\hat{R}_K$ is the estimate of an indicator in household k.
3. Repeat (A) and (B) 500 times.
4. Compute the bootstrap estimate of an indicator and the corresponding bootstrap estimate of variance. The bootstrap estimate of an indicator, $\hat{R}_{BS} = \frac{\sum_{b=1}^{B} \hat{R}^{(b)}}{B} = \frac{\sum_{b=1}^{500} \hat{R}^{(b)}}{500}$, $\hat{R}_{BS} \equiv$ bootstrap estimate of an indicator, $\hat{R}^{(b)} \equiv$estimate of an indicator in resample b, and $B \equiv$ total number of resamples $(B = 500)$. Note that $\hat{R}_{BS}$ is just the average of the 500 means computed in steps (2) and (3). The bootstrap estimate of variance was computed as:

$$\sigma_{BS}^2 = \frac{\sum_{b=1}^{B}(\hat{R}^{(b)} - \hat{R}_{BS})^2}{B} = \frac{\sum_{b=1}^{500}(\hat{R}^{(b)} - \hat{R}_{BS})^2}{500 - 1}$$

where: $\sigma_{BS}^2 2 \equiv$ bootstrap estimate of variance, $\hat{R}_{BS} \equiv$ bootstrap estimate of an indicator, $\hat{R}^{(b)} \equiv$estimate of an indicator in resample b, and $B \equiv$ total number of resamples $(B = 500)$. The annualized provincial totals was computed as the weighted average of the four quarters, that is, $\hat{Y}_p = \sum_{t=1}^{4} W_{pt}\hat{Y}_{pt},\ t = 1,2,3,4\ (quarter)$ where: $\hat{Y}_p \equiv$ estimate of Y for province $p, \hat{Y}_{pt} \equiv$ estimate of Y for quarter t in province $p, W_{pt} = \frac{\frac{1}{\hat{\sigma}_{pt}}}{\left(\frac{1}{\hat{\sigma}_{p1}}+\frac{1}{\hat{\sigma}_{p2}}+\frac{1}{\hat{\sigma}_{p3}}+\frac{1}{\hat{\sigma}_{p4}}\right)} \equiv$ weight of quarter t in province $p, \hat{\sigma}_{pt} \equiv$ bootstrap standard error of Y for quarter t in province p. Its variance, $V(\hat{Y}_p) = W_{p1}^2\hat{\sigma}^2(\hat{Y}_{p1}) + W_{p2}^2\hat{\sigma}^2(\hat{Y}_{p2}) + W_{p3}^2\hat{\sigma}^2(\hat{Y}_{p3}) + W_{p4}^2\hat{\sigma}^2(\hat{Y}_{p4})$.

2.3 Measure of Precision

The precision of the annualized provincial estimates for totals and rates were evaluated using the coecient of variation. Estimates with coecient of variations which are less than or equal to 10% implies precision, and hence, reliable estimates.

3. Results

3.1 Simple Averaging Method

Table 1 shows the number of provinces/HUCs by LFS Indicators (Levels) and CVs using simple averaging method. For *employment*, 91.4 percent of the provinces in the country have CVs less than 10% and the remaining 8.6 percent of the provinces have CVs between 10% and 20%. On the other hand, 26 cities have CVs less than 10%, six (6) have CVs between 10% and 20% and one (1) city with CV more than 20%. For *unemployment*, 7 out of 10 (or 70%) of the provinces have CVs less than 10%, 23.5 percent of the provinces with CVs between 10% and 20%, and the remaining five (5) provinces with CVs larger

than 20%. As to HUCs, 15 cities have CVs lower than 10%, 17 with CVs between 10% and 20% and one (1) city with CV greater than 20%. For *underemployment*, 96.3 percent of the provinces and 69.7 percent of the HUCs with CVs below 10%. For *visibly underemployment*, 59 of the 81 provinces and 10 of the 33 HUCs have CVs lower than 10%. For *labor force*, 74 out of 81 provinces and 25 of the 33 HUCs have registered with CVs lower than 10%. A similar observation was noticed for *not in the labor force*. For *population 15 years old and over, less OCW*, 80 of the 81 provinces and all HUCs have CVs below or equal to 10%. The boxplots in Figure 1, show the CVs of provincial LFS indicators (Levels). For all indicators, the CVs ranges from 0.21% to 68.97%. More than 75% of the provincial employment, underemployment, total labor force, not in the labor force level estimates are reliable while almost all of the provincial total population 15 years old and over less OCW level estimates are reliable.

Table 1: Number of Provinces/HUCs by LFS Indicators (Levels) and CVs using Simple Averaging Method

Indicators	CVs (%) of Provinces			CVs (%) of HUCs		
	< 10	(10, 20]	> 20	< 10	(10, 20]	> 20
Employment	74	7		26	6	1
Unemployment	57	19	5	15	17	1
Underemployment	78	2	1	23	8	2
Visibly Underemployment	59	20	2	10	18	5
Total Labor Force	74	7		25	7	1
Not in the Labor Force	74	7		26	6	1
Popn 15 years old &	80	1		33		

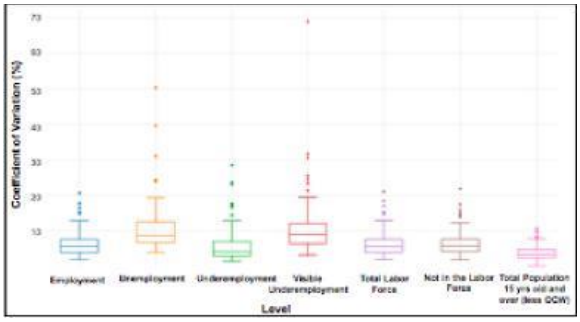


Figure 1: CVs of LFS Indicators (Levels)

Table 2 presents the number of provinces, HUCs by LFS Indicators (Rates) and CVs using simple averaging method. For *employment rate* (ER) or the proportion of the total employed persons to the total labor force is 82.7 percent of the provinces have noticed reliable CVs but in the case of HUCs only 39.4 have CVs less than or equal to 10%. Almost 50.0 percent of the provinces in the country and only 24.2 percent of the HUCs have CVs lower than or equal to 10% for the proportion of total unemployed persons to the

total labor force or *Unemployment Rate* (UR). For the *Underemployment Rate* (UndER) or the proportion of total underemployed persons to the total employed persons and for the *Labor Force Participation Rate* (LFPR) or the proportion of total labor force to the total household population 15 years old and over, similar results were obtained, that is, 70 out of 81 provinces have registered CVs less than or equal to 10% and 46 out of 81 provinces for the proportion of the total visibly underemployed persons to the total employed persons or for *Visibly Underemployment Rate* (VUndER). The CVs of provincial LFS indicators (Rates) is shown in Figure 2. For all indicators, the CVs ranges from 2.30% to 23.24%. Around 75% of the provincial employment and underemployment rates are reliable while more than 75% for the provincial LFPR estimates are reliable.

Table 2: Number of Provinces/HUCs by LFS Indicators (Rates) and CVs using Simple Averaging Method

Indicators	CVs (%) of Provinces			CVs (%) of HUCs		
	< 10	(10, 20]	> 20	< 10	(10,20]	> 20
ER	67	11	3	13	18	2
UR	40	34	7	8	21	4
UndR	70	9	2	11	17	5
VUndR	46	29	6	6	20	7
LFPR	70	9	2	21	12	

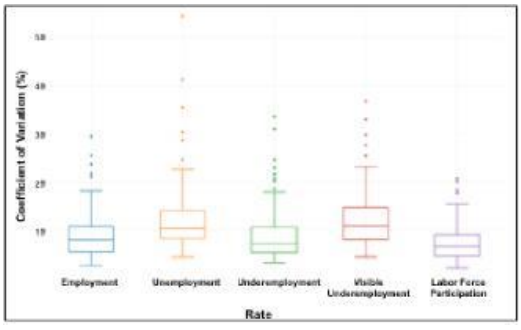


Figure 2: CVs of LFS Indicators (Rates)

a. Bootstrap Method

Table 3 presents the number of provinces and HUCs by LFS Indicators (Levels) using bootstrap method. For the LFS indicators such as *employment, labor force, not in the labor force,* and *population 15 years old and over, less OCW,* indicated that all provinces and HUCs were found to be reliable. Figure 3 shows the boxplots of CVs of the estimates of LFS indicators (levels) for all provinces. The CVs of all estimates range from 0.76% to 29.04%. Estimates for employment levels (ranges from 0.76% to 3.80%), total labor force levels (1.35% to 3.62%), population 15 years old and over,

less OCW (1.10% to 3.01%), and not in the labor force levels (1.85% to 5.00%) in all provinces/HUCs were reliable.

Table 3: Number of Provinces, HUCs by LFS Indicators (Levels) and CVs using Bootstrap Method

Indicators	CVs (%) of Provinces			CVs (%) of HUCs		
	< 10	(10, 20]	> 20	< 10	(10, 20]	> 20
Employment	81			33		
Unemployment	34	43	4	22	11	
Underemployment	78	3		26	7	
Visibly Underemployment	73	7	1	13	17	3
Total Labor Force	81			33		
Not in the Labor Force	81			33		
Population 15 year old &	81			33		

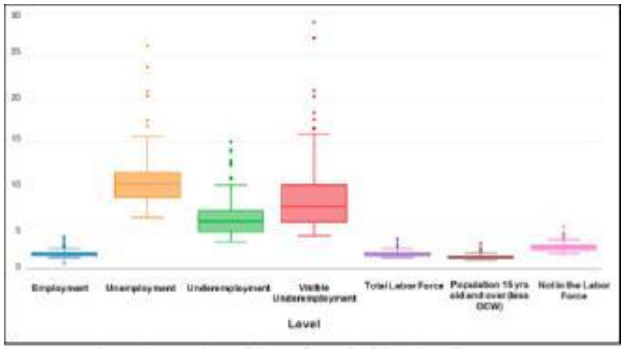


Figure 3: CVs of LFS Indicators (Levels)

Table 4 shows the total number of provinces/HUCs by major LFS indicators (rates) and CVs. Unemployment rate had the least number of provinces/HUCs (46 out of 114) with reliable estimates. Figure 4 shows the boxplots of coefficient of variations of the estimates of major LFS indicators (rates) for all provinces nationwide. The CVs of all estimates range from 0.11% to 29.57%. The CVs of employment rate (0.11% to 1.87%) and labor force participation rate estimates (0.89% to 2.30%) in all provinces/HUCs were reliable.

Table 4: Number of Provinces/HUCs by LFS Indicators (Rates) and CVs using Bootstrap Method

Indicators	CVs (%) of Provinces			CVs (%) of HUCs		
	< 10	10 - 20	~ 20	< 10	10 - 20	~ 20
ER	81			33		
UR	32	45	4	19	14	
UndR	77	4		25	8	
VUndR	72	8	1	12	18	3
LFPR	81			33		

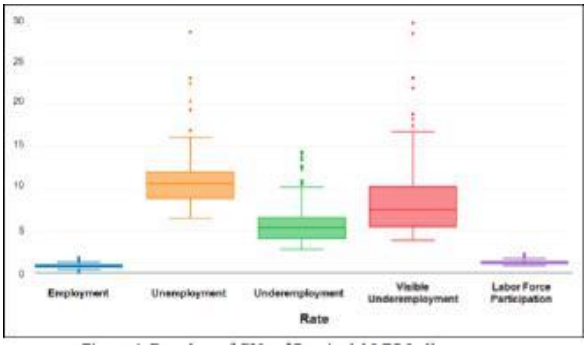


Figure 4: CVs of LFS Indicators (Rates)

4. Discussion and Conclusion

The use of the 2013 Master Sample design in the conduct of the Labor Force Survey calls for an estimation procedure for the computation of annualize provincial estimates of the LFS indicators.

Using simple averaging method for annualized provincial estimates, no LFS indicators in all provinces/HUCs are reliable for with and without disaggregation. However, using bootstrap method, for without disaggregation of the LFS indicators such as employment, labor force, not in the labor force and population 15 years old and over (less OCW) and for with disaggregation in the case of employment, labor force and not in the labor force by sex and by age group are reliable.

The bootstrap method (i.e., the weighted average of the four bootstrap quarter estimates) is better compared to the simple averaging method since it generally yields more reliable estimates.

References

1. Labor Force Survey (2016). Philippine Statistics Authority.
2. Master Sample Design (2003). Philippine Statistics Authority.
3. Census of Population and Housing (2010). Philippine Statistics Authority.
4. Master Sample Design (2013). Philippine Statistics Authority.
5. Cowpertwait, Paul S.P. & Andrew V. Metcalfe (2009). *Introductory Time Series with R*. New York: Springer.
6. Chernick, M.R. (1999). *Bootstrap Methods: A Practitioner's Guide*. New York: John Wiley & Sons, Inc.

Redesigning of the corn production survey in the Philippines

Abubakar S. Asaad[1,2], Sarah B. Balagbis[1], Jay M. Manlapaz[1]
*Melanie C. Estrada[1], Divina Gracia L. del Prado[3], Erniel B. Barrios[4]
[1]Statistical Methodology Unit, Philippine Statistics Authority
[2]College of Public Health, Department of Epidemiology and Biostatistics, University of the Philippines Manila
[3]Economic Sectoral Statistics Section, Philippine Statistics Authority
[4]School of Statistics, University of the Philippines Diliman

Abstract

Redesigning of the Philippine Statistics Authority's (PSA's) nationwide quarterly Corn Production Survey (CPS) was done to reflect the current behavior of corn production in each province. Simulation results show that stratified Lavalle-Hidiroglou (LH) algorithm is the best stratification with probability proportional to size systematic (PPS-SYS) and simple random sampling without replacement (SRSWOR), as the best selection methods for sample barangays and sample households, respectively. Further, design-based estimation is recommended over bootstrap method, in terms of precision and accuracy.

Keywords

Sampling frame; sampling design; design-based; bootstrap method

1. Introduction

The CPS is a nationwide quarterly agricultural survey conducted by the PSA during the months of April, July, October and December. The CPS aims to generate provincial estimates and forecasts on corn production, area and yield as inputs for government policies and programs on corn1. The CPS uses the 1991 Census of Agriculture and Fisheries (CAF) as a sampling frame, with the total farm area based on the 1980 Census of Agriculture (CA). This served as the stratification variable and the number of strata is either 10 or 5 depending on domain size. The coverage of the survey include Isabela, Laguna and Bukidnon and National Capital Region (NCR)2. The histogram of corn production and corn farm area (1st Semester) in the Philippines using the 2013 CPS3 data and 2012 CAF4 respectively, were positively skewed for each province. The correlation coefficient of corn production and corn farm area by quarter is from moderate to high.

The redesigning of the survey was done to reflect the current behavior of corn production in each province in the country. The redesigning of CPS is aimed to come up with a better sampling designs that will generate reliable

corn production estimates at the provincial level. Specifically, this paper aims to perform simulation to determine the best sample size (barangays and households), the best sampling design, the best estimation procedure and model for estimation of production of unsurveyed quarters.

2. Methodology

2.1 Phase 1: Determination of Initial Number of Sample Barangays

The LH algorithm is an iterative procedure that allocate samples into strata5. Different combinations of number of strata (3, 5 and 10 strata) and coefficient of variations (1%, 5% and 8%) were explored to compute for the initial number of sample barangays.

2.2 Phase 2: Determination of Best Sampling Design

2.2.1 Probability Proportional to Size-Systematic (PPS-Systematic)

Corn farm area was used as the measure of size. Barangays were selected using PPSSYS and households were selected using either SRSWOR or systematic. The designbased weight, $w_i = w_{ij} = \frac{\sum_{i=1}^{A} X_i}{aX_i} \times \frac{N_i}{n_i}, w_i \equiv$weight for all household in barangay $i, w_{ij} \equiv$ weight for all household j in barangay i by quarter, $X_i \equiv$ corn farm area in barangay i by quarter, $A \equiv$ total number of barangays per quarter, $a \equiv$ number of sample barangays per quarter, $N_i \equiv$ number of corn farm households in barangay i by quarter, and $n_i \equiv$ number of sample corn farm households in barangay i by quarter. The total corn production in the province is computed as: $\hat{Y} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} w_{ij} y_{ij}, y_{ij} \equiv$ corn production of household j in barangay i per quarter. The variance was computed as: $V(\hat{Y}) = \left(1 - \frac{a}{A}\right) a s_1^2 + \frac{a}{A} \sum_{i=1}^{n_i} \left(1 - \frac{n_i}{N_i}\right) n_i s_2^2$ where: $s_1^2 \equiv$ first stage sample variance (provincial sample variance), $s_2^2 \equiv$ second stage sample variance (barangay sample variance).

i. Stratified using Iterative Method

This method is known as cumpf rule which is based on constructing equal intervals on the cumulative of the square roots of the frequencies of the stratification variables6. Proportional allocation was employed and stratum boundary was determined using the iterative method.

Case 1: PPS selection of barangays, and SRSWOR and systematic selection of households. Note, $w_{hi} = w_{hij} = \frac{\sum_{i=1}^{A_h} X_{hi}}{a_h X_{hi}} \times \frac{X_{hi}}{n_{hi}}, w_{hi} \equiv$ weight for all household in barangay i at stratum $h, X_{hi} \equiv$ corn farm area in barangay i in stratum $h, A_h \equiv$ total number of barangays in stratum $h, a_h =$total number of sample barangays in stratum $h, N_{hi} \equiv$ total number of corn farm households in

stratum h, barangay $i, n_{hi} \equiv$ total number of sample corn farm households in stratum h and barangay i.

Case 2: Selection of Barangays - SYS and selection of Households - SRSWOR & SYS:$w_{hi} = \frac{A_h}{a_h} \times \frac{N_{hi}}{n_{hi}}$. The total corn production for stratified using iterative method was computed as: $\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$, $\hat{Y}_h = \sum_{i=1}^{a_h} \sum_{j=1}^{n_{hi}} w_{hij} y_{hij}$ where: $\hat{Y} =$ sumof weighted total corn production in all strata, $\hat{Y}_h =$ weighted total corn production in stratum $h, w_{hij} =$ weights for household j in barangay i, stratum h, and $y_{hij} =$ corn production in household j, barangay i, stratum h. The variance is computed as:

$$V(\hat{Y}_h) = \left(1 - \frac{a_h}{A_h}\right) a_{h s^2_{h1}} + \frac{a_h}{A_h} \sum_{i=1}^{n_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}}\right) n_{hi s^2_{h2}}.$$

ii. Stratified using LH Algorithm

Case 1: PPS selection of barangay, and SRSWOR and systematic selection of households. Same as Case 1 of Stratified using Iterative Method.

Case 2: Systematic selection of barangays, and SRSWOR and systematic selection of households. Note, $w_{hi} = \frac{A_h}{a_h} \times \frac{N_{hi}}{n_{hi}}$. The total corn production is computed as: $\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h + \hat{Y}_1, \hat{Y}_h = \sum_{i=1}^{a_h} \sum_{j=1}^{n_{hi}} w_{hij} y_{hij} \equiv$ weighted total production of stratum h, $\hat{Y}_1 = \left(\frac{\%\ share\ of\ strat\ 1\ to\ X}{1 - \%\ share\ of\ strat\ \%\ 1\ to\ X}\right) \sum_{h=1}^{L} \hat{Y}_h \equiv$ total production for the take none stratum, X ≡ corn farm area, The variance is, $\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \hat{V}(\hat{Y}_h) + \hat{V}(\hat{Y}_0), \hat{Y}_0 = \left(\frac{\%\ share\ of\ stratum\ 1\ to\ X}{1 - \%\ share\ of\ stratum\ \%\ 1\ to\ X}\right)^2 \sum_{h=1}^{L} \hat{V}(\hat{Y}_h)$,
$\hat{V}(\hat{Y}_h) = \left(1 - \frac{a_h}{A_h}\right) a_{h s^2_{h1}} + \frac{a_h}{A_h} \sum_{i=1}^{n_{hi}} \left(1 - \frac{n_{hi}}{N_{hi}}\right) n_{hi s^2_{h2}}$.

2.2.4 Recreation of Frame for Simulation

Frame for simulation was recreated by initially building a regression model for the non-zero corn production data of 2013 CPS: $Y_{ij} = \beta_i X_{ij} + \epsilon_{ij} \equiv$ corn production of province i for quarter j and $X_{ij} \equiv$ corn farm area of province i for quarter j. The regression model was used to predict corn production of each province per quarter in the 2012 CAF. This recreated frame served as the sampling frame for simulation. Results of the simulation are compared in terms of accuracy and precision using mean absolute percentage error (MAPE) and average coecient of variance (CV).

2.3 Phase 3: Selection of Best Estimation Procedure

2.3.1 Modified Bootstrap 1st Stage

Resampling is done at the first stage (selection of sample barangays). Steps: (1) Draw sample barangays of size α from A for each province and for

each quarter using the design chosen from Phase 2 of the simulation (e.g., PPS-SYS). (2) Draw sample households from each barangay of size $n = (n_1, n_2, \cdots, n_\alpha)$. (3) Compute for $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_\alpha)$from the original sample where $\hat{\theta}_i = \alpha\left(\frac{T_i}{\phi_i}\right), T_i = \sum_{j=1}^{n_i} \frac{N_i}{n_i} y_{ij}$ and ϕ_i is the probability of selection of barangay i computed as the reciprocal of the design weight for barangay i. Each $\hat{\theta}_i$ will serve as an estimate of PSU total and $\hat{\theta}$ will serve as the pseudo-population. (4) Generate size K bootstrap sample from $\hat{\theta}$ using SRSWOR.

2.3.2 Modified Bootstrap 2nd Stage

Resampling is done at the second stage (selection of sample corn farm households). Steps: (1) Same steps as in 1 and 2 of Modified Bootstrap 1st Stage. (2) Let $N^* = (k_1^*)n_i, (k_2^*)n_2, \ldots, (k_\alpha^*)n_\alpha$ where k^* is the vector of rounded-up elements of $k = \frac{N_1}{n_1}, \frac{N_2}{n_2}, \ldots, \frac{N_\alpha}{n_\alpha}$ where $n_1, n_2, \ldots, n_alpha$ is the number of sample corn farm households in barangays 1,2,…,α while $N_1, N_2, \ldots, N_\alpha$ is the total number of corn farm households in barangays 1,2,…,α. (3) Generate $U^*$by copying each households from each sampled barangays k_α^* times and generate new households IDs for these. (4) Draw $K = 500$ samples of size n households using SRSWOR from each barangay from U^*.

The bootstrap estimator of total corn production is $\hat{\theta}_K = \frac{\sum_{k=1}^{K} \hat{\theta}_k}{K}$, $k = 1,2,3, \ldots, K$, $\hat{\theta}_k = \sum_{i=1}^{a} w_i \sum_{j=1}^{n_i} y_{ij}$, $w_i \equiv$ weight of barangay i computed via PPS-SYS and $y_{ij} =$ corn production for household j, barangay i. The estimator of variance is $V(\hat{\theta}_K) = \frac{\sum_{k=1}^{K} (\hat{\theta}_k - \theta_K)^2}{K-1}$.

2.4 Phase 4: Model for Estimation of Corn Production

Corn production in the country is highly a↵ected by seasonal patterns brought about by variations in climate and soil types8. Provinces that contribute to the 97% of the total corn production were considered top producing were surveyed. However, nontop producing provinces of a quarter were un-surveyed, and their corn production was estimated using models. Three models were developed for the estimation.

2.4.1 Model 1

A mixed effect model with standing crop area as fixed e↵ects and municipal as random effect. $Y_{it} = \alpha + \beta Z_{it-1} + \gamma_{it} + \epsilon_{it}$, $\epsilon_{it} \backsim N(0, \sigma^2), Y_{it} \equiv$ total production at municipality i on quarter $t, Z_{it-1} \equiv$ total standing crop of municipality i at quarter $t-1$ and $\gamma_{it} \sim N(0, G)$, where G≡ variance-covariance matrix of the random effect.

2.4.2 Model 2

A mixed effect model with total corn area of previous quarter as fixed e↵ects and municipal random effect. $Y_{it} = \alpha + \beta X_{it-1} + \gamma_{it} + \epsilon_{it}$, $\epsilon_{it} \backsim N(0, \sigma^2), X_{it-1} \equiv$ corn area production at municipality i at quarter $t-1$ and $\gamma_{it} \sim N(0, G)$.

2.4.3 Model 3

A mixed effect model with municipal random effect and total corn area and published provincial yields as fixed effect. $Y_{it} = \alpha + \beta X_{it-1} * W_{it+}\gamma_{it} + \epsilon_{it}$, $\epsilon_{it} \backsim N(0, \sigma^2), W_{it} \equiv$ provincial average yield at municipality i at quarter $t-1$ and $\gamma_{it} \sim N(0, G)$. Each model was developed using CPS survey data collected from years 2013 to 2015 and bootstraps estimates of production with 200 replications were computed.

3. Results

3.1 Determination of Initial Number of Sample Barangay

For the first phase of this research study, the initial sample barangays was determined. Corn farm area was used as the stratification variable when using the LH algorithm since it is highly correlated with corn production. Among all combinations of CVs (1%, 5% and 8%) and strata (3, 5, and 10), five (5) strata and 5% CV yielded the optimal sample size. Results revealed that using 10 strata would result to too many zero cells while using three (3) strata on the other hand would result to too large sample size per strata. For the CVs, the sample size would be too small if 8% CV was used while using 1% CV would result to too large sample size. The total computed initial number of barangays is 1,952.

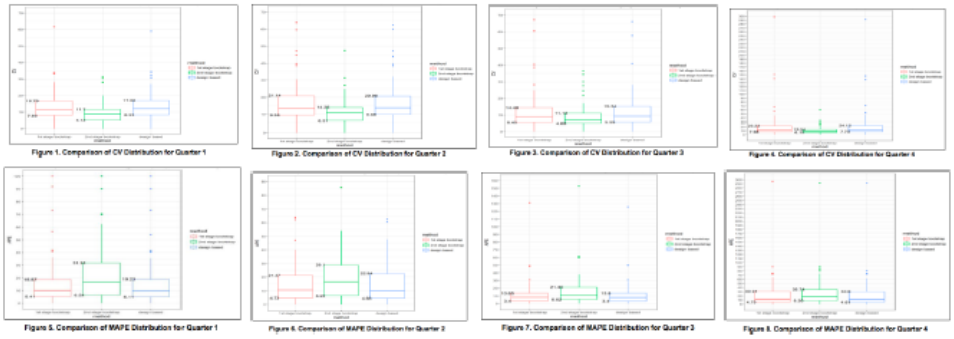
3.2 Determination of the Best Sampling Design

For the second phase of this study, the di↵erent scenarios for the determination of best sampling design was performed using 5% CV and five (5) strata. Results showed that stratified-LH is the best stratification in terms of precision and accuracy with PPS Systematic as the best sampling selection method for sample barangays, in general. Also, SRSWOR is the best selection method for sample households. Further, it was observed that accuracy and precision generally increase as the number of sample household increases. Hence, of the three (3) sample sizes considered (10, 15 and 20) households), a sample of 20 households is the best in terms of CVs and MAPEs. Also, the CVs and MAPEs are high for minor producing provinces/HUCs.

3.3 Determination of the Best Estimation Procedure

Results for the third phase showed that modified bootstrap 2nd stage has the lowest trimed CVs in all quarters with CVs ranging from 4.8% to 14.3%,

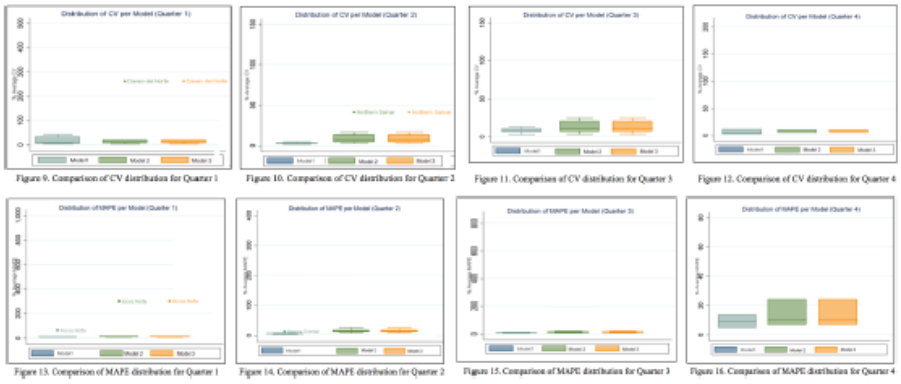
while CVs of the modified bootstrap 1st stage and design-based ranged from 5.4% to 24.2% (Refer to Figures 1-4). Although, modified bootstrap 2nd stage is the most precise method, it



is, however, the least accurate method, registering the highest MAPEs in all quarters (See Figures 5-8). Design-based and modified bootstrap 1st stage are the most accurate methods in all quarters with trimmed MAPEs ranging from 3.4% to 31.0%, while MAPEs of modified bootstrap 2nd stage ranged from 5.5% to 35.8%.

3.4 Determination of the Best Model

Results for the fourth phase showed that Model 1 has lower dispersion compared to Models 2 and 3 for both quarters 2 and 3. However, for quarters 1 and 4, dispersion of CVs were lower for Models 2 and 3 (Figures 9-12). Clearly, the performance of the three models were varied per quarter. The accuracy of estimates for the three models revealed that Model 1 outperformed the other two models as it consistently yielded estimates with generally lower MAPE per quarter (Figures 13-16).



4. Discussion and Conclusion

Stratified-LH is the best stratification in terms of precision and accuracy with PPSSYS and SRSWOR as the best selection method for sample barangays and sample households, respectively. Twenty (20) sample households seem to

be optimal. Design-based estimation and modified bootstrap 1st stage are comparable in terms of accuracy and are both better than modified bootstrap 2nd stage. Althought the best sampling design, based on simulation results, is stratified-LH with PPS-SYS as the barangay selection method and SRSWOR as the sample household selection method, PPS-SYS with corn farm area as the measure of size is better over LH since the CVs of PPS-SYS and LH are very close and the estimation procedure for PPS-SYS is easier to implement. It follows that the selection method for barangays is systematic and SRSWOR is the selection method for sample households.

References

1. countrystat.psa.gov.ph.
2. Compilation of Sampling Methodologies used in Agriculture and Fisheries Survey. Survey Designs Research and Development Section, Philippine Statistics Authority, March 2015
3. Corn Production of the Philippines (2013). Philippine Statistics Authority.
4. Census of Agriculture and Fisheries (2012). Philippine Statistics Authority.
5. Lavallee, P. and Hidiroglou, M.A. (1988). On the stratification of skewed populations. Survey Methodology, 14, 33-43
6. Dalenius, T. and Hodges, J. L. (1959), "Minimum Variance Stratification," Journal of the American Statistical Association, 54, 88-101.
7. Chernick, M.R. (1999). Bootstrap Methods: A Practitioner's Guide. New York: John Wiley & Sons, Inc.
8. Gerpacio, R.V., J.D. Labios, R.V. Labios, and E.I. Diangkinay (2004). Maize in the Philippines: Production Systems, Constraints and Research Priorities. Mexico.

The evolution of the OECD countries after the 2008 financial crisis simultaneous data analysis of the "How's Life" datasets between 2009 and 2015

Paulo Jorge Gomes, Joao Pedro Delgado
Nova IMS, Universidade Nova de Lisboa, Portugal

Abstract

The financial crisis of 2008 affected virtually every country in the World due to the connectivity of the global markets. Despite the significant contrasts in the starting points, there is the common perception that different economies recovered at distinct paces at least in part due to the policies and methods adopted by the authorities to address the financial crisis. In this context, the OECD "How's Life" datasets were analyzed with the objective of trying to detect trajectories in countries that could partially be explained by the macroeconomic measures adopted after the crisis. With the support of the OECD secondary data for the period 2009-2015, this novel study involved not only univariate, bivariate, and cluster evaluations but also a three-way data analysis based on the STATIS method. Among the existing multivariate methodologies, STATIS is the most comprehensive and flexible method to assess the evolution of a large (and possibly varying) number of individuals and variables over several years. With the identification of country trajectories in association with the evolution of variables, the findings may be relevant for business organizations with regard to defining strategic directions and making operational decisions.

Keywords

OECD How's Life/Better Life; PCA; Three-Way Data Analysis; STATIS; 2008 Financial Crisis

1. Introduction

Although the financial crisis of 2008 was not an entire surprise for people from within the industry with a critical mindset, the reality is that the large majority of the insiders and outsiders perceived the developments as a "Black Swan": something totally unpredictable and thus, unavoidable. Regardless of the differences in perspectives, the 2008 crisis started in the USA but quickly propagated and contaminated not only the European but also the Asian markets due to the global connectivity and scale of the financial and business operations.

The global financial crisis affected several countries in different ways and to varying extents. Furthermore, the impacted countries were in different positions in terms of macroeconomic aspects among other dimensions, which resulted in a multitude of different starting points for the post-crisis recovery.

Nonetheless, the analysis of the growth path of the OECD countries based on the "How's Life" datasets unveiled a number of distinct progressions associated to the different evolution of variables dependent on the policies adopted by governments and authorities to address the critical financial circumstances. The identification of different recovery trajectories and variables' evolution may provide valuable information for the processes of business decision-making.

At the request of the President of France in 2010, a team led by Joseph Stiglitz produced a report on the measurement of social and economic progress. This seminal paper represented a breakthrough in relation to the traditional and common way of gauging progress based on GDP alone, which reinforced the OECD initiative related to the collection of data on multiple types of variables linked to the quality and conditions of life. Since 2005, the OECD "How's Life" program has been gathering data and information in relation to the member countries (currently 35) and some partner countries.

From 2011 onwards, the "How's Life" program has been supporting the "Better Life Index" initiative that permits the individual weighting of the different variables to generate results that aretailored to meet the priorities of each user. Although the OECD approach permits to depart from a narrow and limited GDP perspective as discussed by various authors in several papers, the evolution of the multiple variables in the 35 member countries (plus six partners) permits to produce a space analysis over time. In addition to a global and intra-country assessment, a multivariate three-way data analysis provides trajectories for the evolution of the countries in the context of the selected variables.

The available OECD data relates to the current well-being variables (25) in the period from 2005 to 2015 (or 2016 in some cases) but presents several gaps for a few countries and in some years. This secondary data is credible, consistent, and reliable which permits to have confidence in the results obtained through a multivariate spatial analysis. Even though the OECD "How's Life" reports are frequently used as an important reference for the 11 well-being dimensions, the datasets permit to develop a multivariate analysis at three dimensions in order to characterize the evolution of the current well-being variables and assess the recovery of the countries after the 2008 crisis.

2. Methodology

The STATIS (Structuration des Tableaux À Trois Indices de la Statistique) (Escoufier, 1987; Lavit, 1988) method permits to analyze cubes of data and obtain a joint assessment of a set of quantitative tables. In particular, this technique is useful for the analysis of data evolution over time and so, it is related to techniques such as DPCA (Double Principal Components Analysis) and MFA (Multiple Factorial Analysis). The currently available computing

capacity allows the analysts to avoid the complexity resulting from the evaluation of each table and variable by employing an integrated graphic representation of the data collected on periodic occasions. The focus on the relative position of the individuals provided by the STATIS analysis results from the graphic displays that summarize the most important aspects related to large data sets involving multiple variables. Despite the loss of some information detail, the representations resulting from a multidimensional method (such as STATIS) are easy to interpret visually which permits to unveil the main features of the data.

For a set of S data tables, the STATIS method represents each study by an object W_s and the study is defined by three elements (X_s, Q_s, D) with D (observations weight) being constant and with Qs being equal to either I_p or $(diagV)^{-1}$ (for normalized data). For a table X_s (n x p) (with s = 1, ..., S), the representative object is obtained by: $W_s = X_s Q_s X^t_s$(size n x n). For the object distances and graphical representation of the tables, the STATIS method uses the Hilbert-Schmidt inner-product which indicates the existing degree of association between data tables:$\langle Ws\ Ws \rangle$ $(WS\ IWS')$ HS = $T_r(DW_sDW_{s'})$, where T_r (trace) is the sum of the diagonal elements. The joint analysis of multiple data tables permits to have a varying number of variables (STATIS, for object relations) or objects (Dual-STATIS, for variable relations) over time and to collect data with or without a defined periodicity.

This method involves four stages: (i) global analysis based on an interstructure comparing the data table structures with the support of the existing distances and graphic representation; (ii) identification of a compromise table W representing all the data tables in order to avoid the complexity of analyzing the various tables in an independent and separate way; (iii) detailed analysis resulting from the study of the intrastructure which permits to evaluate the similarities and differences between the tables based on their compromise positions; and (iv) analysis of the trajectories presented by each component (objects or variables) of the various data tables over time to appraise the evolution.X

3. Results

The OECD data related to the "How's Life" program for the member and associated countries (35 plus 6 countries in total) involved a varying number of observations and variables during the period from 2009 to 2015. Likewise, it was decided to focus the study on 34 member countries (excluding Chile and the associated countries due to their extensive data gaps) and to use the data for the 15 most complete variables only. Although there were some missing values (c. 5.5% that were imputed through maximum likelihood estimates or correlations), it was possible to produce a joint analysis of the

several data tables based on the STATIS and PCA (Principal Components Analysis) methods with a focus on the various individual countries.

The tables related to quantitative data collected for the same countries (34) and variables (15) in different years (7), and permitted to perform the simultaneous analysis and exploration of the entire set of data tables. The study individuals were the countries (Australia, Austria, Belgium, Canada, Switzerland, Czech Republic, Germany, Denmark, Spain, Estonia, Finland, France, United Kingdom, Greece, Hungary, Ireland, Iceland, Israel, Italy, Japan, Korea, Luxembourg, Latvia, Mexico, Netherlands, Norway, New Zealand, Poland, Portugal, Slovak Republic, Slovenia, Sweden, Turkey, and United States) while the variables involved several of the indicators measured by the initiative in accordance with the datasets of the "How's Life" report of 2017.

The study variables are: Household net-adjusted disposable income (USD at PPP, per capita, 2015); Employment rate (age 15 to 64, as % population with same age); Average annual gross earnings per full-time employee (USD at 2016 PPP); Labor market insecurity (monetary loss from unemployment, share previous earnings); Long-term unemployment rate (% labor force unemployed more than one year); Rooms per person (average number); Household expenditure on housing (% household gross adjusted disposable income); Dwellings without basic sanitary facilities (% people w/o dedicated flushing toilet); Employees working very long hours (% employees working more than 50h/week); Life expectancy at birth (years); Perceived health status (% adults self-reporting above "good"); Upper secondary education attainment per adults (% people 25-64); Social support (% people that can rely on friends or relatives); Satisfaction with water quality (% people in the population); and Feelings of safety when walking alone at night (% people).

The analysis produced at a global level permitted to obtain a view on the general evolution and trends with regard to the conditions of life in the OECD countries during the period from 2009 to 2015 (i.e., after the 2008 global financial crisis). For this purpose, each of the years in the analysis period was treated as an observation (center of gravity), and the study variables were the selected indicators (15) of the "How's Life" program. The statistical effect of the outlier observations related to Mexico and Turkey (on four variables each), Korea (on three variable), and Spain and Greece (on two variables each) was attenuated due to the standardization of data given the different units of the study variables.

In this context, the PCA conducted to eigenvalues (and associated eigenvectors) for the correlation matrix indicating that the first two axes largely explained the results given their combined variability (85.6% of the total inertia). The representation on the first principal plan (Figure 1) indicated that the first axis related to the evolution over time of the dimensions associated with the quality and material conditions of life. In the period 2009

to 2011, the stimulus packages in the OECD countries permitted an evolution of the variables, but there was a stagnation between 2011 and 2013 mainly due to aspects related to unemployment and income. The growth phase was resumed in the years 2014 and 2015. In relation to axis 2, there was a contrast between the initial and final years (mainly 2009 and 2015) and the intermediate years (2011 to 2013, with 2010 and 2014 being almost neutral). This trough (Guttman effect) revealed a decline in essential aspects after the 2008 global crisis until 2012 (pick year of the crisis), which was gradually recovered and surpassed by the OECD (as a whole) in 2015.

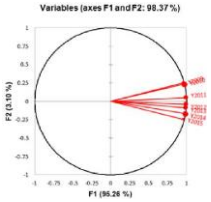


Fig. 2 – Representation of interstructure

The study used normalized objects to analyze the interstructure. The first two axes represented 98.37% of the inertia (with the first axis alone contributing 95.26%) and so, it was viable to assess the interstructure based on the first principal plan. The representation on the first principal plan (Figure 2) revealed that there was a common structure for all the objects (representing the data tables) in the period from 2009 and 2015. Apart from being possible to detect a sequential evolution from 2009 to 2015 with a good quality of the representations (the projected norms on the first axis were close to 1), it was interesting to notice that objects 2009 to 2011 were in opposition to the data tables of 2012 to 2015 in terms of axis 2 (despite its reduced inertia).

With a view to obtaining the compromise Euclidean image, the PCA of the compromise table produced the eigenvalues and associated inertias. For the purpose of the study, it was decided to focus on the interpretation of the first two axes which represented a combined 56.9% inertia. The meaning of each axis could be interpreted based on the correlation coefficient between the principal component of compromise and the initial variables. In terms of axis 1, there was an opposition between variables ranging from the indispensable needs (on the left) to the quality and conditions of life (on the right) and so, axis 1 could be understood as the level of development from a social and collective progress point-of-view. The aspects more exposed to axis 1 were the absence of basic facilities, unemployment, and labor security in opposition to employment, water quality, security, salary, and household income. In addition, axis 2 addressed aspects that were dependent on personal welfare and wealth and thus, ranged from the requirements that were independent of financial means and capabilities to dimensions that were

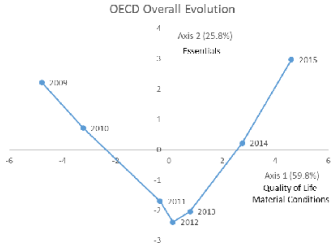


Fig. 1 – OECD evolution in 2009-2015

impacted by the circumstance at an individual level. In particular, the axis 2 presented secondary education, employment, housing expenditures, and water quality in opposition to labor security and unemployment (with negative impact) plus salary, income, health status, and life expectancy (positively affecting the individuals).

With the interpretation of the axes, it was possible to present the compromise positions of the various countries on the first principal plan (intrastructure) which represented the average positions of the countries during the study period (Figure 3). Based on K-means clustering, it was interesting to note a cluster (#1) of Central and Northern European plus North American and Australasia countries. In addition, there was a cluster (#2) of countries including the Southern and some Central European countries, and another cluster (#3) of Eastern European countries plus Korea. Finally, there were three countries (Mexico, Turkey, and Greece) in a cluster (#4) of their own. On axis 1, there was a clear progression of the compromise positions (from cluster #4 towards cluster #1) in terms of the social progress and development (with cluster #2 being positioned in a somewhat more neutral position). In particular, countries as Turkey, Mexico, Greece, and Latvia were positioned on the "Basics" and "Elementary" quadrants of the indispensable aspects in terms of social progress. On the other hand, countries as Switzerland, Norway, Canada, and the USA were located on the "Essentials" and "Aspirational" quadrants of social progress relative to the society quality-of-life and material conditions.

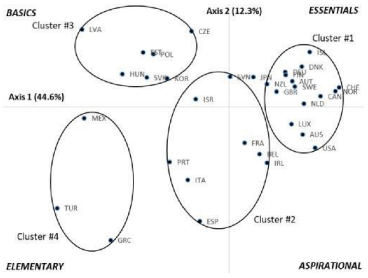


Fig. 3 – Countries' compromise positions and clusters

In terms of axis 2, cluster #3 appeared to be located at the level of assurance of the basic aspects regardless of individual wealth circumstances while cluster #4 seemed to be facing conditions where the personal welfare was decisive. Although clusters #1 and #2 were located in a more intermediate position in relation to axis 2, there were some significant country oppositions within each of these two clusters. In fact, there were countries with compromise positions indicating that the quality of individual life was more independent from the personal circumstances (perhaps due to the existing government policies) while others were more impacted by the wealth at an individual level. In particular, Latvia and Czech Republic ("Basics" quadrant) plus Iceland and Denmark ("Essentials" quadrant) displayed positions that were the least dependent on personal wealth despite the significant opposition at a social level, which could reflect insipid vs. developed social mechanisms where the individual welfare either could not be achieved with or did not require private financial means. At the other extreme of axis 2, Turkey

and Greece (even more than Spain and Italy) were countries where the personal wealth was decisive in terms of the impact on the circumstances and welfare at the individual level, which suggested that the physical infrastructure could exist but was available only to those whom could afford the associated costs.

The trajectories of the various countries permitted to have a more detailed appreciation of the evolution of each country during the seven-year period of the analysis. A long trajectory indicated a country that had developed more in terms of the variables structure than the average of the variables for the OECD countries, while a short trajectory revealed that the country had progressed in line with the variables' averages for the countries in the OECD. In this context, it was relevant to note that the countries with the most differentiated evolution were part of cluster #4 (Turkey, Greece, and Mexico) while two countries in cluster #2 (Spain and Italy) and three countries in cluster #3 (Estonia, Latvia, and Slovakia) also presented a significant evolution. In addition, there were seven countries in cluster #1 (Germany, Iceland, Netherlands, New Zealand, Norway, UK, and the USA) and one country in cluster #2 (Ireland) that presented a noticeable evolution.

However, it was worth noting that the cluster #1 countries (plus Ireland) evolved primarily along axis 2 in the direction of reducing the dependency on individual wealth to ensure the essential dimensions at a personal level (except for New Zealand). At the same time, the countries with the most significant evolutions in clusters #2, #3, and #4 displayed progression along not only axis 2 but also axis 1. Having said that, some of these countries (Latvia, Estonia, Slovakia, and Turkey) developed towards a higher quality and conditions of life at the society level (, Mexico, Greece, Italy, and Spain. With regard to axis 2, these countries displayed a trend towards an increased dependency on personal wealth to secure the necessary dimensions at the individual level (with the exception of Latvia and Estonia). Overall, the cluster #1 countries were located in the "Essentials" quadrant and reinforcing this position (with the USA and Ireland in the "Aspirational" quadrant but moving in the "Essentials" direction). Similarly, the Cluster #3 countries (Latvia, Estonia, and Slovakia) were in the "Basics" quadrant and progressing towards the "Essentials" area while the cluster #2 and #4 countries were located in the "Elementary" area but moving away from the "Essentials" (with the exception of Turkey and the recent recovery of some countries such as Italy, Spain, and Greece).

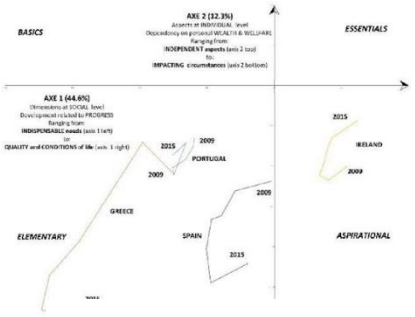


Fig. 4 – Trajectories of bailed-out countries

The trajectories of Ireland, Greece, Portugal, and Spain (Figure 4) were of particular interest given the bail-out programs and associated restructuring plans plus austerity measures implemented in these countries during the course of the study period (in addition to the initial stimulus packages adopted by the OECD countries after 2008). The trajectories were complemented by an analysis of the variables' variation in each country to appreciate the impact and extent of the local government and European measures and policies. In the case of Ireland, there was a trajectory inflection from 2010 onwards mainly based on the favorable movement of variables as labor insecurity, unemployment, house expenses, working long-hours, and secondary education. At the same time, Greece and Spain presented long trajectories with inflections from 2014 resulting from favorable employment, unemployment, working long-hours, secondary education, and water satisfaction (in Greece) plus household income, employment, labor insecurity, unemployment, water satisfaction, and feeling safe (in Spain). Although Portugal also benefited from a bail-out program, the country trajectory was much shorter and presented an inflection from 2013 onwards which was mainly due to favorable (but limited) movements in relation to employment, labor insecurity, unemployment, house expenses, secondary education, and feeling safe.

4. Discussion and Conclusion

Although the OECD "How's Life" datasets were not complete for all member and partner countries nor the entire set of variables, it was possible to select 34 countries and 15 variables over a period of seven years with a reduced number of data gaps. The global analysis of the datasets permitted to detect a Guttman effect in the evolution of the OECD countries given the general progress (with a stagnation phase from 2011 to 2013) in terms of quality and conditions of life over time (axis 1). At the same time, there was a significant decline in relation to essential aspects (due to the volatility of the variables associated to the individual, family, and government budgets) until 2012 but the general recovery afterward permitted to exceed the 2009 levels by 2015 (axis 2).

In this context, the STATIS interstructure revealed not only the existence of a common structure for the objects representing the annual data tables, but also a sequential evolution with a good representation of the years. However, it was insightful to notice the contrasting interstructure opposition of the years 2009-2011 relative to 2012-2015. Furthermore, the correlation coefficients of the compromise principal components and the initial variables permitted to interpret the meaning of the axes based on the variables' oppositions. So, axis 1 relates to the social quality and conditions of life while axis 2 is associated with the dependency on the wealth circumstances at an individual level.

The compromise positions revealed the countries with more prominent positions. With regard to axis 1 (social aspects), there were countries with a more distinctive position in terms of not only the "Basic" and "Elementary" aspects but also with regard to the "Essentials" and "Aspirational" dimensions. The compromise positions and annual data permitted to also identify the existence of four main clusters with a distribution along axis 1 ranging from the "Elementary" (cluster #4) and Basics" quadrants (cluster #3) to the more central positions (cluster #2) and the "Essentials" and "Aspirational" quadrants (cluster #1). There were countries with long trajectories in all clusters but most countries tended to evolve towards the "Essentials" quadrant and so, the cluster #1 countries progressed mainly along axis 2 and the cluster #3 countries had a more predominant evolution in relation to axis 1, while the clusters #2 and #4 countries displayed a mixed evolution on both axes.

Although the OECD "How's Life" program departed from an economic and financial perspective mainly rooted on GDP, many of the decisive variables with regard to the longest country trajectories appeared to be directly (or at least semi-directly) related to the income and revenues generated at the individual, family, and government levels. Nonetheless, the identified critical variables revealed the importance of complementary aspects for the efficiency of business operations and results, namely in relation to social aspects and environmental priorities. It is worth noting that the countries with the largest stimulus packages in 2008-2010 (with the objective of emerging stronger out of the crisis) did not appear to have started to benefit from these investments yet which might be a reflection of their starting positions and/or an indication of insufficient time elapsed.

Apart from the fiscal and financial stimulus, government investments and expenditures, and support to families and businesses, some countries benefited from international bail-out programs (over and above the stimulus packages of 2008-2010) from approx. 2011 onwards. It is clear from the results that the bailed-out countries (Greece, Ireland, Portugal, and Spain) have been able to inflect their downward trajectories (albeit to distinct extents) and thus, have started to make progress towards the "Essentials" quadrant at different paces. In this process, the variables identified as being the most important and decisive represented a compromise and require a well-judged balance involving aspects of an economic, financial, social, and environmental nature.

References

1. Des Plantes, H. H. (1976). Structuration des Tableaux a Trois Indices de la Statistique. These de troisieme cycle. Université de Montpellier II, 94.
2. Escoufier, Y. (1987). Three-mode data analysis: the STATIS method. Methods for multidimensional data analysis, 259-272.
3. Lavit, C. (1988). Analyse conjointe de plusieurs matrices. Masson, Paris.

4. OECD (2017). How's Life? 2017: Measuring Well-being. OECD Publishing, Paris.
5. OECD (2009). Responding to economic crisis: Fostering industrial restructuring and renewal. OECD, Paris.
6. Stiglitz, J. E., Sen, A., & Fitoussi, J.P. (2010). Report by the commission on the measurement of economic performance and social progress. Paris: Commission on the Measurement of Economic Performance and Social Progress.

Redesigning of the Semi-Annual Survey of Dairy Enterprises in the Philippines

Ibarra Aaron R. Poliquit[1], Jay R. Manlapaz[1], Divina Gracia L. del Prado[2], Abubakar S. Asaad[1,3], Erniel B. Barrios[4]

[1]Statistical Methodology Unit, Philippine Statistics Authority
[2]Economic Sectoral Statistics Section, Philippine Statistics Authority
[3]College of Public Health, University of the Philippines Manila
[4]School of Statistics, University of the Philippines Diliman

Abstract

Listing of Dairy Farms was conducted in eight (8) priority provinces to generate an updated list of dairy enterprises. A probability proportional size systematic sample of barangays with the number of animals on the milk line as size measure and a dairy farm sample size of 25 per sampled barangay generated coefficient of variation and absolute percentage error below 10% for major-producing provinces which are the provinces that have barangays with 25 or more dairy farms. The simulation results provided an evidence that the number of animals on the milk line as the size measure is better than the number of female breeders as the size measure in terms of reliability of provincial milk production estimates.

Keywords

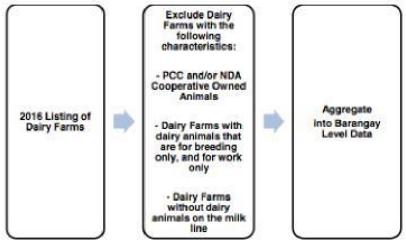
Probability Proportional to Size Systematic; LH Algorithm; Simulation

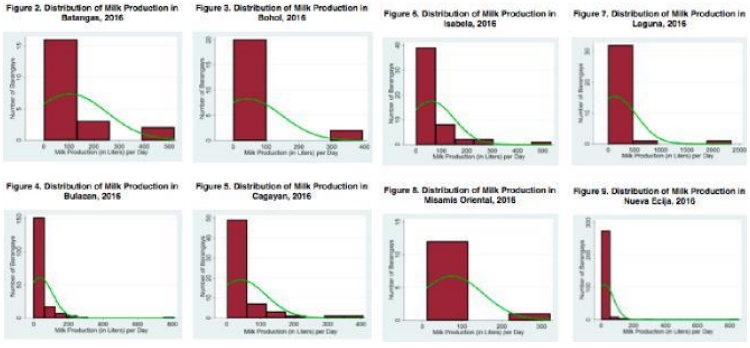
1. Introduction

The Semi-Annual Survey of Dairy Enterprises (SSDE) is one of the regular surveys conducted by the Livestock and Poultry Statistics Division (LPSD) of the Philippine Statistics Authority (PSA). This survey aims to generate data on dairy production and inventory of dairy animals nationwide1. Currently, the sampling frame used is based on the Dairy Enterprise Inventory Profiling (DEIP) 2002 which consists of 46 provinces where dairying exists. Due to the undercoverage of the sampling frame constructed in 2002, the estimate for milk production and dairy animal inventory are underestimated2. On August 2016, Listing of Dairy Farms (LDF) was conducted in eight (8) priority provinces identified by the Philippine Carabao Center (PCC) and the National Dairy Authority (NDA). This activity aims to generate an updated list of dairy enterprises3. Along with the updated list, a new estimation procedure is necessary to determine the appropriate number of sample barangays for each province and number of sample dairy enterprises for each sampled barangay in eight (8) priority provinces.

The LDF in selected provinces is an activity of the LPSD which aims to provide an updated and reliable sampling frame for the SSDE. The 2016 LDF covers household and commercial dairy farms in eight (8) priority provinces namely: Cagayan, Isabela, Bulacan, Nueva Ecija, Batangas, Laguna, Bohol, and Misamis Oriental. These were the priority provinces where dairying activity were extensive as determined by the Philippine Carabao Center (PCC) and National Dairy Authority (NDA). In preparing the barangay level frame, the steps are detailed in Figure 1. Figures 2 to 9 illustrate the distribution of milk production across the barangays per provinces. The figures show that the distribution of milk production per province is positively skewed. The number of animals on the milk line and milk production has the highest correlation (r = 0.9355), the number of female breeders with milk production as the second highest correlation (r = 0.6963), followed by inventory (r = 0.6934) and lastly with farm capacity (r = 0.3336). The histograms of the distribution of milk production was skewed and its correlation with some possible stratification variables were high, suggesting the need to use the Lavallee-Hidiroglou (LH) algorithm[4].

Figure 1. Process Flow from Operator to Barangay Level Data





2. Methodology

2.1 Determination of Barangay Sample Size per Province

The barangays or PSUs were stratified based on the number of animals on the milk line using the LH algorithm. Barangays contributing significantly to the number of animals on the milk line are part of the "must take" stratum and were considered as a certainty stratum. In contrast, barangays that contribute an insignificant number of animals on the milk line were classified under the

"take none" stratum and no sample barangay were drawn from this group. A simulation using R[5] for different scenarios were explored to determine the number of sample barangays for both animals on the milk line and female breeders as stratification variables using different combinations of number of strata (3, 4 and 5 strata) and CVs (1.0%, 5.0% and 8.0%) per province.

2.2 Probability Proportional to Size (PPS) Systematic

Treating the stratification variable as size measure, a PPS systematic sample of barangays were composed mostly of barangays with large size measures and fewer number of barangays with small size measures. This means that the sample barangays selected via PPS systematic sampling were similar to the sample barangays selected via stratified design with "take none" and "take all" strata. So, PPS systematic was used as the 1st stage sampling method for the selection of sample barangays[6].

2.3 1^{st} and 2^{nd} Stage Sample Selection Scheme

The secondary sampling unit (SSU) for this study is the dairy farm or dairy enterprise. The steps for the 1st and 2nd stage sample selection as follows:

1. Compute $n = 2 \times n_0$ where n_0 is the sample size computed such that the stratum number is 5, and the CV is 0.05 for a specific stratification variable, while the constant 2 is the design effect adjustment due to cluster sampling.
2. Compute the PPS Systematic sampling of barangays.
3. Label a barangays as the i^{th} sampled barangay when its corresponding cumulative size measure is less than or equal to $R + (i + 1) \times K$Replace the sample duplicates by doing steps 2 and 3 for the unselected barangays and by setting sample size equal to the number of duplicates.
4. Repeat step 4 until all barangays in the sample are distinct.
5. Divide the sample into 2, one for each semester, such that each division inherits the characteristics of a sample taken using PPS Systematic.
6. In each sampled barangay, select dairy farms via Simple Random Sampling Without Replacement (SRSWOR).
7. Explore different dairy farm sample sizes (i.e., m=5, 10, m=15, m=20, and m=25) in each sampled barangay.

a. Estimation Procedure

The design-based estimator of total milk production is computed as follows:

$$\hat{Y} = \sum_{i=1}^{n} \sum_{i=1}^{mi} w_i y_{ij}, \quad w_i = \frac{\sum_{i=1}^{N} X_i}{nX_i} \times \frac{M_i}{m_i}$$

where: $w_i \equiv$ final weight of i^{th} sampled barangay, $y_{ij} \equiv$ total milk production of the j^{th} sampled diary farm in i sample barangay, and $M \equiv$ total number of dairy farms in i^{th} sampled barangay. The design-based estimator of variance of total milk production is computed as follows:

$$V(\hat{Y}) = \left(1 - \frac{n}{N}\right) n s_1^2 + \frac{n}{N} \sum_{i=1}^{n} \left(1 - \frac{m_i}{M_i}\right) m_i s_2^2$$

where: $s_1^2 \equiv$ first stage sample variance, $s_2^2 \equiv$ second stage sample variance within i^{th} sample barangay.

2.5 Measure of Precision

The precision of the estimates for totals were evaluated using the coefficient of variation (CV) and mean absolute percentage of error (MAPE). Estimates with coefficient of variations which are less than or equal to 10% implies precision, and hence, reliable estimates.

3. Results

Tables 1 and 2 show the computed number of sample barangays across different strata and CVs per province with the number of animals on the milk line and number of female breeders as a stratification variable. The number of animals in the milk line as stratification variable computes a larger number of sample barangays than using the number of female breeders across all scenarios. The reason for this is that the distribution of animals on the milk line has a larger coefficient of variation than the distribution of female breeders. Only one scenario was considered for simulations, where the number of strata is 5 and the target CV is equal to 5.0%. The computed sample sizes for this particular scenario was deemed as the right compromise between the level of precision of estimates and the cost of operations. The total number of sample barangays for all the eight (8) provinces for the considered scenario is 99 sample barangays when using the number of animals on the milk line as a stratification variable, and 88 sample barangays when using the number of female breeders as stratification variable, respectively.

Table 1: Sample Barangays across Strata and CVs per Province with Number of Animals on the Milk Line as Stratification

Province	Barangays	Sample of Barangays								
		3 Strata			4 Strata			5 Strata		
		1%	5%	8%	1%	5%	8%	1%	5%	8%
Batangas	21	9	9	5	13	6	4	13	6	5
Bohol	22	13	6	5	10	5	4	10	5	5
Bulacan	180	110	53	40	110	35	22	88	24	14
Cagayan	62	26	19	15	39	16	10	32	12	7
Isabela	52	27	18	15	37	14	10	31	12	8
Laguna	34	14	9	7	17	7	6	11	8	6
Misamis Oriental	13	11	8	6	10	6	5	10	5	5
Nueva Ecija	288	**166**	79	48	**166**	39	23	124	27	15
Total	672	376	201	141	402	128	84	319	**99**	65

Only Bulacan, Cagayan, Isabela and Nueva Ecija have barangays with 25 or more dairy farms(See Table 3). These provinces will be referred as the major-producing provinces for the rest of the paper. Also, with respect to the sampling frame, simulated replicates for the 1st and 2nd semester are distributed independently and identically so no semestral analysis was done. Increasing dairy farm sample size per sampled barangay should improve the estimates. Estimates for provinces whose barangays have few dairy farms do not benefit much by increasing dairy farm sample size since setting a small dairy farm sample size is enough to cover all dairy farms

Table 2: Sample Barangays across Different Strata and CVs per Province with Number of Female Breeders as Stratification

Province	Barangays	Sample of Barangays								
		3 Strata			4 Strata			5 Strata		
		1%	5%	8%	1%	5%	8%	1%	5%	8%
Batangas	21	8	8	6	11	7	5	12	6	5
Bohol	22	16	6	5	11	5	4	10	5	5
Bulacan	180	105	48	34	108	38	21	71	23	15
Cagayan	62	20	18	15	25	13	11	29	8	6
Isabela	52	29	19	16	31	15	9	32	10	8
Laguna	34	9	9	6	14	6	6	14	7	5
Misamis Oriental	13	10	4	3	9	4	4	8	5	5
Nueva Ecija	268	165	68	41	122	34	20	122	24	13
Total	672	362	180	126	331	122	80	298	88	62

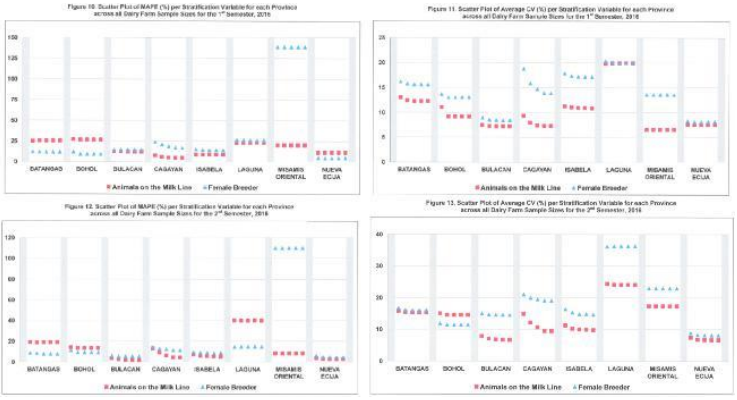
per sampled barangay. A dairy farm sample size of 5 is enough to cover all dairy farms per sampled barangay in Misamis Oriental where no barangays contain 5 or more dairy farms. Dairy farm sample size equal to 25, average CV's and MAPE's for the major-producing provinces have values below 10%.

Table 3: Barangays per Province in terms of Number of Dairy Farms

Province	Barangays	Number of Barangays with:				
		5 or more Dairy Farms	10 or more Dairy Farms	15 or more Dairy Farms	20 or more Dairy Farms	25 or more Dairy Farms
Batangas	21	2	1	1	0	0
Bohol	22	3	1	0	0	0
Bulacan	180	58	24	11	8	4
Cagayan	62	19	10	7	5	2
Isabela	52	32	18	14	12	11
Laguna	34	4	1	0	0	0
Misamis Oriental	13	0	0	0	0	0
Nueva Ecija	288	69	17	7	5	2
Total	672	187	72	40	30	19

Regarding the reliability of estimates based on the two size measure variables, Figures 10 to 13 indicate that for Misamis Oriental and the major-producing provinces, estimates are more accurate and precise when number

of animals on the milk line is the size measure variable. Additionally, for both semesters of Misamis Oriental, MAPE of estimates exceeded 100% when the number of female breeders is the size measure variable. The results are varied for the provinces of Batangas, Bohol and Laguna. These results are supported by the fact that the number of animals on the milk line has the highest correlation to total milk production among all possible size measure variables. However, these results might be inconclusive since total sample size is larger when the number of animals on the milk line is the size measure variable than when number of female breeders.



4. Discussion and Conclusion

The objective of computing barangay sample size using LH-algorithm was achieved. However, a stratified design using LH-algorithm was found to be inappropriate since for minor-producing provinces, some computed stratum sample sizes were equal to one which makes variance estimation impossible. A PPS systematic sample of barangays with number of animals on the milk line as size measure and a dairy farm sample size of 25 per sampled barangay, generate CV and APE below 10% for major-producing provinces (i.e., provinces that have barangays with 25 or more dairy farms).

Simulation results provide an evidence that using the number of animals on the milk line as size measure may be better than using the number of female breeders as size measure in terms of reliability of provincial milk production estimates.

References

1. Compilation of Sampling Methodologies used in Agriculture and Fisheries Survey (2015). Survey Designs Research and Development Section, Philippine Statistics Authority.
2. Santos, K.C. (2014). Improvement of Methodology in Semi-Annual Survey of Dairy Enterprises (SSDE) Phase I. University of the Philippines, Diliman, Quezon City.

3. Listing of Dairy Farms Field Operations Manual (2016). Philippine Statistics Authority, Quezon City.
4. Lavallee, P. and Hidiroglou, M. (1988). On the Stratification of Skewed Populations. Survey Methodology. 14(1), p. 33-34.
5. Baillargeon, S. and L.P. Rivest (2014). Univariate Stratification for Survey Population.https://cran.r-project.org/web/packages/stratification/stratification.pdf
6. Lohr, S. L., (2010). Sampling: Design and Analytics, Second Edition. Arizona State University, USA.

Inference on P(X < Y) for bivariate normal distribution based on censored data

Manoj Chacko
Department of Statistics, University of Kerala, Trivandrum, India

Abstract

In this paper, we consider the problem of estimation of $R = P(X < Y)$, when (X, Y) follows bivariate normal distribution and measurement on one variable is dificult. The maximum likelihood estimates (MLEs) and Bayes estimates (BEs) of R are obtained based on censored data, in which censoring is done based on the easily measurable variate. BEs are obtained based on both symmetric and asymmetric loss functions. The percentile bootstrap and HPD confidence intervals for R are also obtained. Monte Carlo simulations are carried out to study the accuracy of the proposed estimators. The inferential procedure developed in this paper is also illustrated using water quality data.

Keywords

Maximum likelihood estimation; Bayesian estimation; importance sampling method; order statistics.

1. Introduction

Censored sample arises in a life-testing experiment whenever the experimenter does not observe the failure times of all units placed on a life-test. In medical or industrial studies, researchers have to treat the censored data because they usually do not have sufficient time to observe the lifetime of all subjects in the study. There are different types of censoring. The most common censoring schemes are type-I and type-II censoring schemes. In this paper, we consider a type-II censoring scheme in which the experiment continues until a pre-specified number of failures, $r(\leq n)$ occur. The remaining (n-r) items are regarded as censored data. Let (X, Y) be an absolutely continuous random vector with cummulative distribution function (cdf) $F(x, y)$ and joint probability density function (pdf) $f(x, y)$. Let $(X_i, Y_i), i = 1, 2, \ldots, n$ be a random sample of size n drawn from the distribution of (X, Y). If the sample is ordered by the $X_i's$ i s then the Y -variate associated with the rth order statistic $X_{r:n}$ is called the concomitant of the rth order statistic and is denoted by $Y_{[r:n]}$ (see, David, 1973). Suppose that we only observe $r(\leq n)$ smallest X-sample and their associated values of the Y -sample, then $(X_{(i:n)}, Y_{[i:n]}), i = 1, 2, \cdots, r$ is called a type-II right-censored sample. The joint pdf of

$(\mathrm{X}_{(\mathrm{r})}, \mathrm{Y}_{[\mathrm{r}]}) = ((\mathrm{X}_{(1:\mathrm{n})}, \mathrm{Y}_{[1:\mathrm{n}]}), (\mathrm{X}_{(2:\mathrm{n})}, \mathrm{Y}_{[2:\mathrm{n}]})), \ldots, (\mathrm{X}_{(\mathrm{r}:\mathrm{n})}, \mathrm{Y}_{[\mathrm{r}:\mathrm{n}]}))$ is given by

$$f_{(X_{(r)},Y_{[r]})}(X_{(r)},y_{[r]}) = \prod_{i=1}^{r} f(y_{[i]}|x_{(i)}) f_{1,2,\dots,r:n}(x_{(1)},x_{(2)},\dots,x_{(r)}), \text{ where } f_{1,2,\dots,r}(x_{(1)},x_{(2)},\dots,x_{(r)}) \quad (1)$$

where $f_{1,2,\dots r}(x_{(1)},x_{(2)},\dots,x_{(r)})$ is the joint pdf of first r order statistics of random sample of size n and is given by

$$f_{1,2,\dots r:n}(x_{(1)},x_{(2)},\dots,x_{(r)}) = \frac{n!}{(n-r)!}[1-F(x_{(r)})]^{n-r}\prod_{i=1}^{r} f(x_{(i)}).$$

In the context of reliability, the stress-strength model describes the life of a component which has a random strength Y and is subjected to a random stress X. The component fails at the instant that the stress applied to it exceeds the strength and the component will function satisfactorily whenever $X < Y$. Thus, $R = P(X < Y)$ is a measure of component reliability. It has many applications especially in engineering concepts such as structures, deterioration of rocket motors, static fatigue of ceramic components, fatigue failure of aircraft structures and the aging of concrete pressure vessels. Some examples are as follows. If X represents the maximum chamber pressure generated by ignition of a solid propellant and Y represents the strength of the rocket chamber, then R is the successful firing of the engine. Let Y and X be the remission times of two chemicals when they are administered to two kinds of mechanical systems, then inferences about R present a comparison of the effectiveness of the two chemicals. If X and Y are future observations on the stability of an engineering design, then R would be the predictive probability that X is less than Y. Similarly, if X and Y represent lifetimes of two electronic devices, then R is the probability that one fails before the other.

The applications of $R = P(X < Y)$ is not limited to reliability and engineering, it has lot of applications in medicine, psychology, environmental studies etc. For example in medical studies, if X and Y represent the outcome of control and experimental treatments, then R can be interpreted as the effectiveness of the treatment. In the study of water quality in freshwater, if Y represent the concentration of dissolved trace metals such as zinc, copper or lead in water and X represents the corresponding worldwide water quality standards of that metal in water, then $P(Y < X) = 1 - R$ can be considered as the probability that the metal concentration in freshwater is lower than the corresponding worldwide standard.

Suppose in a study, measurement on one variable, say Y is difficult or expensive but measurement of other variable X is not difficult, then we can reduce the number of measured units on Y by applying censoring on the X observations. For example, in the water quality study, quantification of dissolved trace metals such as zinc, copper and lead in freshwater is very difficult and expensive whereas the determination of water quality standard of metals is easy as it is a function of magnesium and calcium which can be easily

quantified. In such situations, the procedures developed in this paper can be effectively utilized to make inference on $R = P(X < Y)$.

The estimation of R has been extensively investigated in the literature when X and Y are independent random variables belonging to the same bivariate family of distributions. However there is a relative little work when X and Y are dependent random variables. The problem of estimating R, when the X and Y are dependent, was considered by Abu-Salih and Shamseldin (1988), Awad et.al. (1981), Jana and Roy (1994) and Cramer (2001). Estimation of R when (X, Y) follows bivariate normal has been discussed by Enis and Geisser (1971) and Mukherjee and Saran (1985). Jana(1994) and Hanagal (1995) discussed the estimation procedure of R when (X, Y) follows Marshall-Olkin bivariate exponential. Hanagal (1997) discussed the estimation of R when (X, Y) has a bivariate Pareto distribution. In this paper, we focus on estimation of $R = P(X < Y)$ based on first r $(r \leq n)$ order statistics and its concomitants, when (X, Y) follows bivariate normal distribution (BVND). A random variable (X, Y) follows BVND if its pdf is given by

$$f(x,y) = \begin{cases} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}, \\ (x,y) \in R \times R;\ -\infty < \mu_1, \mu_2 < \infty; \sigma_1, \sigma_2 > 0; |\rho| < 1. \\ 0, \quad otherwise \end{cases} \quad (2)$$

If (X, Y) follows BVND with pdf defined in (2), then R is given by

$$R = P(X < Y) = \Phi\left[\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}}\right], \quad (3)$$

where Φ is the cdf of standard normal distribution. If we denote $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2\ \rho)$ be the vector of parameters then we can write

$$R = R(\theta) = \Phi\left[\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}}\right],$$

In this paper, We obtain both maximum likelihood estimator (MLE) and Bayes estimator of R and make an efficiency comparison between them. The Bayes estimators for R under different loss functions are obtained using importance sampling method. A real data on water quality is used to illustrate the inferential procedure developed in this paper.

2. Methodology

We consider both maximum likelihood estimation and Bayes estimation for R using censored order statistics and its concomitants. Percentile bootstrap

confidence intervals are also obtained in this section. For Bayes estimation for R we use importance sampling method.

a. Maximum Likelihood Estimation

In this section, we obtain the MLE of R for BVND given in (2). Let $(X_{(i)}, Y_{[i]}), i = 1, 2, \ldots, r$ be the vector of first r order statistics and its concomitants of a random sample of size n arising from BVND with parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Then from (1) the likelihood function is given by

$$L(\mu_1, \mu_2, \sigma_1 \sigma_2, \rho) = \prod_{i=1}^{r} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} exp \frac{-1}{2(1-\rho^2)} \{(\frac{x_{(i)} - \mu_1}{\sigma_1})^2 - 2\rho(\frac{x_{(i)} - \mu_1}{\sigma_1})\left(\frac{y_{[i]} - \mu_2}{\sigma_2}\right) + (\frac{y_{[i]} - \mu_2}{\sigma_2})^2\}\left(1 - \Phi(\frac{x_{(r)} - \mu_1}{\sigma_1}\right)^{n-r}. \quad (4)$$

The MLEs of $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ can be obtained by solving the non-linear normal equations. Let $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2$ and $\hat{\rho})$ be the MLEs of $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ obtained by solving the above non-linear equations, then the MLE of R is given by

$$\hat{R}_{ML} = \Phi\left[\frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2}}\right]. \quad (5)$$

b. Bootstrap Confidence Interval

In this section, we consider percentile bootstrap confidence interval for R based on MLEs. For that we do the following.

1. Compute the MLEs $\hat{\mu}_1^{(0)}, \hat{\mu}_2^{(0)}, \hat{\sigma}_1^{(0)}, \hat{\sigma}_2^{(0)}$ and $\hat{\rho}^{(0)}$ of $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ using original observations and put k=1.
2. Generate a bootstrap sample using $\hat{\mu}_{(k)}^{(k-1)}, \hat{\mu}_2^{(k-1)}, \hat{\sigma}_1^{(k-1)}, \hat{\sigma}_2^{(k-1)}$ and $\hat{\rho}^{(k-1)}$ and obtain the MLEs $\hat{\mu}_1^{(k)}, \hat{\mu}_2^{(k)}, \hat{\sigma}_1^{(k)}, \hat{\sigma}_2^{(k)}$ and $\hat{\rho}^{(k)}$ using the bootstrap sample.
3. Obtain the MLE of $\hat{R}_k = R(\hat{\mu}_1^{(k)}, \hat{\mu}_2^{(k)}, \hat{\sigma}_1^{(k)}, \hat{\sigma}_2^{(k)}, \hat{\rho}^{(k)})$.
4. put k=k+1.
5. Repeat steps (2)to(4) B times to have $\hat{R}_k$ for $k = 1, 2, \ldots, B$.
6. Arrange $\hat{R}_k$ for $k = 1, 2, \ldots, B$ in ascending order as $\hat{R}_{(1)} \leq \hat{R}_{(2)}, \ldots, \leq \hat{R}_{(B)}$. Then the $100(1 - v)$ percentile bootstrap CI for R is given by $\left(\hat{R}_{(B(v/2))}, \hat{R}_{(B(1-v/2))}\right)$.

2.3 Bayesian Estimation

In this section, we consider Bayesian estimation of R for BVND under symmetric as well as asymmetric loss functions such as squared error, LINEX and entropy loss functions. Let $(X_{(i)}, Y_{[i]}), i = 1, 2, \ldots, r$ be the vector of order statistics and its concomitants arising from BVND$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Assume

that the prior distributions of $\mu_1|\sigma_1 \sim N(\mu_{01}, \sigma_1^2), \mu_2|\sigma_2 \sim N(\mu_{02}, \sigma_2^2), \sigma_1^2 \sim$ Inverse Gamma$(a, b), \sigma_2^2 \sim$ Inverse Gamma(c, d) and $\rho \sim u(-1, 1)$. Thus the joint prior distribution of $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ is given by

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2}\frac{a^b}{\Gamma(b)2^{b-1}}\frac{c^d}{\Gamma(d)2^{d-1}}\sigma_1^{-2b-1}\sigma_2^{-2d-1}$$
$$exp\frac{-1}{2}\{(\frac{\mu_1-\mu_{01}}{\sigma_1})^2 + \left(\frac{\mu_2-\mu_{02}}{\sigma_2}\right)^2 + \frac{a}{\sigma_1^2} + \frac{c}{\sigma_2^2}\}.$$

Then the joint posterior density of θ given data is obtained as

$$\pi^*(\theta|data) = \frac{L(\theta)\pi(\theta)}{\int_\theta L(\theta)\pi(\theta)d\theta}. \quad (6)$$

It is not possible to compute the Bayes estimates explicitly. Thus we propose importance sampling method to find the Bayes estimators for R. we consider the sequential importance sampling method to generate samples from the posterior distribution and then find the Bayes estimate of R (see, Tokdar and Kass, 2010). We also construct HPD interval under SEL for R as described in Chen and Shao (1999).

3. Results

3.1 Simulation Study

In this section, we carry out a simulation study to asses the performance of different estimators developed in previous sections. First we obtain the MLE of R using (5) for different combinations of μ_1, μ_2 and ρ by fixing $\sigma_1 = 1$ and $\sigma_2 = 1$. We have obtained the bias and MSE of MLEs. The bootstrap CI for R is also obtained. The average interval length (AIL) and coverage probability (CP) are also obtained. For the simulation study for Bayes estimators we took the hyper parameters $\mu_{01} = 2$, $\mu_{02} = 2$, $a = 2, b = 2, c = 2$ and $d = 2$. We have obtained the Bayes estimators for R of BVND under SEL, LL (with h=1) and EL (with q=1) functions.

We repeat the simulation procedure for different values of μ_1, μ_2, ρ and $r = 10, 15, 20$. The bias and MSE of Bayes estimators for different combinations of μ_1, μ_2 and ρ are obtained. The AIL and CP for HPD interval are also obtained. Based on the simulation studies we have the following results.

1. The bias and MSE of all estimators decrease when the number of uncensored observations r increases.
2. Bias and MSE of Bayes estimators are smaller than that of MLEs.
3. Among the Bayes estimators, estimator under SEL function possess minimum bias and MSE when $\mu_1 \neq \mu_2$ and estimator under EL function possess minimum bias and MSE when $\mu_1 = \mu_2$.

4. AILs of HPD intervals are smaller and the associated CPs are higher than that of bootstrap confidence intervals.

3.2 Illustration using real data

We also illustrate an application for the inferential procedures on $R = P(X < Y)$ developed in the previous sections. For that, we consider an example given in Hoffman and Johnson (2015) on water quality level of freshwater streams across the commonwealth of Virginia in USA. We have obtained the MLE and Bayes estimates of P(Y < X) based on first r order statistics of X observation and its concomitants on Y variates and shows that the estimated values of $P(Y < X)$ is more or less 0.9 for all the cases. Therefore we can claim that the concentration level of Zinc in freshwater across Virginia is relatively lower than the worldwide quality standard of zinc in water.

4. Conclusion

In this work, we have considered the problem of estimation of $R = P(X < Y)$ for bivariate normal distribution using type II censored order statistics and its concomitants. The maximum likelihood and Bayes estimators have been obtained for R. For obtaining the Bayes estimates, importance sampling method has been applied. Based on the simulation study we have the following conclusions. The bias and MSE of all estimators decrease when the number of uncensored observations r increases. Bayes estimators perform better than that of MLEs in terms of bias and MSE. Among the Bayes estimators, estimator under SEL function perform better when $\mu_1 \neq \mu_2$ and estimator under EL function perform better when $\mu_1 = \mu_2$. AILs of HPD intervals are smaller and the associated CPs are higher than that of bootstrap confidence intervals.

References

1. Abu-Salih, M. S., Shamseldin, A. A.(1988) Bayesian estimation of P(X < Y) for bivariate exponential distribution, Arab Gulf J. Sci. Res. A. Math. Phys. Sci., 6(1), 17-26.
2. Awad, A., Azzam, M., Hamdan, M. (1981) Some inference results on P(Y < X) in the bivariate exponential model, Communications in Statistics Theory and Methods, 10, 2515-2525.
3. Chen M.H., Shao Q.M. (1999) Monte Carlo estimation of Bayesian credible and HPD intervals, Journal of Computational and Graphical Statistics, 8, 69-92.
4. Cramer, E. (2001) Inference for stress-strength models based on wienman multivariate exponential samples, Communications in Statistics - Theory and Methods, 30, 331-346

5. David, H. A. (1973) Concomitants of order statistics. Bulletin of the International Statistical Institute 45, 295-300.
6. Enis, P., Geisser, S. (1971) Estimation of the probability that Y < X, Journal of American Statistical Association, 66, 162-186.
7. Hanagal, D.D. (1995) Testing reliability in a bivariate exponential stress-strength model, Journal of the Indian Statistical Association, 33, 41-45.
8. Hanagal, D.D. (1997) Estimation of reliability when stress is censored at strength, Communication in Statistics : Theory and Methods, 26(4), 911-919.
9. Hoffman, H. J., Johnson, R. E. (2015) Pseudo-likelihood Estimation of Multivariate Normal Parameters in the Presence of Left-Censored Data, Journal of Agricultural, Biological, and Environmental Statistics, 20, 156–171.
10. Jana, P.K. (1994) Estimation of P(Y < X) in the bivariate exponential case due to Marshall-Olkin, Journal of the Indian Statistical Association, 32, 35-37.
11. Jana,P.K., Roy,D. (1994) Estimation of reliability under stress-strength model in a bivariate exponential set-up, Calcutta Statistical. Association Bulletin, 44, 175-181.
12. Mukherjee, S.P. and Saran, L.K. (1985) Estimation of failure probability from a bivariate normal stress-strength distribution, Microelectronics and Reliability, 25, 692-702.
13. Tokdar, S. T., Tass, R. E. (2010) Importance sampling: a review Wiley Interdisciplinary Reviews: Computational Statistics 2, 54-60.

Bayesian inversion into soil types with Kernel-Likelihood Models

Selamawit Moja[1]; Henning Omre[2]
[1]Hawassa University, Hawassa, Ethiopia
[2]Norwegian University of Science and Technology, Trondheim, Norway

Abstract
Knowledge of the sub-surface characteristics is crucial in many engineering activities. Sub-surface soil classes must for example be predicted from indirect measurements in narrow drill holes and geological experience. In this study the inversion is made in a Bayesian framework by defining a hidden Markov chain. The likelihood model for the observations is assumed to be in factorial form and they are assessed from a calibration well by kernel estimators. The prior Markov model are defined to be either a traditional stationary Markov chain or a trend Markov chain. The methodology is demonstrated on one case study for offshore fundamentation of wind-mills. We conclude that a suitable choice of kernel likelihood model is of at most importance, and that using a trend Markov prior model improve the predictions even more.

Keywords
Sub-surface layer prediction; circular uniform kernel; Gaussian kernel

1. Introduction

In the previous study Moja et al. (2018), we constructed a prediction rule for the sub-surface facies characteristics from well-log observations. This rule is based on a non-stationary prior Markov chain model with a Gaussian likelihood model on factorial form. However, the Gaussian likelihood model does not capture the bimodal nature of our CPT data. In the current study we define a non-parametric kernel model in order to capture the bimodal nature of the observations which offers an alternative to traditional parametric models (Izenman 1991). We define a non-stationary prior Markov chain model with two different types of kernel likelihood models. These models are circular uniform kernel likelihood and Gaussian kernel likelihood. Therefore, the current study allow us to handle the bimodality nature of the data in the likelihood model. We compare results from the inversion based on the Gaussian likelihood, and the circular uniform kernel likelihood and Gaussian kernel likelihood model. The sensitivity of the likelihood model to the choice of kernel band width is also explored. Thereafter, an estimate of the optimal band width based on the maximum cross-validation pseudo-likelihood criterion is obtained, and the corresponding likelihood models and posterior pdfs are presented.

2. Model Definition

The data we consider for this study is a well log from Sheringham Shoal wind farm and we consider data only from one well. The well is discretized by a one-dimentional grid, $T : \{1, \ldots, T\}$. The categorical facies class $\kappa_t \in \Omega_k : \{1, \ldots, K\}$ is assigned to one of K classes. The facies profile along the discretized well is represented by $\kappa = (\kappa_1, \ldots, \kappa_T)'$ and the corresponding data, well-log observations, is denoted by $d = (d_1, \ldots, d_T)'$. The sub-surface layer prediction based on the observed well logs is solved by Bayesian inversion and settled in a Bayesian framework

$$p(\boldsymbol{\kappa}|\mathbf{d}) = \text{const} \times p(\boldsymbol{\kappa}|\mathbf{d})\, p(k) \tag{1}$$

where, $p(\mathbf{k}|\mathbf{d})$ is the likelihood model, $p(\boldsymbol{\kappa})$ is the prior model, and 'const' is a normalizing constant. The likelihood and prior models are defined and the corresponding posterior model is developed.

Observation likelihood

We assume the likelihood model on factorial form given by

$$p(\mathbf{d}|\boldsymbol{\kappa}) = \prod_{t \epsilon \tau} p(d_t|\kappa) = \prod_{t \epsilon \tau} p(d_t|k_t) \tag{2}$$

with an assumption about conditional independence and single site dependence. Here $p(\mathbf{d}|\boldsymbol{\kappa})$ is the likelihood model and it defines the link between the observations, d and the variable of interest, κ.

Prior experience

The prior pdf $p(\kappa)$ captures the primary knowledge we have about the variable of interest κ. In our study this information can primarily be obtained from certain geotechnical studies. Consider a first order Markov chain prior model which allows spatial coupling in the sub-surface profile. This coupling is represented by the conditional probability of the facies state κ_t, given all the previous states, which only depends on the single previous state $\kappa_t - 1$, and it is defined as

$$p(\kappa) = p(\kappa_1) \prod_{t \in \tau_{-1}} p(\kappa_t|\kappa_{t-1}, \ldots, \kappa_1) = p(\kappa_1) \prod_{t \in \tau_{-1}} p(\kappa_t|\kappa_{t-1}), \tag{3}$$

The prior pdf $p(\kappa)$ can be defined by the initial pdf $P_1 = [p(k_1)]_{k_1 \in \Omega_k}$ and the set of transition matrices $\mathbf{P}_{t-1,t} = [p(\kappa_t|\kappa_{t-1})]_{\kappa_{t-1},\ \kappa_t \in \Omega_\kappa}; t \in T_{-1}$. By recursion the set of marginal pdfs is defined by

$$\mathrm{P}_t = [p(\kappa_t)]_{\kappa_t \in \Omega_k} = \mathrm{P}'_{t-1,\ t} P_{t-1}; t \in \tau_{-1}. \tag{4}$$

Posterior model

The posterior pdf constitute the solution to the Bayesian inversion and it is defined by the prior pdf and the likelihood function. It can be presented as

$$\begin{aligned} p(\kappa|\mathrm{d}) &= \text{const} \times p(d_t|\kappa_t) \times p(\kappa_1) \prod_{t\in\mathcal{T}_{-1}} p(\kappa_t|\kappa_{t-1}) \\ &= \text{const} \times p(d_1|\kappa_1)p(\kappa_1) \times \prod_{t\in\mathcal{T}_{-1}} p(d_t|\kappa_t)p(\kappa_t|\kappa_{t-1}) \\ &= p(\kappa_1|\mathrm{d}) \prod_{t\in\mathcal{T}_{-1}} p(\kappa_t|\kappa_{t-1}, \mathrm{d}_{t:T}) \end{aligned} \tag{5}$$

$$\begin{aligned} p(\boldsymbol{\kappa}|\mathbf{d}) &= \text{const} \times \prod_{t\in\mathcal{T}} p(d_t|\kappa_t) \times p(\kappa_1) \prod_{t\in\mathcal{T}_{-1}} p(\kappa_t|\kappa_{t-1}) \\ &= \text{const} \times p(d_1|\kappa_1)\, p(\kappa_1) \times \prod_{t\in\mathcal{T}_{-1}} p(d_t|\kappa_t)\, p(\kappa_t|\kappa_{t-1}) \\ &= p(\kappa_1|\mathbf{d}) \prod_{t\in\mathcal{T}_{-1}} p(\kappa_t|\kappa_{t-1},\ \mathbf{d}_{t:T}). \end{aligned} \tag{5}$$

By defining a first order Markov chain prior model in Eq.3 and a factorial form likelihood model in Eq.2, our posterior pdf model will be a hidden Markov model. The posterior model will also be Markovian and it can be assessed by the highly efficient recursive Forward-Backward algorithm see Moja et al. (2018). The posterior pdf can be used to predict the facies profile by a MAP-criterion, κˆMAP . In order to evaluate the inversion results we may compare the predictions with the reference profile κ r . Two test statistics are defined,

$$c_1 = \frac{1}{T}\sum_{t=1}^{T} \mathbf{I}\left(\hat{\kappa}_{MAP_t} \neq \kappa_t^r\right) \tag{6}$$

and

$$c_2 = \frac{1}{K}\sum_{k=1}^{K} |\, \hat{P}_{k,k}^{MAP} - P_{k,k}^{r}\,| \tag{7}$$

where $\hat{P}_{k,k}^{MAP}$ and $P_{k,k}^{r}$ are the estimated diagonal terms of the transition matrix from the MAP predictions and the reference profile respectively. The statistics c1 represents the location wise mis-match between the prediction and the reference facies, while c_2 is defined by the absolute deviation between the diagonal terms in the transition matrix of the prediction and the reference transition matrix, see Lindberg et al. 2015. When the MAP predicts the facies profile well, both of the test statistics result in values close to zero.

Likelihood Model

The observations, normalized cone resistance and sleeve friction, are available along the well, and so is the true soil classes, see Fig.1a. These

observations provide the bi-variate display in Fig. 1b, where the colors of the points corresponds to the soil classes. The points for some soil classes appear in more than one cluster, which indicates a bimodal likelihood model $p(d_t|\kappa_t)$. In Moja et al. (2018) a Gaussian likelihood model is assumed which does not capture bimodality, see Fig.1b.

Consider the likelihood model $p(d|\kappa)$ as a pdf of $d \in R_2$ given $\kappa \in \Omega_k$, and denote the related observations $d^k = (d_1^k, d_2^k \cdots, d_{n_k}^k)$. Then $p(d|\kappa)$ may be inferred by a kernel-estimator, see Izenman 1991, defined as:

$$\hat{p}_k(\mathbf{d}|\kappa) = \frac{1}{n_\kappa h_{n_\kappa}} \sum_{i=1}^{n_\kappa} k\left(\frac{\mathbf{d} - \mathbf{d}_i^\kappa}{h_{n_\kappa}}\right);\ \kappa \in \Omega_\kappa \tag{8}$$

where $k(\tau)$; $\tau \in R^2$ is the kernel function and h_{n_k} is the band width which should be dependent on the number of observations. A frequently used measure for inference precision is the mean integrated square error (MISE), see Izenman 1991,

$$MISE_\kappa(h_{n_\kappa}) = E_p\left\{\int_{\Omega_d}\left[\hat{p}_k(\mathbf{d}|\kappa) - p(\mathbf{d}|\kappa)\right]^2 d\mathbf{d}\right\} \tag{9}$$

Certain asymptotic results are available for kernel estimators, see Izenman 1991. If the kernel function $k(\tau)$ is a pdf itself, then $\hat{P}_k(\mathbf{d}|\kappa)$ will always be a valid pdf, and if $h_{nk} \to 0$ for $n_k \to \infty, \hat{P}_k(\mathbf{d}|\kappa)$ is a consistent estimator for $p(\mathbf{d}|\kappa)$, regardless of $k(\tau)$ and $p(\mathbf{d}|\kappa)$. Moreover, the $MISE_k(h_{n_k}) \to 0$ at a rate of $o\left(n_k^{-1/3}\right)$. A suitable band width h_{n_k} may be determined by a cross validation psuedo-likelihood (CVL) approach. Define the estimator $\hat{P}_{k_{(-i)}}(\mathbf{d}|\kappa)$ as the kernel estimator based on the observation vector d_{-i}^k, hence with observation no i removed. Define the CV psuedo-likelihood of h_{n_k} by,

$$CVL(h_{n_\kappa}) = \prod_{i=1}^{n_\kappa} \hat{p}_{k_{(-i)}}(\mathbf{d}_i^\kappa|\kappa) \tag{10}$$

A reasonable estimator for the optimal band width is then

$$\hat{h}_{n_\kappa} = \arg\max_{h_{n_\kappa}} \{\ log\ CVL(h_{n_\kappa})\ \} \tag{11}$$

In the current study, the following test design is used:

Case A: Circular uniform kernel model

The likelihood estimates are denoted $\hat{p}u(\mathbf{d}|\kappa)$: $\kappa \in \Omega_k$, and they are based on a uniform kernel function,

$$k_U(\boldsymbol{\tau}) = \left[\pi r_0^2\right]^{-1} I\left(|\ \boldsymbol{\tau}\ | < r_0\right);\ \boldsymbol{\tau} \in \mathcal{R}^2, \tag{12}$$

with $r_0 = 0.2$ and $I(A)$ being the indicator function taking value 1 if A is true and value 0 else. The associated band width h_{n_k} remain to be defined.

Case B: Gaussian kernel model

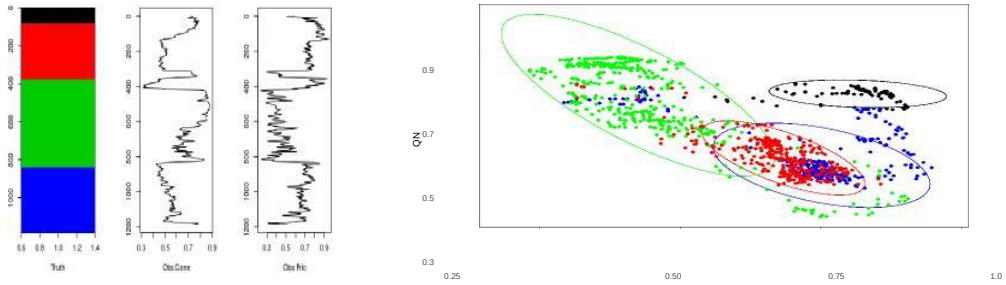
The likelihood estimates are denoted $\hat{\mathcal{P}}G(\mathbf{d}|\kappa): \kappa \in \Omega_k$, and they are based on a Gaussian kernel function,

$$k_G(\boldsymbol{\tau}) = \varphi_2\left(\boldsymbol{\tau};\ 0\mathbf{i}_n,\ \hat{\boldsymbol{\Sigma}}_\kappa\right);\ \boldsymbol{\tau} \in \mathcal{R}^2 \qquad (13)$$

with the (2 × 2)-matrix $\hat{\Sigma}_k$ being the empirical covariance matrix of the the corresponding set of observations d^k. The associated band width h_{n_k} remain to be defined. For the traditional Markov chain model with stationary class proportions along the profile, we use,

$$\mathbf{P} = \begin{bmatrix} 0.85 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.85 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.85 \end{bmatrix} \qquad \mathbf{p}_s = (0.25 \quad 0.25 \quad 0.25 \quad 0.25)'$$

As prior trend Markov chain model we use the marginals $p_t^0; t \in T$, and the transition matrix above as reference matrix P_r. The trend prior model $p(\kappa)$ is computed by the approach defined in Moja et al. (2018).



(a) (b)

Figure 1: Case study data and estimated likelihood model

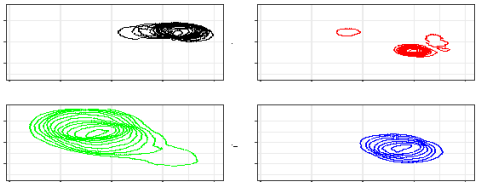
3. Results and Discussion

Firstly, we consider the major results presented in Moja et al. (2018) see Fig.??. The upper display presents results based on the traditional prior model, while the lower display is based on a trend prior model. Note, however, that both displays are based on a Gaussian likelihood model, see Fig.1b.

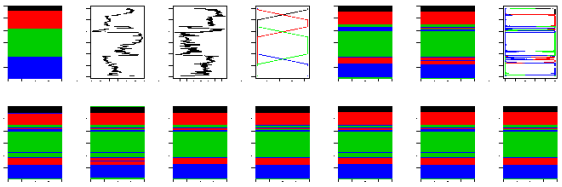
In this study we evaluate the effect of using two alternative kernel likelihood models with varying band widths. In order to measure the sensitivity

of the band width on posterior model, we have used three different level of band width which are low, optimum, and high respectively. Their performance are tested on prediction from the posterior model. The prediction will actually heavily depend on the choice of band width. The effect of varying band width on the circular uniform kernel likelihood model is illustrated in Fig.2. It has two parts, the upper part contains contour plots of the likelihood model for each soil class made from the circular uniform kernel while the lower part contains posterior model result based on the trend prior model and the corresponding likelihood model.

The reference profile $\boldsymbol{\kappa}^r$, observations $\mathbf{d}^0$, the prior marginal pdf profile, the MAP, MMAP, and the posterior probability profile $\mathbf{p}_k$ are presented in the lower display going from left to right respectively. Seven realizations which are sampled from the posterior pdf are also presented in the lower display in the same figure. The MAP, MMAP, and posterior marginal pdf based on circular uniform kernel likelihood model with optimum band width are presented in the upper right of Fig.2b. The prediction largely appear with the correct sequence of the classes compared to the true profile at the top left in the same figure. However, there appears to be some overlaps of the classes. For example, the red class is predicted in approximately half of the blue class at the bottom of the profile. The main reason for this may be found in the likelihood model. The red classes have masked some part of the blue classes. Observe also the bimodality in the likelihood model, see Fig.1b and 2a. The corresponding realization in the bottom line of Fig.2b are fairly similar, and display little uncertainty, resulting from the bimodal likelihood model see Fig.2a.

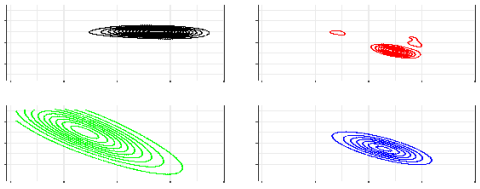


(a) Likelihood Model: Circular Uniform Kernel with Optimal bandwidth *r=0.2*

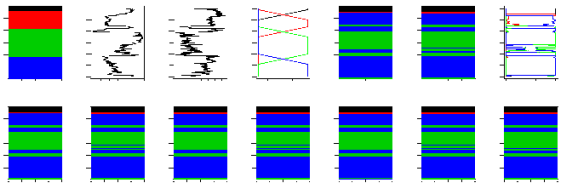


(b) Results from Trend Prior Model

Figure 2: Results from Bayesian Inversion of Case Study



(a) Likelihood Model: Gaussian Kernel with Optimal Band width



(b) Result from Trend Prior Model

Figure 3: Results from Bayesian Inversion of Case Study

Fig.3 contain results from the trend prior model with the Gaussian kernel likelihood model. The lay-out of the figure is the same as for Fig.2.

The predictions, MAP, MMAP, and posterior marginal pdf based on Gaussian kernel likelihood model with optimum band width are presented in Fig.3b. The inversion results are not very reliable since the blue class seems to replace the red class. The red class is almost not present. The posterior realizations displayed in the bottom line appears with little variability. 5. Conclusions Bayesian inversion for prediction of sub-surface facies profile based on the observed well log observations is made. The posterior model is assessed by the Forward-Backward and Viterbi algorithms.

For prediction of the sub-surface layer, both circular uniform kernel and Gaussian kernel likelihood model are defined and evaluated. The model parameter band width in the kernel likelihood is estimated by a cross-validation psuedo-

likelihood estimation criterion. The sensitivity of the kernel likelihood model are tested out on the prediction of posterior model based on the traditional and trend prior models.

The conclusion is that the trend prior model and the circular uniform kernel likelihood model provide the best prediction results. Using a reliable kernel likelihood model appears as the most important feature. Inversion based on Gaussian kernel likelihood models provide the by far less reliable predictions. Lastly, in order to have sufficiently good prediction of the sub-surface profiles, one may need to draw attentions to the choice of band width.

References

1. Baum. L.E., Petrie. T., Soules. G., Weiss. N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* 41(1):164-171
2. Eidsvik. J., Mukerji. T., Switzer. P. (2004). Estimation of geological attributes from a well log: An application of hidden Markov chains. *Mathematical Geology* 36(3):379-396
3. Izenman. A. (1991). Review papers: Recent developments in non parametric density estimation. *Journal of the American Statistical Association* 86(413):205-224.
4. Krumbein. Y.C., Dacey. M.F. (1969). Markov chains and embedded Markov chains in geology. *Mathematical Geology* 1(1):79-96 *Journal of Petroleum Science and Engineering* 134:237-246
5. Moja. S.S., Asfaw. Z.G., Omre. H. (2018). *Math Geosci.* https://doi.org/10.1007/ s11004-018-9752-z
6. Robertson. P.K. (2010). Soil behavior type from the CPT: an update. *California, USA*
7. Viterbi. A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory* 13(2):260-269

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

Fethi Şaban Özbek
Turkish Statistical Institute

Abstract

Forecasting agricultural prices is useful for farmers, policymakers, and agribusiness industries. In this study, wheat, barley, maize (seed), and raw cotton, widely sown in arable land of Turkey (57% of total agricultural area excluding fallow area), were selected for forecasting by using artificial neural networks (ANN). The data of monthly price from 2009 January to 2018 September were used as sample data (network training) and historical prediction evaluation (network evaluation). In each forecasting, the input was the last observation, and the output was the predicted value of agricultural commodity price. ANN models were effective for forecasting agricultural commodity prices in that all accuracies were very high; 99.2, 90.9, 99.0, 91.7 for wheat, barley, maize (seed), and raw cotton, respectively. The results of forecasting models show that the prices of wheat fluctuate between 0.97 TL/KG and 0.93 TL/kg between 2018-October and 2019-December. And the prices fluctuate between 0.84 TL/KG and 0.81 TL/kg, 0.87 TL/kg and 0.84 TL/kg, 2.26 TL/kg and 2.24 TL/kg for maize (seed), barley and raw cotton, respectively.

Keywords

Agricultural prices; Artificial neural network; Forecasting; Turkey agriculture

1. Introduction

It is well known that forecasting agricultural prices is useful for farmers, policymakers, and agribusiness industries. Wheat, barley, maize (seed), and raw cotton, widely sown in arable land of Turkey (57% of total agricultural area excluding fallow area), were selected for forecasting.

In recent years, machine learning techniques, such as decision trees and artificial neural networks (ANN), because they are quick, powerful, and flexible tools for prediction, classification, optimization, and decision support system, are used increasingly in agriculture. ANN has provided great benefits to the researchers for forecasting in economics and finance (Zhang et al., 1998; Jha et al., 2009), particularly in agricultural price forecasting (Li et al., 2010; Jha and Sinha, 2013). Also, artificial neural network (ANN) has been proposed as an efficient tool for modelling and forecasting (Ozbek and Fidan, 2008; Li et al., 2010; Jha and Sinha, 2013; Al-Maqaleh et al., 2016 etc.).

The purpose of this paper is to evaluate the effectiveness of artificial neural network as a forecasting tool, and to forecast the agricultural commodity prices in Turkey for the period from October 2018 to December 2019.

2. Methodology

ANN, represents a nonlinear statistical modelling tool that is based on the concept of a biological neural network (Shahinfar et al., 2012), was used for forecasting prices of wheat, barley, maize (seed), and raw cotton in the agriculture of Turkey. The data of monthly price from 2009 January to 2018 September were used as sample data (network training) and historical prediction evaluation (network evaluation). In each forecasting, the input was the last observation, and the output was the predicted value of agricultural commodity price.

Different ANN models for different problem structures have been developed (Akkol et al., 2017). The most used networks in literature are known as single and multilayer perception, vector quantization models (LVQ), self-organizing model (SOM), adaptive resonance theory models (ART), Hopfield networks, Elman network and radial based networks (Öztemel, 2006). In this study, the multilayer perception model of ANN was used. The multilayer perception network is a function of predictors (also called inputs or independent variables) that minimize the prediction error of target variables (also called outputs). This structure is known as feedforwardarchitecture because the acquaintances in the network flow forward from the input layer to the output layer without any feedback loops (Giri et al., 2014).

Node in the hidden layer contains hyperbolic tangent activation function (Eq. 1). It takes real-valued arguments and transforms them to the range (–1, 1).

$$f(y_i) = \frac{e^{y_i} - e^{-y_i}}{e^{y_i} + e^{-y_i}} \qquad \text{Eq. 1}$$

And they take a weighted sum of all input variables:

$$y_i = \sum_j w_{ji} x_i \qquad \text{Eq. 2}$$

where x_i is an input variable and w_{ji} is corresponding weight in layer j.

Identity function was used in output layer activation. Identity function has the form in Eq. 3. It takes real-valued arguments and returns them unchanged (IBM SPSS Neural Networks 22, 2018).

$$f(y_i) = y_i \qquad \text{Eq. 3}$$

IBM's SPSS Modeler 15 software package was used for developing and running for all networks. This package automatically selects an optimal network structure and settings. However, networks with alternate structures were also tested by using the same inputs for all types of crops. The networks did not improve the performance, so the Modeler suggested form with one hidden layer of one node was used. Thus, for each multilayer perception, there were one input, one hidden layer with one node, and one output. In this structure, the value used for the random seed was 229176228, and 30% of the training set was used to prevent over-fitting. The training algorithm used by Modeler stops after 15 minutes, or when the error in the over-fit prevention set does not decrease after each cycle, if the relative change in the training error is small, or if the ratio of the current training error is small compared to the initial error (Malliaris and Malliaris, 2013).

3. Results

The detailed information on ANN models are given below:

Fields
- Target
  - Price
- Predictors(Inputs)
  - Price -1
- Use partitioned data: false

Build Options
- Objectives
  - What do you want to do?: Build new model
  - What is your main objective?: Create a standard model
- Basics
  - Neural network model: Multilayer Perceptron (MLP)
  - Hidden Layers: Automatically compute number of units
  - Hidden layer 1: 1
  - Hidden layer 2: 0
- Stopping Rules
  - Use maximum training time (per component model): true
  - Minutes: 15
  - Customize number of maximum training cycles: false
  - Use minimum accuracy: false
- Advanced
  - Overfit prevention set(%): 30.0
  - Replicate Results: true
  - Random seed: 229176228
  - Missing values in predictors: Delete listwise

Training Summary

Method: Neural Networks
Records used in training: 116
Model type: Classification
User: ---
Application: IBM SPSS Modeler Common 15.0.0.3
Date built: October 31, 2018 2:36:57 PM AST
Predictors used in model
Price -1

ANN models were effective for forecasting agricultural commodity prices in that all accuracies were very high; 99.2, 90.9, 99.0, 91.7 for wheat, barley, maize (seed), and raw cotton, respectively.

Applying the above models of ANN, the forecasting results for the period from October 2018 (2018-Oct.) to December 2019 (2019-Dec.) are shown in Figure 1. The results of forecasting models show that the prices of wheat fluctuate between 0.97 TL/KG and 0.93 TL/kg between 2018-October and 2019-December. And the prices fluctuate between 0.84 TL/KG and 0.81 TL/kg, 0.87 TL/kg and 0.84 TL/kg, 2.26 TL/kg and 2.24 TL/kg for maize (seed), barley and raw cotton, respectively.

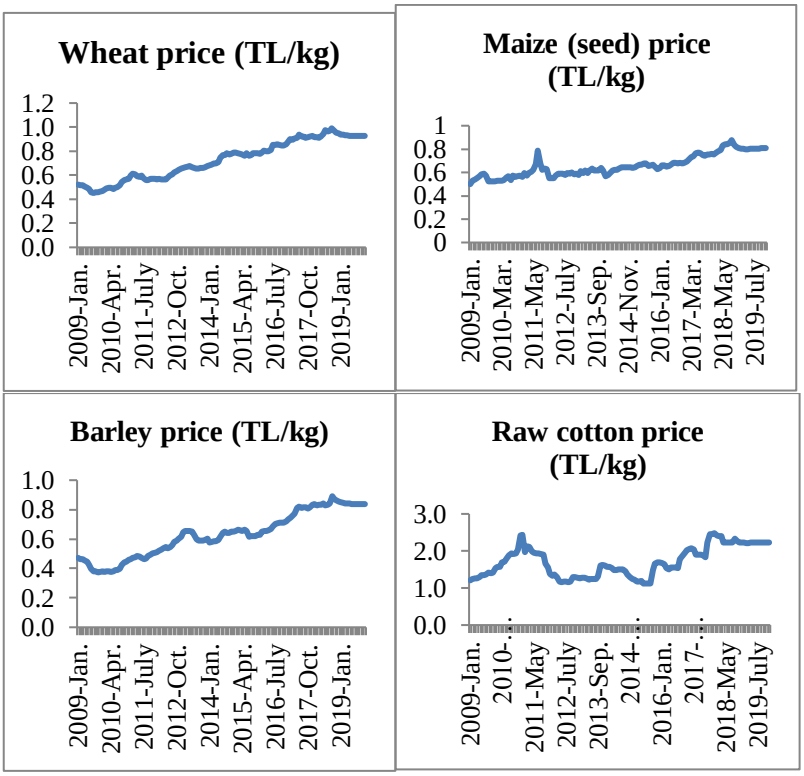


Fig. 1. Prices of wheat, barley, maize (seed), and raw cotton, 2009-Jan.-2019-Dec., TL/kg

4. Discussion and Conclusion

Considering high effect of agricultural commodities on inflation in Turkey, it is needed efficient and reliable food price forecasting models to develop efficient policies to combat inflation. Following of price fluctuation in food products is highlighted in the New Economic Programme of Turkey (2019-2021). The results showed that the model accuracies are good enough to forecast prices of agricultural commodities, so the model can be used in forecasting and evaluating the agricultural prices.

It is also crucial to analyse future prices in terms of observing price anomalies, which is an important indicator of global sustainable development goals (SDG). This study ensures evaluating future price anomalies of agricultural commodities in Turkey. According to model results, there are no price anomalies for wheat, barley, maize, and raw cotton up to December 2019.

For all crop types, the model estimates the prices will decrease after September 2018. This can be explained that the prices decrease over years after they reach the peak values as seen from Fig. 1. The government take some measures to decrease producer prices of agricultural commodities e.g. providing subsidies. This makes positive effect on decreasing prices.

References

1. Akkol. S., Akıllı. A.& Cemal. İ. (2017). Comparison of Artificial Neural Network and Multiple Linear Regression for Prediction of Live Weight in Hair Goats. YYÜ TAR BİL DERG (YYU J AGR SCI). 27(1): 21-29.
2. Al-Maqaleh. B. M., Al-Mansoub. A. A. & Al-Badani. F. N. (2016). Forecasting using Artificial Neural Network and Statistics Models. I.J. Education and Management Engineering. 2016 (3): 20-32.
3. Giri. J. P., Tatwawadib. V. H., Mahakalkar. S. G. & Modak. J. P. (2014). Comparative analysis of Multilayer Perceptron and Radial Basis Function ANN for prediction of cycle time of Structural subassembly Manufacturing. International Journal of Science, Spirituality, Business and Technology (IJSSBT). 3 (1).
4. Jha. G. K., Sinha. K. (2013). Agricultural Price Forecasting Using Neural Network Model: An Innovative Information Delivery System Agricultural Economics Research Review. 26:2.
5. Jha. G. K., Thulasiraman. P. & Thulasiram. R. K. (2009). PSO based neural network for time series forecasting. Proceedings of the International Joint Conference on Neural Networks. Atlanta, USA. pp. 1422-1427.
6. Li. G., Xu. S. & Li. Z. (2010). Short-Term Price Forecasting For Agro-products Using Artificial Neural Networks. Agriculture and Agricultural Science Procedia. 1 (2010): 278–287.

7. Malliaris. M. & Malliaris. A. G. (2013). Neural Network Forecasting with the S&P 500 Index Conference Proceedings.
8. Özbek. F. Ş. & Fidan. H. (2009). Estimation of pesticides usage in the agricultural sector in Turkey using Artificial Neural Network (ANN). Journal of Animal & Plant Sciences. 4 (3): 373 – 378.
9. Öztemel, E. (2006). Yapay Sinir Ağları. İstanbul, Papatya Yayıncılık.
10. Shahinfar. S., Mehrabani-Yeganeh. H., Lucas. C., Kalhor. A., Kazemian. M. & Weigel. K. A. (2012). Prediction of Breeding Values for Dairy Cattle Using Artificial Neural Networks and Neuro-Fuzzy Systems. Computational and Mathematical Methods in Medicine Volume 2012.
11. IBM SPSS Neural Networks 22 (2018). http://www.sussex.ac.uk/its/pdfs/SPSS_Neural_Network_22.pdf. Accessed on 12.11.2018.
12. Zhang. G., Patuwo. B. E. & Hu. M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting. 14: 35-62.

Bayesian inference of multiple structural breaks in multiple regimes threshold autoregressive model

Varun Agiwal, Jitendra Kumar
Central University of Rajasthan

Abstract

This paper provides a Bayesian setup of multiple regimes threshold autoregressive model with possible break points. A full conditional posterior distribution is obtained for all model parameters using the suitable prior information's, including threshold and break point variables that are not attain standard form distributions. In order to compute posterior distributions, we applied Gibbs sampler with Metropolis-Hastings algorithm. A variety of loss function is considered for optimizing the risk associated with each parameter. For empirical evidence, simulation study and real data illustration are carried out.

Keywords

Threshold autoregressive model; Prior & Posterior Distribution; MCMC method; Structural Break

1. Introduction

Linear time series model are very popular for investigation of dynamic structure of a series over time. There are several linear time series model which adequately explore the characteristics such as stationarity, unit root, de-trending, co-integration and so on to recognize the process efficiency and improving the prediction (Dickey and Fuller(1979), Nelson & Plosser (1982), Watson (1986)). However, series is non-linear in nature due to non-stationarity or non-Gaussian of the stochastic trend. This possibility often arises during the permanent change (structural break), temporary effect (outlier) or varying structure relationship. For that reasons, a range of non linear time series model was introduced and used for better analysis. The concept of break point(s) in time series addressed by Albert and Chib (1993), Wang and Zivot (2000), Meligkotsidou *et al.* (2011) in various univariate and multivariate AR model for detection, estimation and testing purpose. All are considering structural breaks in the time horizon. Although, series generation framework may change during own past realizations. First, Tong (1983) introduced a standard non linear time series model known as threshold autoregressive model (TAR) and Chan (1993) derived the limiting distribution of the least square estimator. Gonzalo and Wolf (2005) proposed a subsampling methodology to obtain the consistent confidence interval of threshold and regression parameter when

model is discontinuous. Liu *et al.* (2011) considered the limiting distributions of the least-squares estimators for non-stationary TAR model and Li *et al.* (2011) extended Liu *et al.* (2011) works for TARMA model. Rather than classical methodology, it is also attracting researchers under Bayesian framework. Chen and Lee (1995) obtained its Bayesian estimation for two regimes through Gibbs sampler and M-H algorithm. Chen (1998) and Charif (2003) constructed a Bayesian framework for generalized TAR model and TMA model, respectively. Yu (2012) obtained the likelihood based inference in threshold regression model that allow heteroskedasticity and threshold effect in both mean and variance. Xia *et al.* (2012) generalized TARMA with explanatory variables model and considered iterative least square and two stage MCMC methods for estimation. Pan et al. (2017) discussed the multiple thresholds autoregressive model and obtained the threshold dependent sequence using stochastic search selection method. TAR model is also extended with structural break when change in state space parallel occurs with change in time domain. This is very serious problem in time series which is not much explored by researchers. Yau *et al.* (2015) derived a minimum descriptive length principle to estimate the TAR parameters and detecting the breakpoints. Gao and Ling (2018) considered the least square estimation of TAR model with structural break and shows that threshold and break points are n-consistent and converge weakly to a Poisson process and two-sided random walk respectively.

Above literature shows that inference about TAR model with structural change was done in classical approach but no one have perform this under Bayesian framework. So, the main objective of this paper is to develop a Bayesian setup of multiple regimes TAR model with multiple structural breaks. A conditional posterior distribution is obtained for parameter estimation form a Bayesian point of view. For building better inference, we used various symmetric and asymmetric loss functions. A simulation and empirical study is carried out using Gibbs sampler and N-H algorithm techniques to record the performance of the Bayesian perspective of multiple breaks and multiple regimes TAR model.

2. Threshold Autoregressive model with Structural Break

Let $\{y_t ; t=1,2,\ldots,T\}$ be a stochastic process having a multiple unknown change points and each change point interval follows a multiple-regimes threshold autoregressive (TAR) model. Then, the mathematical form of m-breaks and k-regimes TAR model (MB-TAR(m,k)) is obtained as

$$y_t = \theta^{(ij)} + \sum_{l=1}^{p_{ij}} \phi_l^{(ij)} y_{t-l} + e_t^{(ij)} \qquad r_{ij-1} < y_{t-d_i} \le r_{ij} \quad T_{i-1} < t \le T_i \tag{1}$$

for i=1,2,...,m; j=1,2,...,k_i, where d_i is the i^{th} delay parameter of the TAR model acquire some positive integer value from $\{1,2,\ldots,d_{i0}\}$, d_{i0} is the maximum delay

lag considered and r_{ij} is the threshold parameter of the j^{th} regime of the i^{th} break point constitute a partition on the real line and satisfy $-\infty = r_{i0} < r_{i1} < \cdots < r_{ik_i} = \infty$. The locations of structural breaks are $0 = T_0 < T_1 < \cdots < T_m = T$ that partition the time series into TAR models. The number of structural breaks (m) having partition in such a way that each break segment having k_i regimes TAR model with delay parameter d_i. The error term $\left\{e_t^{(ij)}\right\}$is independent and identically normal distributed random variable with mean zero and unknown variance σ_{ij}^2 . Also, TAR model coefficients $\left\{\phi_l^{(ij)}\right\}$and lags order $\{p_{ij}\}$ are also different with respect to change in threshold and delay parameter as well as the presence of break points. In time series, sequential observations is valuable for making effective statistical inference so reorganized the MB-TAR generating observations in a group of each regimes for a particular break interval. This separation appears in an order form and does not need to know the threshold value. Let π_{ij} be the time index of the j^{th} smallest observation in i^{th} break point in this series. Then, each regime has (s_{ij}-s_{ij-1}) observations from (y_{p+1-d}, y_{p+2-d},..., y_{Ti-d}) where p=max(p_{ij}; i=1,2,...,m; j=1,2,...,k_i). Therefore, we can rewrite the model (1) as given by

$$y_{\pi_{w_{ij}}+d_i} = \theta^{(ij)} + \sum_{l=1}^{p_{ij}} \phi_l^{(ij)} y_{\pi_{w_{ij}}+d_i-l} + e^{(ij)}_{\pi_{w_{ij}}+d_i} \qquad s_{ij-1} < w_{ij} \le s_{ij} \qquad T_{i-1} < s_{ij} \le T_i \tag{2}$$

Given the first p-observations, likelihood function of the model (2) can be obtained as

$$L(\Theta \mid y) = \prod_{i=1}^{m}\prod_{j=1}^{k_i}\left(\left(2\pi\sigma_{ij}^2\right)^{-\frac{s_{ij}-s_{ij-1}}{2}} \exp\left[-\frac{1}{2\sigma_{ij}^2} \sum_{w_{ij}=s_{ij-1}+1}^{s_{ij}} \left(y_{\pi_{w_{ij}}+d_i} - \theta^{(ij)} - \sum_{l=1}^{p_{ij}} \phi_l^{(ij)} y_{\pi_{w_{ij}}+d_i-l} \right)^2 \right] \right)$$

$$\propto \prod_{i=1}^{m}\prod_{j=1}^{k_i}\left(\left(\sigma_{ij}^2\right)^{-\frac{n_{ij}}{2}} \exp\left[-\frac{1}{2\sigma_{ij}^2} \left(Y_{ij} - \Phi^{(ij)} X_{ij}\right)' \left(Y_{ij} - \Phi^{(ij)} X_{ij}\right) \right] \right) \tag{3}$$

where s_{ij} satisfy $y_{\pi_{s_{ij}}} \le r_{ij} < y_{\pi_{s_{ij}+1}}$, $s_{i0} = T_{i-1}, s_{ik_i} = T_i$, $n_{ij} = s_{ij} - s_{ij-1}$;

$$X_{ij} = \begin{pmatrix} 1 & y_{\pi_{s_{ij-1}+1}+d_i-1} & \cdots & y_{\pi_{s_{ij-1}+1}+d_i-p_{ij}} \\ 1 & y_{\pi_{s_{ij-1}+2}+d_i-1} & \cdots & y_{\pi_{s_{ij-1}+2}+d_i-p_{ij}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_{\pi_{s_{ij}}+d_i-1} & \cdots & y_{\pi_{s_{ij}}+d_i-p_{ij}} \end{pmatrix} \quad Y_{ij} = \left(y_{\pi_{s_{ij-1}+1}+d_i}, y_{\pi_{s_{ij-1}+2}+d_i}, \ldots, y_{\pi_{s_{ij}}+d_i} \right);$$

3. Bayesian Inference

In general, Bayesian inference provides some additional information about the unknown parameter rather than existing data information named as prior information. Selection of a prior is a main assignment in Bayesian study because on the basis of suitable prior, distinguish characteristic of the

unknown parameters. However, present study targets only for estimation, so we have considered the following priors for the parameters which are as listed below.

(i) $\Phi^{(ij)}$ follows an independent multivariate normal distribution $N\left(\Phi_0^{(ij)}, \sigma_{ij}^2 I_{ij}\right)$.

(ii) σ_{ij}^2 follows an independent inverse gamma distribution $IG\left(u_{ij}, v_{ij}\right)$.

(iii) r_{ij} follows a uniform distribution $U\left(a_{ij}, b_{ij}\right)$.

(iv) d_i follows a discrete uniform distribution $\{1,2,\ldots,d_{0i}\}$.

(v) Break point location (T_B) is equal to $P(T_B \mid m) = \dfrac{1}{\binom{T-1}{m}}$

In Bayesian inference, joint posterior distribution of the proposed model is expressed by combining the likelihood function with assumed prior distributions.

$$P\left(\Phi^{(ij)}, \sigma_{ij}^2, r_{ij}, d_i, T_B \mid Y\right) = \frac{\prod_{i=1}^{m} d_{i0}^{-1}}{\binom{T-1}{m}} \prod_{i=1}^{m} \prod_{j=1}^{k_i} \left(\frac{(2\pi)^{-\frac{n_{ij}+p_{ij}+1}{2}} \left(\sigma_{ij}^2\right)^{-\frac{n_{ij}+p_{ij}+1}{2}+u_{ij}+1} v_{ij}^{u_{ij}}}{\left(b_{ij}-a_{ij}\right)\Gamma(u_{ij})} \exp\left[-\frac{1}{2\sigma_{ij}^2} \left\{ \left(Y_{ij} - \Phi^{(ij)} X_{ij}\right)' \right.\right.\right. \tag{4}$$

$$\left.\left.\left. \left(Y_{ij} - \Phi^{(ij)} X_{ij}\right) + \left(\Phi^{(ij)} - \Phi_0^{(ij)}\right)' I_{ij}^{-1} \left(\Phi^{(ij)} - \Phi_0^{(ij)}\right) + 2v_{ij} \right\} \right] \right)$$

For getting the estimator, conditional posterior distributions for each parameter are derived by solving (4). The forms of conditional posterior distributions are obtained as given below

- The conditional posterior distribution of independent $\Phi^{(ij)}$ follows multivariate normal distribution having form

$$\Phi^{(ij)} \mid Y, \sigma_{ij}^2, r_{ij}, d_i, T_B \sim N\left(\left(X_{ij}' Y_{ij} + \Phi_0^{(ij)} I_{ij}^{-1}\right)\left(X_{ij}' X_{ij} + I_{ij}^{-1}\right)^{-1}, \sigma_{ij}^2 \left(X_{ij}' X_{ij} + I_{ij}^{-1}\right)^{-1} \right) \tag{5}$$

- The conditional posterior distribution of independent σ_{ij}^2 follows inverse gamma distribution having form

$$\sigma_{ij}^2 \mid Y, \Phi^{(ij)}, r_{ij}, d_i, T_B \sim IG\left(\frac{n_{ij}+p_{ij}+1}{2} + u_{ij}, S_{ij} \right) \tag{6}$$

- The conditional posterior distribution of r_{ij} is proportion to likelihood function

$$P\left(r_{ij} \mid Y, \Phi^{(ij)}, \sigma_{ij}^2, d_i, T_B\right) \propto \left(\sigma_{ij}^2\right)^{-\frac{n_{ij}}{2}} \exp\left[-\frac{1}{2\sigma_{ij}^2} \left(Y_{ij} - \Phi^{(ij)} X_{ij}\right)' \left(Y_{ij} - \Phi^{(ij)} X_{ij}\right) \right] \tag{7}$$

- The conditional posterior distribution of d_i is a multinomial distribution with probability mass function is

$$P\left(d_i \mid Y, \Phi^{(ij)}, \sigma_{ij}^2, r_{ij}, T_B\right) = \frac{L(\Theta \mid y)}{\sum_{d_i=1}^{d_{0i}} L(\Theta \mid y)} \tag{8}$$

- The conditional posterior distribution of T_m is

$$P\left(T_B \mid Y, \Phi^{(ij)}, \sigma_{ij}^2, r_{ij}, d_i, m\right) = \frac{P\left(T_B, Y \mid \Phi^{(ij)}, \sigma_{ij}^2, r_{ij}, d_i, m\right)}{P\left(Y \mid \Phi^{(ij)}, \sigma_{ij}^2, r_{ij}, d_i, m\right)} \tag{9}$$

where

$$S_{ij} = \frac{1}{2}\left[\left(Y_{ij} - \Phi^{(ij)} X_{ij}\right)'\left(Y_{ij} - \Phi^{(ij)} X_{ij}\right) + \left(\Phi^{(ij)} - \Phi_0^{(ij)}\right)' I_{ij}^{-1}\left(\Phi^{(ij)} - \Phi_0^{(ij)}\right) + 2v_{ij}\right]$$

$$P\left(T_B, Y \mid \Phi^{(ij)}, \sigma_{ij}^2, r_{ij}, d_i, m\right) = \frac{\prod_{i=1}^{m} d_{i0}^{-1}}{\binom{T-1}{m}} \prod_{i=1}^{m} \prod_{j=1}^{k_i} \frac{(2\pi)^{-\frac{n_{ij}}{2}} v_{ij}^{u_{ij}} \Gamma\left(\frac{n_{ij}}{2} + u_{ij}\right)}{\left(b_{ij} - a_{ij}\right)\Gamma(u_{ij})\left|X_{ij}' X_{ij} + I_{ij}^{-1}\right|^{\frac{1}{2}}} \left[\frac{1}{2}\left\{Y_{ij}' Y_{ij} + 2v_{ij} + \right.\right.$$

$$\left.\left. \Phi_0^{(ij)'} I_{ij}^{-1} \Phi_0^{(ij)} - \left(X_{ij}' Y_{ij} + I_{ij}^{-1} \Phi_0^{(ij)}\right)'\left(X_{ij}' X_{ij} + I_{ij}^{-1}\right)^{-1}\left(X_{ij}' Y_{ij} + I_{ij}^{-1} \Phi_0^{(ij)}\right)\right\}\right]^{-\frac{n_{ij}}{2} + u_{ij}}$$

$$P\left(Y \mid \Phi^{(ij)}, \sigma_{ij}^2, r_{ij}, d_i, m\right) = \sum_{T_1=1}^{T-m} \cdots \sum_{T_m = T_{m-1}+1}^{T-1} P\left(T_B, Y \mid \Phi^{(ij)}, \sigma_{ij}^2, r_{ij}, d_i, m\right)$$

From equation (5) to (9), we observed that except threshold and break location parameters, conditional posterior distributions of other parameters are come out in standard distribution form. For analysis purpose, random walk Metropolis-Hastings (M-H) algorithm is used for computing the estimates of non standard distribution form parameters where as applied Gibbs sampler method for standard form distribution. For better selection of an estimator, consider various loss functions which explain the minimum average risk associated with an estimator. Therefore, we have taken squared error loss function (SELF), absolute loss function (ALF) and Linex loss function (LLF) for better inference about the parameters.

4. Real Data Analysis

This section implemented the proposed methodology to a real data series. We want to draw inference from yearly time series of tree rings of Qilian Juniper which is taken from the northeastern Tibetan Plateau region of China. This data set is obtained from the NOAA paleoclimatology database and consist 930 observations from the period 1079 to 2009. For the same series, Gao and Ling (2018) employed a TAR and structurally changed TAR model for making classical inference. First, they recorded the results of non-linearity, then an AR-order and delay parameter was summarized by information criterion that consist single regime TAR with d=8. After that, structural break was found at point 578 and write down the estimated values of parameters using least square estimation. With the help of these findings, we are modeling this under Bayesian setup. Here, we applied our proposed methodology to identify the break point and then estimated the model parameters. For the given time series, first identify the break point through posterior probability where it attains maximum value. Using the conditional distribution of T_m, recorded the probability of each time point shown in Figure 8. From this figure,

one can observe that maximum probability is at time point 572 which is near to the 578 break point identified by Gao and Ling (2018). For Bayes estimation, use of their estimates as an initial value of parameters to generate the posterior samples. We carried out 1000 iterations and burn-in 200 to get the approximate estimated values and recorded that generate sample is convergent and stationary using Gelman-Rubin and Geweke test. Then, the corresponding outcomes of posterior estimators and its standard deviation are recorded in Table 1-2. From Table 1-2, one can easily conclude that standard deviation is not much wider and its truly indicating the estimated parameter values. We also obtained the credible interval in table 3 for establishing the confidence interval of the estimated value and observed that some parameters have not much significant affect on the series because its intervals contain zero value.

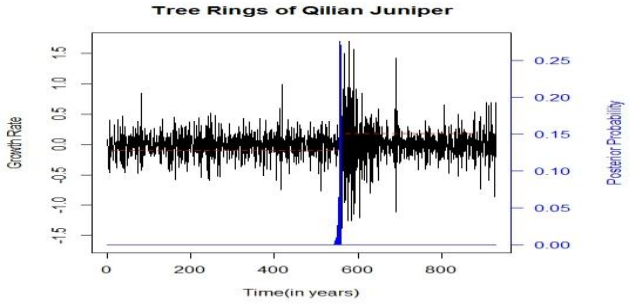


Figure 1: Tree ring series with posterior distribution of the break date

Table 5: Bayes estimate for TAR model parameters in a Tree ring series							
Parameter	SELF	ALF	LLF	Parameter	SELF	ALF	LLF
r_1	-0.0959	-0.1081	-0.0977	r_2	0.1772	0.1664	0.1754
θ_{11}	-0.014	-0.0143	-0.0169	θ_{21}	0.0005	0.0002	0
$\phi_1^{(11)}$	-0.5414	-0.5437	-0.5501	$\phi_1^{(21)}$	-0.5394	-0.5378	-0.5489
$\phi_2^{(11)}$	-0.3431	-0.3449	-0.3545	$\phi_2^{(21)}$	-0.3256	-0.3223	-0.3382
$\phi_3^{(11)}$	-0.0993	-0.1007	-0.1138	$\phi_3^{(21)}$	-0.1995	-0.1977	-0.2126
$\phi_4^{(11)}$	0.1091	0.1105	0.0949	$\phi_4^{(21)}$	-0.134	-0.135	-0.1482
$\phi_5^{(11)}$	0.0039	0.0032	-0.0074	$\phi_5^{(21)}$	0.0237	0.0241	0.0111
$\phi_6^{(11)}$	0.0669	0.0687	0.0581	$\phi_6^{(21)}$	-0.0564	-0.0572	-0.068
$\phi_7^{(11)}$	0.2512	0.2511	0.2437	$\phi_7^{(21)}$	-0.0146	-0.0137	-0.0261
$\phi_8^{(11)}$	0.1228	0.1221	0.1096	$\phi_8^{(21)}$	0.0416	0.0396	0.0255
σ_{11}^2	-0.0189	-0.0187	-0.0191	σ_{21}^2	-0.0373	-0.0361	-0.0423
θ_{12}	-0.4767	-0.4774	-0.4784	θ_{22}	-0.7211	-0.722	-0.7473
$\phi_1^{(12)}$	-0.3158	-0.3161	-0.3178	$\phi_1^{(22)}$	-0.4252	-0.4237	-0.4711
$\phi_2^{(12)}$	-0.2031	-0.2027	-0.2052	$\phi_2^{(22)}$	-0.2185	-0.2201	-0.2673
$\phi_3^{(12)}$	-0.178	-0.1782	-0.1802	$\phi_3^{(22)}$	0.1218	0.1241	0.0789
$\phi_4^{(12)}$	-0.2775	-0.2768	-0.28	$\phi_4^{(22)}$	0.1825	0.1851	0.1308
$\phi_5^{(12)}$	-0.2117	-0.2114	-0.2145	$\phi_5^{(22)}$	-0.1701	-0.1703	-0.23
$\phi_6^{(12)}$	-0.1317	-0.1311	-0.1342	$\phi_6^{(22)}$	-0.0052	-0.0043	-0.0581
$\phi_7^{(12)}$	0.1154	0.116	0.1127	$\phi_7^{(22)}$	0.0767	0.0753	0.0389
$\phi_8^{(12)}$	0.2314	0.2298	0.2309	$\phi_8^{(22)}$	0.138	0.1378	0.1379
σ_{12}^2	0.1247	0.125	0.1247	σ_{22}^2	0.2953	0.2921	0.2939

Table 8: Bayesian credible interval of fitted TAR model parameters							
	SELF	ALF	LLF		SELF	ALF	LLF
r_1	(-0.11, -0.08)	(-0.13, -0.08)	(-0.12, -0.08)	r_2	(0.16, 0.20)	(0.14, 0.20)	(0.16, 0.20)
θ_{11}	(-0.03, 0.00)	(-0.03, 0.00)	(-0.03, -0.00)	θ_{21}	(-0.01, 0.01)	(-0.01, 0.01)	(-0.01, 0.01)
$\phi_1^{(11)}$	(-0.57, -0.52)	(-0.57, -0.52)	(-0.57, -0.53)	$\phi_1^{(21)}$	(-0.56, -0.51)	(-0.57, -0.51)	(-0.57, -0.52)
$\phi_2^{(11)}$	(-0.37, -0.31)	(-0.38, -0.30)	(-0.38, -0.33)	$\phi_2^{(21)}$	(-0.36, -0.29)	(-0.36, -0.29)	(-0.37, -0.31)
$\phi_3^{(11)}$	(-0.13, -0.07)	(-0.14, -0.06)	(-0.15, -0.08)	$\phi_3^{(21)}$	(-0.23, -0.17)	(-0.23, -0.16)	(-0.24, -0.18)
$\phi_4^{(11)}$	(0.08, 0.14)	(0.07, 0.15)	(0.06, 0.13)	$\phi_4^{(21)}$	(-0.17, -0.10)	(-0.18, -0.10)	(-0.18, -0.12)
$\phi_5^{(11)}$	(-0.02, 0.04)	(-0.03, 0.04)	(-0.03, 0.02)	$\phi_5^{(21)}$	(-0.01, 0.06)	(-0.01, 0.06)	(-0.02, 0.04)
$\phi_6^{(11)}$	(0.04, 0.09)	(0.04, 0.10)	(0.03, 0.08)	$\phi_6^{(21)}$	(-0.08, -0.03)	(-0.09, -0.02)	(-0.10, -0.04)
$\phi_7^{(11)}$	(0.23, 0.27)	(0.22, 0.28)	(0.22, 0.27)	$\phi_7^{(21)}$	(-0.04, 0.02)	(-0.05, 0.02)	(-0.06, 0.00)
$\phi_8^{(11)}$	(0.09, 0.16)	(0.08, 0.16)	(0.08, 0.14)	$\phi_8^{(21)}$	(0.01, 0.08)	(-0.00, 0.08)	(-0.01, 0.06)
σ_{11}^2	(-0.02, -0.02)	(-0.02, -0.01)	(-0.02, -0.02)	σ_{21}^2	(-0.06, -0.02)	(-0.06, -0.01)	(-0.06, -0.02)
θ_{12}	(-0.49, -0.47)	(-0.49, -0.46)	(-0.49, -0.47)	θ_{22}	(-0.76, -0.68)	(-0.78, -0.67)	(-0.79, -0.70)
$\phi_1^{(12)}$	(-0.33, -0.30)	(-0.33, -0.30)	(-0.33, -0.31)	$\phi_1^{(22)}$	(-0.49, -0.36)	(-0.50, -0.36)	(-0.53, -0.41)
$\phi_2^{(12)}$	(-0.22, -0.19)	(-0.22, -0.19)	(-0.22, -0.19)	$\phi_2^{(22)}$	(-0.28, -0.16)	(-0.30, -0.15)	(-0.33, -0.20)
$\phi_3^{(12)}$	(-0.19, -0.17)	(-0.19, -0.16)	(-0.19, -0.17)	$\phi_3^{(22)}$	(0.07, 0.18)	(0.05, 0.19)	(0.02, 0.14)
$\phi_4^{(12)}$	(-0.29, -0.26)	(-0.29, -0.26)	(-0.29, -0.27)	$\phi_4^{(22)}$	(0.12, 0.25)	(0.11, 0.26)	(0.07, 0.19)
$\phi_5^{(12)}$	(-0.23, -0.20)	(-0.23, -0.19)	(-0.23, -0.20)	$\phi_5^{(22)}$	(-0.24, -0.10)	(-0.25, -0.07)	(-0.30, -0.15)
$\phi_6^{(12)}$	(-0.15, -0.12)	(-0.15, -0.11)	(-0.15, -0.12)	$\phi_6^{(22)}$	(-0.06, 0.06)	(-0.09, 0.07)	(-0.13, 0.00)
$\phi_7^{(12)}$	(0.10, 0.13)	(0.10, 0.13)	(0.10, 0.13)	$\phi_7^{(22)}$	(0.03, 0.13)	(0.00,0.14)	(-0.02, 0.09)
$\phi_8^{(12)}$	(0.23, 0.24)	(0.22, 0.24)	(0.22, 0.24)	$\phi_8^{(22)}$	(0.14, 0.14)	(0.13, 0.14)	(0.14, 0.14)
σ_{12}^2	(0.12, 0.13)	(0.12, 0.13)	(0.12, 0.13)	σ_{22}^2	(0.29, 0.31)	(0.28, 0.30)	(0.28, 0.30)

5. Conclusion

In this article, we have proposed a Bayesian framework to analysis the multiple regime TAR model with multiple structural breaks. This methodology consist identification of break points and estimation of MB-TAR model parameters. Under Bayesian inference, conditional posterior distribution is derived for estimation of the unknown model parameters and computed using Gibbs sampler and M-H algorithm for standard and non standard form distribution, respectively. From the numerical illustration, the proposed Bayesian setup is appropriately determine the break points and estimate the parameters associated with model.

References

1. Albert, J. H., & Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business & Economic Statistics*, *11*(1), 1-15.
2. Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The annals of statistics*, *21*(1), 520-533.

3. Charif, M. A. I. H. A. (2003). Bayesian Inference for threshold moving average models. Metron, 61(1), 119-132.
4. Chen, C. W. (1998). A Bayesian analysis of generalized threshold autoregressive models. *Statistics & probability letters*, *40*(1), 15-22.
5. Chen, C. W., & Lee, J. C. (1995). Bayesian inference of threshold autoregressive models. Journal of Time Series Analysis, 16(5), 483-492.
6. Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366a), 427-431.
7. Gao, Z., & Ling, S. (2018). Statistical inference for structurally changed threshold autoregressive models, Statistica Sinica, forthcoming.
8. Gonzalo, J., & Wolf, M. (2005). Subsampling inference in threshold autoregressive models. *Journal of Econometrics*, *127*(2), 201-224.
9. Li, D., Li, W. K., & Ling, S. (2011). On the least squares estimation of threshold autoregressive moving-average models. *Statistics and Its Interface*, 4, 183-196.
10. Liu, W., Ling, S., & Shao, Q. M. (2011). On non-stationary threshold autoregressive models. *Bernoulli*, *17*(3), 969-986.
11. Meligkotsidou, L., Tzavalis, E., & Vrontos, I. D. (2011). A Bayesian Analysis of Unit Roots and Structural Breaks in the Level, Trend, and Error Variance of Autoregressive Models of Economic Series. *Econometric Reviews*, *30*(2), 208-249.
12. Nelson, C. R., & Plosser, C. R. (1982). Trends and random walks in macroeconmic time series: some evidence and implications. *Journal of monetary economics*, *10*(2), 139-162.
13. Pan, J., Xia, Q., & Liu, J. (2017). Bayesian analysis of multiple thresholds autoregressive model. *Computational Statistics*, *32*(1), 219-237.
14. Tong, H. (1983). Threshold models in non-linear time series analysis. Lecture notes in statistics, No. 21.
15. Wang, J., & Zivot, E. (2000). A Bayesian time series model of multiple structural changes in level, trend, and variance. *Journal of Business & Economic Statistics*, *18*(3), 374-386.
16. Watson, M. W. (1986). Univariate detrending methods with stochastic trends. *Journal of monetary economics*, *18*(1), 49-75.
17. Xia, Q., Liu, J., Pan, J., & Liang, R. (2012). Bayesian analysis of two-regime threshold autoregressive moving average model with exogenous inputs. *Communications in Statistics-Theory and Methods*, *41*(6), 1089-1104.
18. Yau, C. Y., Tang, C. M., & Lee, T. C. (2015). Estimation of multiple-regime threshold autoregressive models with structural breaks. *Journal of the American Statistical Association*, *110*(511), 1175-1186.
19. Yu, P. (2012). Likelihood estimation and inference in threshold regression. *Journal of Econometrics*, *167*(1), 274-294.

Prediction for censored lifetimes from Weibull distribution in Khamis and Higgins Step-Stress Model

Indrani Basak
Penn State Altoona, Altoona, USA

Abstract

We consider the problem of prediction of lifetimes of units from the Weibull distribution which are censored under a simple step-stress testing experiment in this article. We considered Progressive Type-II censoring as the form of censoring. Cumulative Exposure Model (CEM) is the most popular model for analyzing step-stress data. In case of the Weibull distribution, the CEM becomes quite complicated. Due to this reason, Khamis and Higgins (1998) proposed a step-stress model based on the hazard functions and we will use that model in this article. Two kinds of predictors - the maximum likelihood predictors (MLP) and the conditional median predictors (CMP) - are derived. These two prediction methods are numerically illustrated using simulation studies along with generating mean squared prediction error (MSPE) and prediction intervals (PI). We then compare the MLP and the CMP with respect to MSPE and PI.

Keywords

Conditional median predictor; Khamis and Higgins Model; Maximum likelihood predictor; Mean squared prediction error; Prediction Interval.

1. Introduction

Items made in industrial production these days have become very reliable with large mean survival times under normal operating conditions. Complete data are not observed by the experimenter for the items under study in order to reduce cost and experimental time. Those items are usually censored. For Weibull distributed lifetimes, inferential methods are developed by Kundu (2007), Banerjee and Kundu (2008) and Mokhtari, Rad and Yousefzadeh (2011) for various kind of censored data. But, the prediction of actual survival times (which are censored) has not been discussed much. Prediction of unobserved or censored observations is an interesting topic. It is usually difficult to obtain adequate information about lifetime distribution and its associated parameters In conventional industrial life-testing experiment, even with these efficient censoring schemes. In order to obtain information on the parameters of the lifetime distribution more rapidly than under normal operating conditions, accelerated life tests allow the experimenter to expose the experimental

units in industrial experiments to increased levels of stress factors such as load, pressure, temperature and voltage. Step-stress tests are special class of accelerated life tests that allow the experimenter to change the stress levels at pre-fixed times during the life-testing experiment. We consider a simple step-stress testing with only two stress levels and this model has been studied extensively in the literature in this article. Developing statistical prediction theory and procedures for step-stress models under progressive Type II censoring schemes based on Weibull distribution is the purpose of this article. Two frequently used predictors which are called the Maximum Likelihood Predictor (MLP) and the Conditional Median Predictor (CMP) will be used for prediction purpose.

2. Methodology

In a simple step-stress testing setup, n identical units are tested at initial stress level of s1 until a pre-fixed time τ, at which point stress level is changed to s2 and the life-test continues on. Let us denote the hazard functions at stress levels s1 and s2 by h1 and h2, respectively. Cumulative Exposure Model is the most popular model for analyzing step-stress data. The main idea is to assume that the remaining lifetime of the items depends only on the current accumulated stress, regardless of how it has been accumulated. In case of the Weibull distribution, the CEM becomes quite complicated. Khamis and Higgins (1998) proposed the following step-stress model for the Weibull distribution:

$$h(y) = \begin{cases} h_1(y) = \frac{\beta}{\theta_1} y^{\beta-1} \text{ for } 0 < y < \tau \\ h_2(y) = \frac{\beta}{\theta_2} y^{\beta-1} \text{ for } \tau < y < \infty \end{cases} \tag{1}$$

The corresponding survival functions are:

$$\bar{F}(y) = \begin{cases} \bar{F}_1(y) = e^{-\frac{y^\beta}{\theta_1}} & \text{for } 0 < y < \tau \\ \bar{F}_2(y) = e^{-\frac{t^\beta - \tau^\beta}{\theta_2} - \frac{\tau^\beta}{\theta_1}} & \text{for } \tau < y < \infty \end{cases} \tag{2}$$

We will assume that, at the stress level 1, lifetimes have the distribution Weibull (β, θ_1) with the shape and scale parameters β and θ_1 and at the stress level 2, lifetimes have the distribution Weibull (β, θ_2) with the shape and scale parameters β and θ_2. We found that the prediction methods become complicated for the original Weibull distribution and therefore we will be working with the variable $T = \ln Y$ where the hazard function and survival function of the

variable Y is given by (1) and (2) respectively. The probability density function *g(t)* and the cumulative distribution function *G(t)* of *T* are given as

$$g(t) = \begin{cases} g_1(t) = \frac{1}{\sigma} e^{\frac{t-\mu_1}{\sigma}} e^{-e^{\frac{t-\mu_1}{\sigma}}} & \text{for } y < \ln\tau \\ g_2(t) = \frac{1}{\sigma} e^{\frac{t-\mu_2}{\sigma}} e^{-e^{\frac{t-\mu_2}{\sigma}}} e^{\left(e^{\frac{\ln\tau-\mu_2}{\sigma}} - e^{\frac{\ln\tau-\mu_1}{\sigma}}\right)} & \text{for } \ln\tau < t < \infty \end{cases} \quad (3)$$

and

$$G(t) = \begin{cases} G_1(t) = 1 - e^{-e^{\frac{t-\mu_1}{\sigma}}} & \text{for } y < \ln\tau \\ G_2(t) = 1 - e^{-e^{\frac{t-\mu_2}{\sigma}} e^{\left(e^{\frac{\ln\tau-\mu_2}{\sigma}} - e^{\frac{\ln\tau-\mu_1}{\sigma}}\right)}} & \text{for } \ln\tau < t < \infty \end{cases} \quad (4)$$

Where $\mu_1 = \frac{1}{\beta}\ln\theta_1, \mu_2 = \frac{1}{\beta}\ln\theta_2$ and $\sigma = \frac{1}{\beta}$.

Suppose a sample of n experimental units are placed on a simple step-stress life test at an initial stress level of s_1 and the stress level is changed to s_2 at a pre-fixed time τ. Then, the progressive Type-II censoring is implemented in this experimental setting in the following manner. At the stress-level s_1 and at the time of the first failure, R_1 of the $n-1$ surviving units are randomly removed from the experiment. At the time of the second failure, R_2 of the $n-2-R_1$ surviving units are randomly removed from the experiment, and similarly the test continues until time τ. Let N_1 be the random number of units that fail at stress level s_1 and $R^{(1)} = \sum_{i=1}^{n_1} R_i$ be the total number of the censored units at stress level s_1 where n_1 denotes the observed value of N_1. Then, after time τ (at stress level s_2), at the time of the (n_1+1)-th failure, R_{n+1} of the $n-n_1-R^{(1)}-1$ surviving units are randomly removed from the experiment. At the time of the (n_1+2) -th failure, R_{n_1+2} of the $n-n_1-R^{(1)}\ R_{n_1+1}-2$ surviving units are randomly removed from the experiment, and similarly the test continues at the stress level s_2. Let $R^{(2)} = \sum_{i=n_1+1}^{m} R_i$ be the total number of the censored units at stress level s_2 for a fixed value of m, the total number of observations. Then, let $N_2 = m - N_1$ denotes the random number of units that fail at stress level s_2. With $m, R_i\ (i = 1, 2, \cdots, m-1)$ fixed in advance, the test continues until the m-th failure at which time all the remaining $n-m-R^{(1)}-\sum_{i=n_1+1}^{m-1} R_i$ surviving units are removed. Note that $n = n_1 + n_2 + R^{(1)} + R^{(2)}$ where n_2 denotes the observed value of N_2. If $R_1 = \cdots = R_{n_1} = 0, R_{n_1+1} = \cdots = R_m = 0$, then $n = m$ which corresponds to the complete sample situation. If $R_1 = \cdots = R_{n1} = 0, R_{n1+1} = \cdots = R_{m-1} = 0$ and $R_m = n - m$ then it corresponds to the conventional Type-II right censoring scheme. Note

that the life-testing experiment is terminated when the m-th failure occurs. With these notations, we will observe the following progressively censored data:

$$t = (t_1, \cdots, t_{n_1}, t_{n_1}+1, \cdots, t_m)$$

with $t_1 < \cdots < t_{n_1} < \tau \leq t_{n_1+1} < \cdots < t_m$. Here, t is the observed values of the variable $T = T_1, \ldots, T_{N_1}, T_{N_1+1}, \ldots, T_m$ denoting the m Type-II progressively right censored order statistics from a population with pdf $g(t) = g(t; \theta)$ where $(t;\theta) \in D = (R^+) \times \Omega$. Also, $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ and Ω is a 4-dimensional parametric space. We now consider the maximum likelihood prediction of $T_{j:R_i}$, j-th order statistic out of a sample of size $R_i; j = 1,2, \ldots, R_i; i = 1,2, \ldots, m$, having observed $\mathbf{T}$. Here, $t = (t_1, \ldots, t_{n_1}, t_{n_1+1,\ldots,} t_m)$ and $t = t_{j:R_i}$ denote the observed value of $\mathbf{T}$ and the unobserved value of $T_{j:R_i}$ respectively. Using the fact that the conditional density of $T_{j:R_i}$ is the same as the density of the j-th order statistic out of R_i units from the density $g(t)/(1 - G(t_i)), t \geq t_i$ the predictive likelihood function (PLF) of $T_{j:R_i}$ and $\boldsymbol{\theta}$ is given by

$$\begin{aligned} L &= (t, \boldsymbol{\theta}; \boldsymbol{t}) \\ &= g_{T_{j:R_i}} | \, T^{(t|\boldsymbol{t},\boldsymbol{\theta})} . g_{\boldsymbol{T}(\boldsymbol{t},\boldsymbol{\theta})} \\ &= g_{T_{j:R_i}} | \, T_i^{(t|t_i,\boldsymbol{\theta})} . g_{\boldsymbol{T}(\boldsymbol{t},\boldsymbol{\theta})} \end{aligned} \tag{5}$$

in which $g_{T_{j:R_i}} | T^{(t|\boldsymbol{t},\boldsymbol{\theta})}$ denotes the conditional density of $T_j R_i$ given the observed value of $\boldsymbol{T} = \boldsymbol{t}$ and $g_{\boldsymbol{T}(t,\theta)}$ denotes the density of $\mathbf{T}$. L in (5) is then given by

$$L = \begin{cases} c_1 \prod_{l=1}^{n_1} g_1(t_l) \prod_{l=1, l\neq i}^{n_1} [1 - G_1(t_l)]^{R_l} [G_1(t) - G_1(t_i)]^{j-1} \\ g_1(t)[1 - G_1(t)]^{R_i - j} \prod_{l=n_1+1}^{m} g_2(t_l) \prod_{l=n_1+1}^{m} [1 - G_2(t_l)]^{R_l} & \text{if } 1 \leq n_1 \leq m-1 \\ & \& \ i = 1, \cdots, n_1, \\ c_2 \prod_{l=1}^{n_1} g_1(t_l) \prod_{l=1, l\neq i}^{n_1} [1 - G_1(t_l)]^{R_l} [G_1(t) - G_1(t_i)]^{j-1} \\ g_1(t)[1 - G_1(t)]^{R_i - j} & \text{if } n_1 = m \\ & \& \ i = 1, \cdots, n_1, \\ c_3 \prod_{l=1}^{n_1} g_1(t_l) \prod_{l=1}^{n_1} [1 - G_1(t_l)]^{R_l} \prod_{l=n_1+1}^{m} g_2(t_l) \\ \prod_{l=n_1+1, l\neq i}^{m} [1 - G_2(t_l)]^{R_l} [G_2(t) - G_2(t_i)]^{j-1} \\ g_2(t)[1 - G_2(t)]^{R_i - j} & \text{if } 0 \leq n_1 \leq m-1 \\ & \& \ i = n_1 + 1, \cdots, m. \end{cases} \tag{6}$$

Here, c_1, c_2 and c_3 denote constant factors. If $T_{j:R_i}^{L} = u(\boldsymbol{T})$ and $\theta^* = v(\boldsymbol{T})$ are statistics for which

$$L(u(\boldsymbol{T}), v(\boldsymbol{T}); \boldsymbol{T}) = \sup_{T,\theta} L(T, \boldsymbol{\theta}; \boldsymbol{T}),$$

then $u(\boldsymbol{T})$ is said to be the MLP of $T_{j:R_i}$ and $v(\boldsymbol{T})$ the predictive maximum likelihood estimator (PMLE) of $\boldsymbol{\theta}$.

Case 1 : $(1 < N_1 < m-1$ **and** $i = 1, \cdot\ \cdot\ \cdot, n_1)$ The logarithm of the predictive likelihood function (log PLF) of $T_{j:R_i} = t$, corresponding to the PLF given by the first equation of (6), is given by

$$\log L = -(m+1)\log\sigma + \sum_{l=1}^{n_1}\left(\frac{t_1-\mu_1}{\sigma}\right) - \frac{M_1}{\sigma} + (j-1)\log\left(1-e^{-\frac{w}{\sigma}}\right) + \left(\frac{t-\mu_1}{\sigma}\right) + \sum_{l=n_1+1}^{m}\left(\frac{t_t-\mu_2}{\sigma}\right) - \frac{M_2}{\sigma} \tag{7}$$

in which $M_1 = \sum_{l=1}^{n_1}(R_l+1)z_{1l} + (R_i-j+1)w$ and $M_2 = \sum_{l=n_1+1}^{m}(R_l+1)z_{2l}$ with $w = w(t, t_i, \mu_1, \sigma) = \sigma\left(e^{\frac{t-\mu_1}{\sigma}} - e^{\frac{t_i-\mu_1}{\sigma}}\right)$ and $\mathcal{Z}_{1l} = \sigma e^{\frac{t_1-\mu_1}{\sigma}}; i = 1, \ldots, n_1$ and $\mathcal{Z}_{2l} = \sigma\left(e^{\frac{t_1-\mu_2}{\sigma}} - e^{\frac{\ln\tau-\mu_2}{\sigma}} + e^{\frac{\ln\tau-\mu_1}{\sigma}}\right); i = n_i+1, \ldots, m.$

The predictive likelihood equations (PLEs) are obtained by differentiating log L in (7) with respect to t, μ_1, μ_2 and σ and these are as follows:

$$\left.\begin{aligned}
\frac{\partial \log L}{\partial t} &= -(R_i-j) + (j-1)\frac{e^{-\frac{w}{\sigma}}}{1-e^{-\frac{w}{\sigma}}} = 0,\\
\frac{\partial \log L}{\partial \mu_1} &= \frac{1}{\sigma}\left[\sum_{l=1}^{n_1}(R_l+1)e^{\frac{t_l-\mu_1}{\sigma}} + \frac{w}{\sigma}\left\{(R_i-j+1) - (j-1)\frac{e^{-\frac{w}{\sigma}}}{1-e^{-\frac{w}{\sigma}}}\right\} + e^{\frac{\ln\tau-\mu_1}{\sigma}}\sum_{l=n_1+1}^{m}(R_l+1) - (n_1+1)\right] = 0,\\
\frac{\partial \log L}{\partial \mu_2} &= \frac{1}{\sigma}\left[\sum_{l=n_1+1}^{m}(R_l+1)e^{\frac{t_l-\mu_2}{\sigma}} - e^{\frac{\ln\tau-\mu_2}{\sigma}}\sum_{l=n_1+1}^{m}(R_l+1) - (m-n_1)\right] = 0,\\
\frac{\partial \log L}{\partial \sigma} &= \frac{1}{\sigma^2}\left[\sum_{l=1}^{n_1}(R_l+1)(t_l-\mu_1)e^{\frac{t_l-\mu_1}{\sigma}} + Q_1\left\{(R_i-j+1) - (j-1)\frac{e^{-\frac{w}{\sigma}}}{1-e^{-\frac{w}{\sigma}}}\right\}\right.\\
&\quad + \sum_{l=n_1+1}^{m}(R_l+1)\left\{(t_l-\mu_2)e^{\frac{t_l-\mu_2}{\sigma}} - (\ln\tau-\mu_2)e^{\frac{\ln\tau-\mu_2}{\sigma}} + (\ln\tau-\mu_1)e^{\frac{\ln\tau-\mu_1}{\sigma}}\right\}\\
&\quad \left.-\sum_{l=1}^{m}t_l - t + (n_1+1)\mu_1 + (m-n_1)\mu_2 - (m+1)\sigma\right] = 0
\end{aligned}\right\} \tag{8}$$

where $Q_1 = (t-\mu_1)e^{\frac{t-\mu_1}{\sigma}} - (t_i-\mu_1)e^{\frac{t_i-\mu_1}{\sigma_1}}$. The MLP $T_{j:R_i}^{L}$ of $T_{j:R_i}$ is obtained by solving the first equation of (8) and is given by

$$T_{j:R_i}^{L} = \mu_1^* + \sigma^* \log\left[e^{\frac{t_i-\mu_1^*}{\sigma^*}} + D\right] \tag{9}$$

in which $D = D(j, R_i) = \log((R_i - 1)/(R_i - j)$ and μ_1^* and σ^* are the PMLEs of μ_1 and σ respectively. After substituting t by $T_{j:R_i}^{L}$ as given in (9), the second equation in (8) becomes

$$\sum_{l=1}^{n_1}(R_1+1)e^{\frac{t_1-\mu_1}{\sigma_1}} + e^{\frac{\ln\tau-\mu_1}{\sigma}}\sum_{l=n_1+1}^{m}(R_1+1) + D = n_1 + 1 \tag{10}$$

The third equation in (8) simplifies to

$$\sum_{l=n_1+1}^{m}(R_1+1)e^{\frac{t_1-\mu_2}{\sigma}} - e^{\frac{\ln\tau-\mu 2}{\sigma}}\sum_{l=n_1+1}^{m}(R_1+1) = m - n_1. \tag{11}$$

σ^* is then obtained by using the third equation of (8), (10) and (11) after substituting t by $T_{j:R_i}^{L}$ as given in (9) as

$$\sigma^* = \frac{\sum_{l=1}^{n_1}(R_l+1)(t_l-\ln\tau)e^{\frac{t_l-\mu_1}{\sigma}} + (t-t_i)e^{\frac{t_i-\mu_1}{\sigma}} + \sum_{l=n_1+1}^{m}(R_l+1)(t_l-\ln\tau)e^{\frac{t_l-\mu_2}{\sigma}} + (t-\ln\tau)D + (m+1)\ln\tau - \sum_{l=1}^{n_1}t_l - t}{m+1}. \tag{12}$$

It can be observed that right hand side of (12) contains σ^* which creates some computational difficulty. Following Thomas and Wilson (1972), we propose the linearized PMLE $\tilde{\sigma}$ of σ. It is given by

$$\tilde{\sigma} = \frac{\sum_{l=1}^{n_1}(R_l+1)(t_l-\ln\tau)\kappa_{1l} + (t-t_i)\kappa_{1i} + \sum_{l=n_1+1}^{m}(R_l+1)(t_l-\ln\tau)\kappa_{2l} + (t-\ln\tau)D + (m+1)\ln\tau - \sum_{l=1}^{n_1}t_l - t}{m+1}. \tag{13}$$

where for $l = 1, 2, \cdots, n_1, \kappa_{1l} = E\left[e^{\frac{t_l-\mu_1}{\sigma}}\right] = \sum_{i=1}^{l} c_i^{-1}$ with $c_i = \sum_{j=i}^{n_1}(R_j+1)$ and for $l = n_1+1, \cdots, m, \kappa_{2l} = E\left[e^{\frac{t_l-\mu_2}{\sigma}}\right] = \sum_{i=n_1+1}^{l} d_i^{-1}$ with $d_i = \sum_{j=i}^{m}(R_j+1)$. Finally, the PMLE $\tilde{\mu}_1$ corresponding to the linearized PMLE $\tilde{\sigma}$ in (13) is obtained by solving (10). The modified MLP (MMLP) is then given by

$$T_{j:R_i}^{L} = \begin{cases} Y_i & \text{if} \quad j = 1 \\ \tilde{\mu}_1 + \tilde{\sigma}\log\left[e^{\frac{t_i-\tilde{\mu}_1}{\tilde{\sigma}}} + D(j, R_i)\right] & \text{if} \quad 1 < j \le R_i. \end{cases} \tag{14}$$

Case 2 : ($n_1 = m$ **and** $i = 1, \ldots, n_1$) The MMLP $T_{j:R_i}^{L}$ of $T_{j:R_i}$ is given by (14) in which PMLE $\mu_1^*$is obtained by solving

$$\sum_{l=1}^{n_1}(R_1+1)e^{\frac{t_1-\mu_1}{\sigma_1}}+D=n_1+1. \tag{15}$$

The linearized PMLE $\tilde{\sigma}$ is given by

$$\tilde{\sigma}=\frac{\sum_{l=1}^{n_1}(R_l+1)t_l\kappa_{1l}+(t-t_i)\kappa_{1i}+tD-\sum_{l=1}^{n_1}t_l-t}{m+1}. \tag{16}$$

in which k_{1l} is as given in Case 1. The PMLE $\tilde{\mu}_1$ corresponding to the linearized PMLE $\tilde{\sigma}$ in (16) is obtained by solving (15).

Case 3 : $(0 \le n_1 \le m-1$ **and** $i=n_i+1,\dots,m)$ Following similar derivations, the modified MLP (MMLP) in this case is given by

$$T_{j:R_i}^{L}=\begin{cases}Y_i & \text{if } j=1\\ \tilde{\mu}_2+\tilde{\sigma}\log\left[e^{\frac{t_i-\tilde{\mu}_2}{\tilde{\sigma}}}+D(j,R_{i)}\right] & \text{if } 1<j\le R_i\end{cases} \tag{17}$$

3.2. Conditional Median Predictor

The median of the conditional distribution of $T_{j:R_i}$, given T_i, can be used as a predictor of $T_{j:R_i}$. This predictor is called the Conditional Median Predictor (CMP) of $T_{j:R_i}$ and will be denoted by $T_{j:R_i}^{C}$. The CMP $T_{j:R_i}^{C}$ of $T_{j:R_i}$ is such that $\int_{t_i}^{T_{j:R_i}^{C}} g(t|t_i)dt=\frac{1}{2}$ in which $g(t|t_i)$ is the conditional density of $T_{j:R_i}$, given $T_i=t_i$. Since $\left[e^{\frac{T_{j:R_i}-\mu}{\sigma}}-e^{\frac{T_i-\mu}{\sigma}}|T_i=t_i\right]$ where μ and σ are the location and scale parameters respectively (in general), has the same distribution as the random variable $Z_{j:R_i}$, the CMP $T_{j:R_i}^{C}$ of $T_{j:R_i}$ is given by

$$T_{j:R_i}^{C}=\mu_1+\sigma\log\left[e^{\frac{t_i-\mu_1}{\sigma}}+M_{j,R_i}\right]$$

Where $Z_{i:R_i}$ is the j-th order statistics out of R_i units from Exp(1) and $\text{Med}\left[Z_{j:R_i}\right]=M_{j:R_i}$.

Case 1 : $1 \le n_1 \le m-1$ **and** $i=1,\dots,n_1$ The CMP $T_{j:R_i}^{C}$ of $T_{j:R_i}$ is given by

$$T_{j:R_i}^{C}=\hat{\mu}_1+\hat{\sigma}\log\left[e^{\frac{t_i-\hat{\mu}_1}{\hat{\sigma}}}+M_{j,R_i}\right], \tag{18}$$

in which $\hat{\mu}_1$ and $\hat{\sigma}$ can be used as the linearized uniformly minimum variance unbiased estimator (UMVUE) of μ_1 and σ to reduce the variability in the

resulting prediction. $\hat{\sigma}$ is given by (13) with the denominator modified to m. Then, the linearized UMVUE $\hat{\mu}_1$ corresponding to $\hat{\sigma}$ is obtained by solving (10) when σ is substituted by this $\hat{\sigma}$.

Case 2 : ($n_1 = m$ **and** $i = 1, \dots, n_{1)}$ The CMP $T_{j:R_i}^C$ of $T_{j:R_i}$ is given by (18) in which the linearized UMVUE $\hat{\sigma}$ given by (16) with the denominator modified to m and the linearized UMVUE $\hat{\mu}_1$ is obtained by solving (15) when σ in there is substituted by this $\hat{\sigma}$.

Case 3 : ($0 \leq n_1 \leq m-1$ **and** $i = n_1 + 1, \dots, m$) The CMP $T_{j:R_i}^C$ of $T_{j:R_i}$ is given by

$$T_{j:R_i}^C = \hat{\mu}_2 + \hat{\sigma} \log\left[e^{\frac{t_i - \hat{\mu}_2}{\hat{\sigma}}} + M_{j,R_i}\right] \tag{19}$$

3. Results

Progressive Type-II censored data under the step-stress setting are generated and the values of MLP and CMP of $T_j R_i$ are computed under Khamis and Higgins Model. A numerical study is carried out to compare the performances of these MLP and CMP in terms of their Mean Square Prediction Errors (MSPEs) and Prediction Intervals (PIs). Derivations of the MSPEs and PIs of the MLP and CMP are complicated and so we used simulations to get those. Using simulation studies, standard errors of these $T_j R_i$ were generated and the PIs were constructed for each of the predictors MLP and CMP.

4. Discussion and Conclusion

In this article, we have derived the MLP and CMP for the survival times of units from the Weibull distribution which are progressively Type II censored under the Khamis and Higgins model for the simple step-stress data. Simulation studies are used to illustrate and compare the methods developed in this article. Simulation studies show that the predicted values for CMP are generally closer to their actual values than the corresponding predicted values for MLP. This is particularly true for larger sample size, larger number of uncensored observations and for delayed censoring scheme as long as the number of predicted observations are not very small.

References

1. Banerjee, A. and Kundu, D. (2008). Inference Based on Type-II Hybrid Censored Data from a Weibull Distribution, *IEEE Transactions on Reliability* 57, pp. 369-378.
2. Khamis, I. H. and Higgins, S. H. (1998). A new model for step-stress testing, *IEEE Transactions on Reliability* 47, pp. 131-134.

3. Kundu, D. (2007). On hybrid censored Weibull distribution, *Journal of Statistical Planning 4 Inference* 137, pp. 2127-2142.
4. Mokhtari, E.B., Rad, A.H. and Yousefzadeh, F. (2011). Inference for Weibull distribution based on progressively Type-II hybrid censored data, *Journal of Statistical Planning 4 Inference* 141, pp. 2824-2838.
5. Thomas, D.R. and Wilson, W.M. (1972). Linear order statistic estimation for the two-parameter Weibull and Extreme Value distributions from Type-II progressively censored samples, *Technometrics* 14, pp. 679-691.

Statistics education in Indian Agricultural Universities and Agro- Technical Institutes

D. S. Hooda[1]; B. K. Hooda[2]
[1]Honorary Professor in Mathematics, G J University of Science and Technology
[2]Department of Mathematics and Statistics, CCS Haryana Agricultural University

Abstract
Statisticians working in State Agricultural Universities (SAU) and Agro-technical Institutes of Indian Council of Agricultural Research (ICAR) play a very important role in the planning of agricultural research programs in India by doing layout of agricultural experiments, collection of data, analysis and drawing valid inferences from these data. In the present paper we discuss the present status of agricultural education and research and teaching of agricultural statistics at all levels in state agricultural universities and agro-technical institutes of the country. The role of computer technology in research and management of agricultural statistics has also been explained. Curricula of various programs in agricultural statistics and evaluation systems are presented. Some ways and means are suggested for further improvement in statistics education at national and international levels.

Keywords
Agro-technical; agricultural statistics; curricula; statistics education and evaluation system

1. Introduction

The innovations in agricultural sciences and information technology have made it possible for us to provide our country with enough quantities of food, fodder, fuel etc. In order to meet the ever-increasing food grains requirement of India its food grains production is to be increased and vis-à-vis the agricultural education in the country need to be strengthened. The goal for increased production can only be achieved through the applications of the latest techniques supported by the trained manpower. The State Agricultural Universities (SAU's) and the Institutes of Indian Council of Agricultural Research (ICAR) have the responsibility of training personnel to meet the research and teaching requirements agricultural research and education the country. Advancement in agricultural education in India started with the establishment of the first agricultural university at Pant Nagar in 1960 followed by Punjab Agricultural University(PAU) at Ludhiana in 1962. Agricultural education got further boost up with the establishment of a division of agricultural education in 1966 in the Indian Council of Agricultural Research.

At present there are 28 State Agricultural Universities and a Central Agricultural University for the north east region.

In addition to SAU's we have four institutes under ICAR system viz., Indian Agricultrual Research Institutes (IARI), New Delhi, Indian Veterinary Research Institutes (IVRI), Izatnagar, National Dairy Research Institutes (NDRI), Karnal and Central Institute of Fishery Education (CIFE), Mumbai which enjoy the status of deemed universities. All these universities and institutes have been established mainly on the pattern of the Land Grant Institutions of USA. The state agricultural universities and the ICAR institutes have the responsibility of under graduate and the post graduate education in agricultural and allied disciplines. The ICAR provides technical and financial support to the State Agricultural Universities.

In order to increase the potential of agricultural research, it is essential that agricultural scientists be adequately exposed to the statistical techniques. The Statisticians working in SAU's and ICAR institutions in India provide statistical and statistically related expertise for a diverse clientele. In addition, many SAU's have the teaching program of Agricultural Statistics at under-graduate (UG), post-graduate (PG) and doctoral levels. The course curriculum, system of education, admission, system of examination and grading all vary from institution to institution. Therefore, there is an urgent need to evaluate the programs of agricultural Statistics at various universities/institutes. Keeping above things in mind, this presentation is planned to highlight Teaching Agricultural Statistics in SAU's and ICAR institutes. Academic programs and their curricula have been discussed. Some ways and means for Improvement in Agricultural Statistics education in India have been suggested.

2. Teaching of Agricultural Statistics in India

During last few decades statistics has penetrated into almost all sciences like agriculture, biology, business, social, engineering, medical, etc. Its wide and varied applications have lead to the growth of many new branches, such as Industrial Statistics, Biometrics, Biostatistics, Agricultural Statistics and the most recently developed Statistical Bioinformatics. These branches have emerged as distinct entities or subjects with a bulk of statistical techniques specific to their application areas. Agricultural Statistics comprises the area of statistical science that deals directly with the problems of field experimentation and interpretation of results in agricultural sciences.

Agricultural Statistics is the most important discipline regarding the training of the research scientists, to help them in planning of their experiments and in the analysis of data and drawing inferences thereof. In addition to statistical research, which is applicable to agricultural problems and consulting activities, agricultural statisticians also participate in collaborative research projects with scientists from other departments.

Statisticians in Agricultural universities and institutions not only provide a strong technical support to other departments and disciplines for agriculture related programs but also conduct their own research in Statistics. In almost all SAU's, Statistics is taught as supporting subject and is compulsory for the UG and PG programs irrespective of their discipline. Students also take statistics as minor from departments such as Agricultural Economics.

In Agricultural Universities, the subject of Statistics is also taught as an independent subject leading to the formal degrees of M.Sc. and Ph.D. in Statistics. We observe that teaching Statistics to agricultural students is a tough job and needs special attention. This is mainly because of the fact that agricultural students are not well trained in mathematical concepts and face difficulty in grasping the mathematical notions and the subject matter itself. Also many students seem to have the feelings that Statistics courses are dry and unattractive and consider Statistics it irrelevant to their own disciplines.

A teaching approach includes teaching tools or methods to be used in teaching such as tutorials, illustrations, case studies, Audio-visual like projectors or LCD's etc. Use of audio visual aids would make the teaching more enjoyable. Audio visual aids like an overhead projector, slide projector and LCD's can help in better presentation and condensing information from bulk and thus provide more information in short time. But majority of SAU's and ICAR institutions lacks in such facilities.

3. Course Curriculum for Statistics Teaching Programs

All SAU's and ICAR institutes having status of deemed universities have the dual responsibilities of conducting research in agricultural Statistics and also of running regular courses leading to the award of M.Sc. and Ph.D. degrees in Statistics. In addition to the PG.-programme in Statistics, the Statistics Departments in SAU,s are also involved in teaching of informal courses designed to cater the needs of agricultural scientists. Despite similarity in education system and objectives of SAU's, we find considerable variation in their course curriculum for M.Sc and Ph.D programs in Statistics. The course curriculum is an important aspect for the success of any program of study and, in view of the advancement in the subject, there is always a continuous need for updating it. Past experience indicates that the job opportunities in agricultural universities are declining day by day. The outgoing students from SAU's have to compete in examinations at the national level conducted by agencies such as ASRB, UPSC, UGC etc. The syllabi for UPSC and UGC examinations are generally formulated keeping in view the syllabi of traditional universities and not the syllabus of M.Sc. Level programs taught in Agricultural Statistics as these do not cover a significant portion of the syllabus framed by national bodies like UGC and UPSC.

Teaching of Statistics in traditional universities have maximum emphasis on course work whereas the students in SAU's have to devote more than half of their study time on minor, supporting and thesis work. Because of the diversion of a sizable portion of the total load towards minor and thesis work, M.Sc. students in SAU's and deemed universities of ICAR are not getting enough time for their major subject. Due to the existence of such disparities, some traditional universities in our country consider the SAU's product as that of a substandard and ignore them for appointments. Majority of the universities in India whether under ICAR or UGC systems have introduced Statistics directly from the Master's level. Consequently, students are not getting good exposure of the subject within a short span of two years. The training in Agricultural Statistics in India is to some extent incomplete and obsolete. Scrapping the minor and thesis system at master's level or by restricting entry in M.Sc. to only those students who have studied Statistics as a subject at B.Sc. level can overcome this dilemma.

4. Eligibility for Admission in PG Programs

In general the eligibility criterion for M.Sc. Statistics in SAU's is graduation with mathematics or/ Statistics at under graduation level as one of the subject. Recently High Level committees of ICAR recommended that only those courses be run in SAU/ICAR institutes in which agricultural graduates are eligible.

Consequent upon the recommendation of this committee the general eligibility qualifications for M.Sc. Statistics programs have been modified to a Bachelor's degree in Agriculture/Horticulture/ Agro forestry/ Science. Agricultural graduate though give first preference to agricultural subjects, yet seek admission in Statistics if they do not get in any other discipline. But their constant effort is to shift to other disciplines as and when they get an opportunity. Consequently, some seats in Statistics program always remain vacant.

The qualification for admission to Ph.D. programme requires Master's degree in Agricultural Statistics/ Statistics/ Mathematical Statistics with at least 60% marks or overall grade point average (OGPA) of 7.00/10.00 or 3.00/4.00. Like other subjects the introduction of Statistics as a full subject at graduation level along with suitable combinations with subjects like Mathematics, Computer Sciences/applications, economics etc. can be helpful in getting talented students for M.Sc. Statistics programs.

5. System of Examination and Evaluation

Examination system plays an important role in imparting education. In majority of SAU's each academic year was initially divided into four parts namely trimester I, II and III with summer trimester. The evaluation of student's

performance was done under the grading system. An undergraduate student was required to secure at least 50% marks to pass in a course where a D grade was awarded to him. A student was awarded C, B and A grades on securing 60-69, 70-79 and more than 80% marks, respectively. At the postgraduate level, the pass percentage was 60 for which a student was awarded a C grade. The conditions for awarding B and A grade remained the same as above.

On completion of degree requirements, a student was awarded Overall Grade Point Average (OGPA) out of a maximum of 4.00. The universities/ institutes have these days switched over to the semester system and each academic year is now comprised of two semesters namely semester I and II. With the requirement of pass percentage remaining unchanged, the grading is done on a 10-point scale and a student is awarded Overall Credit Point Average (OCPA) out of a maximum of 10.00.

For M.Sc. and Ph.D. programs all the students have to take examinations given by the instructors concerned, in various courses registered by them in that particular semester in the major, minor supporting fields. For doctoral students, after having successfully completed the major portion of the course work (at least 75% as per the students' proposed plan of work), a written comprehensive examination is held in both major and minor fields, followed by a pre-comprehensive oral examination. Thereafter, a qualifying viva-voce examination is held to test the students' general mastery of the subject.

6. Training to Statisticians in Agricultural Universities

Statisticians working in state agricultural universities and deemed universities of ICAR, often, assist agricultural scientists and undertake methodological investigations utilizing the data generated under different agricultural experimentation programs. They have to work with people from other professional backgrounds to solve practical problems. They use modern computing methods to process and interpret data. In agricultural research the emphasis is on the data to be understood and the problem to be solved rather than on computing methods for their own sake.

A statistician who works with the agricultural scientists must have an understanding of biological phenomena and should develop the ability to coordinate with other people. He should be trained in analysis of statistics packages such as SPSS, SAS etc. But this is not found in general. Due to ban and limited appointments most of the agricultural statisticians are in old age and are not aware of the recent packages. Moreover, only a few institutions have advanced packages like Mat Lab., SPSS, SAS etc.

7. Suggestions for Improvement

The Education Division of ICAR is constantly making efforts to improve the quality of human resource through funding of training programmes by way of

setting up the Centre of Advanced Studies in various disciplines of Agricultural Sciences including Statistics. A qualitative positive effect of these efforts can now be visualized on the teaching and research in SAUs and ICAR institutes.

Efforts are needed to update these programs and bring uniformity in their course curriculum. Also due to availability of high speed computers and advanced analysis software packages, there is an imminent need to contemplate over the training provided to statisticians and redefine their role in agricultural universities.

Following are a few suggestions for regarding improvements in Agricultural Statistics Education in India:

- Teacher's assigned teaching must have undergone some training in agricultural and related sciences so that they can do applications oriented teaching.
- The system needs to be monitored properly and regularly to improve the quality of teaching.
- The course curriculum of all SAU's and ICAR institutions should be as uniform as possible. It must be at par with the academic programmes running at the international level. The Agricultural Statistics course curriculum at all the levels of B.Sc., M.Sc. and Ph.D. should be revised and updated from time to time.
- Relevant teaching material be prepared and handed over to students in the beginning of the semester. Textbooks may be written and followed keeping in view the objective of the courses and background of the students. Teaching be made experiments oriented and enjoyable supported by live examples through Audio -Visual aids, LCD's etc.
- The admission criteria for M.Sc. and Ph.D. programmes require a fresh look. It has been observed that the students who have not studied mathematics at +2 level face problems in grasping the subject. It is desirable that the students who are admitted to M.Sc./Ph.D. programme in Agricultural Statistics should have studied Mathematics/ Statistics at the graduation level.
- Statistics should be introduced as an elective subject like other subjects at undergraduate level, and preference in admissions be given to only those who have Statistics as a full subject at B.Sc. level.
- There should be uniformity in the grading system and the minimum course curriculum of statistics at B.Sc., M.Sc. and Ph.D. levels.
- The system of education, to the extent possible, should be the same in all the Universities/Institutes and the academic sessions should also start during the same month as far as possible.
- At least six months' professional training be provided to all the newly appointed statisticians in SAU's. During the training periods they may

be asked to handle practical problems, case studies or small research projects of applied nature. Training should also include the preparation of layout for field experiments and their actual implementation in the field conditions.

The use of computers needs to be integrated with the class room teaching. The theory needs to be supported and explained by live practical examples through computers. Students have to be trained in the use of computers and application of various standard statistical software packages for analysis of data using different statistical techniques.

References

1. Proceedings of the National Workshop on Teaching/Training in Agricultural Statistics and Computer Applications held during Nov, 19-20. 1985 at IASRI New Delhi.
2. Proceedings of the Ninth National Conference of Agricultural Research Statisticians held during July 19-21, 1989, at Tamil Nadu Agricultural University, Coimbatore.
3. Proceedings of the Thirteenth National Conference of Agricultural Research Statisticians held during November 6-8, 2001, at Ludhiana.

Clustering Chinese cities' economic growth paths with dynamic time warping

Song Xue, Hong Liu
Zhongnan University of Economics and Law, Wuhan

Abstract

In this research, we use the time-series cluster analysis to study Chinese cities' economic growth patterns. In particular, dynamic time warping algorithm is used to measure the time-series distance between two growth paths. 35 Chinese cities that are economically most important are included in the research. The cluster analysis categorizes the 35 cities into five groups, each of which exhibits distinct economic growth patterns. The five groups are: a. service centers, b. deindustrializing cities, c. balanced-industrializing cities, d. traditional industrial centers, and e. emerging industrial centers. This research shows the potential of applying unsupervised machine learning techniques in development economics, and can be extended in many ways.

Keywords

Dynamic time warping; time series clustering; urban growth; economic development

1. Introduction

China's economic growth in recent decades has been accompanied by unprecedented scale of urbanization (Démurger, 2001). In the last three decades, over half a billion Chinese rural residents have moved to cities. The urban population in China has increased from 19.4% in 1980 to 57.9% in 2017. Given a base population growth from 1 billion to 1.38 billion during the same period, this urbanization surge means over half a billion rural residents have moved to cities in China, more than the population of United States and Japan combined. After year 2006, 60% of China's urbanization occurs in large cities that have more than 4 million residents.

While China's economic boom and the accompanying urbanization have been the themes of voluminous research (Chan and Wan, 2017), the heterogeneity within the economic development paths of Chinese cities is often overlooked in extant literature. Despite their similarities in economic achievements, cities in China are diverse in natural, social, and political factors that would influence their growth paths (Wu, 2016). Some of the cities are coastal harbors that have been the center of trade and commerce for centuries, while some others are hundreds of miles inland, and have become industrialized at the result of central planning (Alder 2016). In terms of the

economic growth path, the differences among these cities are arguably no smaller than the differences between Dubai, UAE and San Jose, CA. Examining the variations in China's cities' economic development paths would shed additional light on how to induce and facilitate urbanization process in developing countries. Using the weights of secondary and tertiary industries as input variables, the time-series clustering has grouped the 35 major Chinese cities into five categories. Cities in each of the five categories have shown interesting similarities in economic growth paths, in spite of some seemingly significant disparities. On the other hand, cities in different categories have distinct growth patterns. This research shows the potential of applying unsupervised machine learning techniques in the field of development economics.

2. Methodology

Clustering is a family of machine learning algorithms that seek to categorize unlabeled data objects into a number of groups, in such a way that objects in the same group are similar and objects in different groups are distinct (Jain, Murty and Flynn, 1999). In contrast to classification algorithms, which assign data objects to predefined groups and hence necessitate labeled training data, clustering algorithms do not require ex ante knowledge on the groups.

Selecting distance measure to evaluate the similarity of data objects is critical to a clustering algorithm and its results. The Euclidean distance is commonly used in clustering as the similarity measure. However, it is not suitable for clustering time-series data. Extant literature have proposed a number of time-series similarity measures, with each of them being appropriate for different applications (Aghabozorgi, Shirkhorshidi and Wah, 2015).

In this research, we use dynamic time warping (DTW, [Jeong, Jeong, and Omitaomu, 2011]) distance for clustering the time-series economic data of 35 Chinese cities. DTW measures the distance between two data series based on their shapes. The method was first applied in speech processing problems by Berndt and Clifford (1994), and soon become popular as a time series distance measure (Izakian, Pedrycz, and Jamal, 2015).

DTW computes the distance between two time series T1 and T2 of length m and n using dynamic programming as follows:

1. $D_{DTW}(T_1, T_2) = 0, if\ m = n = 0$
2. $D_{DTW}(T_1, T_2) = \infty, if\ m = n \neq 0$
3. $D_{DTW}(T_1, T_2) = dist(T11, T21) + MinFactor, otherwise$

Where $minFactor$ is computed as

$$minFactor = min\begin{cases} D_{DTW}(Rest(T_1), Rest(T_2)) \\ D_{DTW}(Rest(T_1), T_2) \\ D_{DTW}(T_1, Rest(T_2)) \end{cases}$$

The research uses the annual macroeconomic data of 35 Chinese cities from year 2007 to year 2016, which are publicly available on the website of The National Bureau of Statistics. The 35 cities are in top 3 tiers in China's administrative level classifications:

1. Direct-controlled municipalities (4): Beijing, Tianjin, Shanghai, Chongqing;
2. Provincial capitals (26): Taiyuan, Guangzhou, Haikou, Nanjing, Hangzhou, Hefei, Fuzhou, Nanchang, Jinan, Zhengzhou, Wuhan, Changsha, Shenyang, Changchun, Harbin, Hohhot, Shijiazhuang, Yinchuan, Xi'an, Lanzhou, Xining, Urumqi, Chengdu, Guiyang, Kunming, Lhasa, Nanning;
3. Cities specifically listed in economic plans[1] (5): Shenzhen, Ningbo, Ningbo, Qingdao, Dalian.

The metrics we selected as input features are those closely reflect a city's economic growth patterns. In particular, the weights of secondary and tertiary sectors in the economy[2] are considered. In the cluster analysis, the values are standardized to z-scores by year.

3. Results

The DTW time series clustering with the standardized industry composition data results in five city groups as shown in Table 1.

Table 1: Five city groups are identified by cluster analysis.

Group	Number of Cities	Cities
A	7	Beijing, Guangzhou, Haikou, Harbin, Hohhot, Shanghai, Urumqi
B	9	Guiyang, Hangzhou, Jinan, Lanzhou, Nanjing, Shenzhen, Taiyuan, Xi'an, Xiamen
C	6	Chengdu, Fuzhou, Kunming, Nanning, Qingdao, Wuhan
D	4	Changchun, Chongqing, Dalian, Shijiazhuang
E	9	Changsha, Hefei, Nanchang, Ningbo, Shenyang, Tianjin, Xining, Yinchuan, Zhengzhou

[1] Chinese: 计划单列市.

[2] The first sector was excluded, for the obvious reason that its values are perfectly linear combination of the other two

Group A cities can be called service centers. They are the only group of cities that have significantly higher proportions of the tertiary industries than that of the secondary industries, and their industry composition remains stable over time. This group includes the top three Chinese cities in terms of economic importance: Beijing, Shanghai, and Guangzhou. Also included are four provincial capitals of the border provinces: Haikou, Harbin, Hohhot, and Urumqi. This result suggests that these cities play the role of service centers in each of their own regions that are either geographically or culturally independent from the national economy.

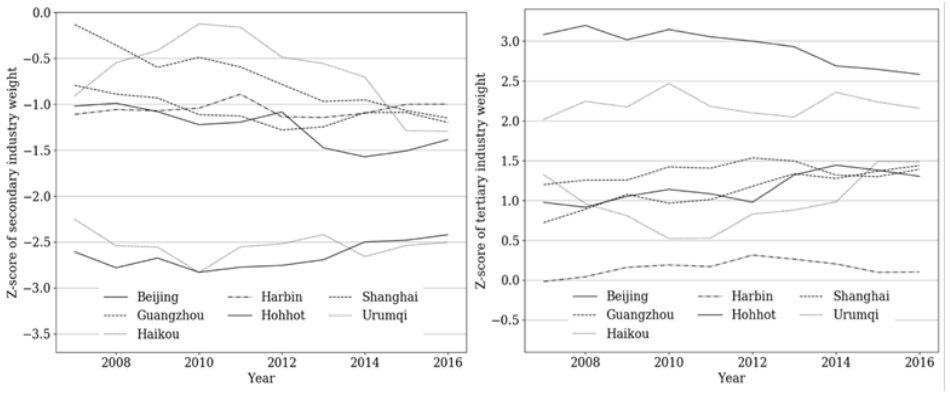


Figure 1: Normalized weights of secondary and tertiary industries for group A cites

Group B cities can be called deindustrializing cities. These cities used to heavily rely on the secondary industries as the economic engine, but their service industries are growing rapidly in recent years. Coastal cities that have been the export-driven manufacturing centers are in this group, including Shenzhen, Xiamen, and Hangzhou. It is interesting that quite a few central and western China cities such as Xi'an and Guiyang are in this group too.

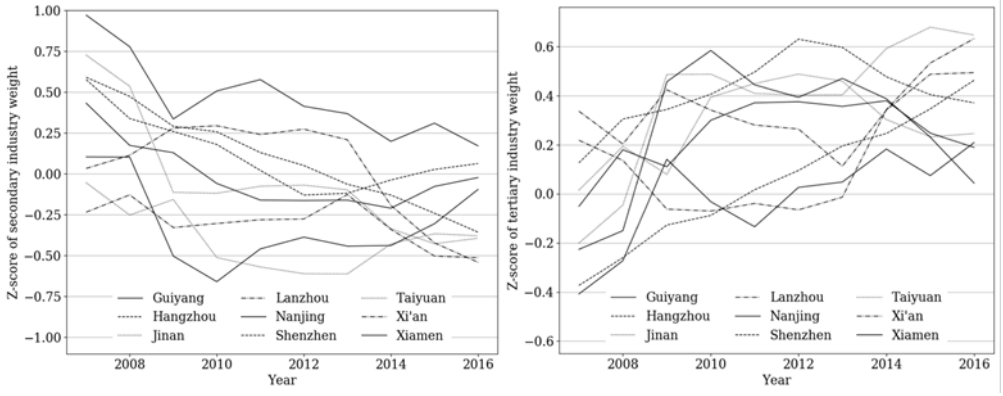


Figure 2: Normalized weights of secondary and tertiary industries for group B cites.

Group C include cities that seem to have balanced economies between secondary and tertiary industries. They also seem to have experiencing slightly faster growth in secondary industries than the service sector. These cities are mostly in central China that have been receiving capital and labor influx from the coast, where the manufacturer costs have been rising in the last decade. This group can be called balanced-industrializing cities.

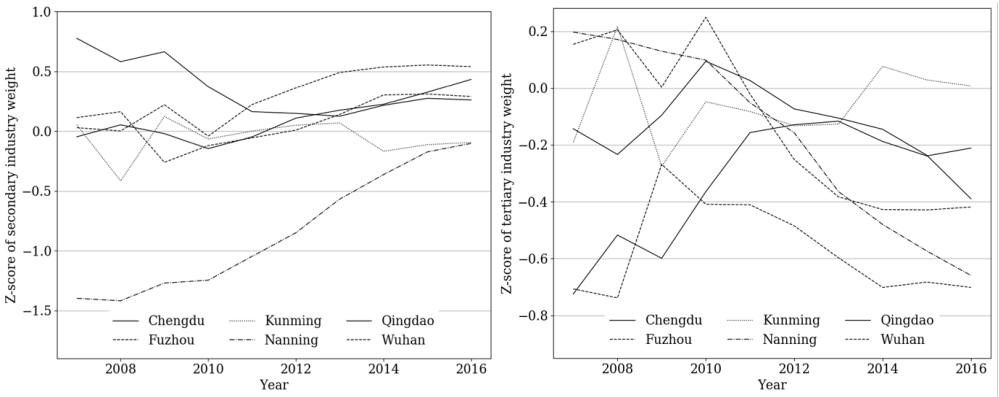


Figure 3: Normalized weights of secondary and tertiary industries for group C cites.

Group D and E are cities with heavy reliance on secondary industries than on the tertiary industries. The major difference between these two groups is the trend in industry composition change. Despite of the fluctuation, there is no observable trnd for group D cities, all of which are traditionally regarded as industrial centers. Hence the group D is called traditional industrial centers.

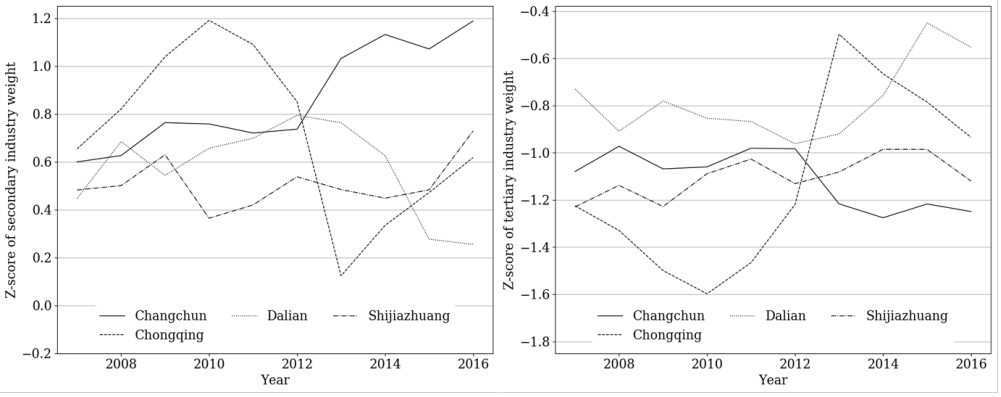


Figure 4: Normalized weights of secondary and tertiary industries for group D cites.

In contrast, group E cities seem to have industrial sectors that are heavy in the economy and is still growing steadily. Similar to Group C, cities in group E are mostly in central China, and are poised to take the industries retired from coastal cities. This strategy seems to be more desirable for group E cities

because of their historical reliance on secondary industries. Group E can be called emerging industrial centers.

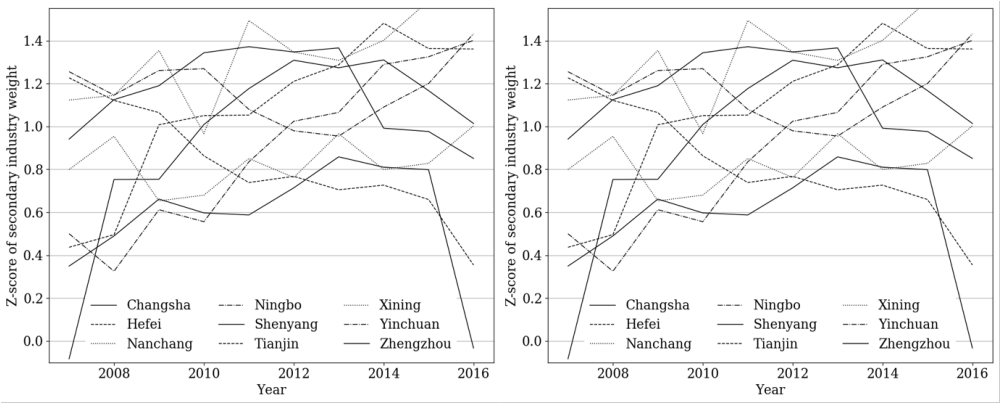


Figure 5: Normalized weights of secondary and tertiary industries for group E cites.

4. Discussion and Conclusion

In this research, we use the time-series cluster analysis to reveal the distinct patterns in Chinese cities' economic growth. In particular, dynamic time warping algorithm is used to measure the time-series distance between two economic growth paths.

Using the weights of secondary and tertiary industries as input variables, the time-series clustering has grouped the 35 major Chinese cities into five categories. According to the industrial composition and the growth trend of each of the five categories, they are labeled as service centers, deindustrializing cities, balanced-industrializing cities, traditional industrial centers, and emerging industrial centers, respectively.

The city categorization as revealed by clustering algorithm shed additional light on how Chinese cities have developed over the last decade. Seemingly distinct cities are shown to share more economical similarities than researchers have thought. Moreover, with the role each city plays become clear, competing and compensating relationships among cities are easier to identify.

This research is a preliminary study on how the unsupervised machine learning techniques can be applied on the field of development economics. It can be extended in various interesting ways. First, more metrics can be included in the cluster analysis, so finer and more complete pictures of economic growth paths can be painted. Second, fuzzy clustering techniques can be applied, since it is common that a city would play multiple roles in an economy. Finally, the economic network among cities is crucial to their development; including graph theory based metrics that capture the network effects into the model would provide additional insight.

References

1. Aghabozorgi, 'S., Shirkhorshidi, A.S. and Wah, T.Y. (2015). Time-series clustering–A decade review. *Information Systems*, 53, 16-38.
2. Alder, S., Shao, L. and Zilibotti, F. (2016). Economic reforms and industrial policy in a panel of Chinese cities. *Journal of Economic Growth,* 4, 305-349.
3. Berndt, D.J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. *KDD workshop*, 16, 359-370.
4. Chan, K.W. and Wan, G. (2017). The size distribution and growth pattern of cities in China, 1982– 2010: Analysis and policy implications. *Journal of the Asia Pacific Economy*, 1, 136-155.
5. Démurger, S. (2001). Infrastructure development and economic growth: an explanation for regional disparities in China. *Journal of Comparative economics*, 29, 95-117.
6. Jain, A.K., Murty, M.N. and Flynn, P.J. (1999). Data clustering: a review. *ACM computing surveys(CSUR),* 3, 264-323.
7. Jeong, Y.S., Jeong, M.K. and Omitaomu, O.A. (2011). Weighted dynamic time warping for time series classification. *Pattern Recognition*, 9, 2231-2240.
8. Izakian, H., Pedrycz, W. and Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance*. Engineering Applications of Artificial Intelligence*, 39, 235-244.
9. Wu, Y. (2016). China's Capital Stock Series by Region and Sector. *Frontiers of Economics in China*,1, 156-172.

Hedonian Brownian motion index for Morocco

Firano Zakaria[1], Filali A. Fatine[1,2]
[1]University of Mohammed V Rabat
[2]CIRPEC[1]

Abstract
In this paper, we propose a new approach to modelling a real estate price index in Morocco, based on the hedonic approach. The basic idea of this paper is to verify the importance of the characteristics of real estate in the real estate price. Thus, based on data from the three major cities of the capital region of Morocco (RABAT Region), we estimated a hedonic model that takes into account spatial autocorrelation. The results obtained through this modelling generally confirm that the surface area and location of real estate (land, house, villa and apartment) have a significant influence on the price of real estate. In addition, and because of the static nature of the database, which refers to one year, we have proposed a new approach to building the real estate price index, namely the stochastic Brownian motion approach. The results claim that this index is in perfect agreement with the real estate price index based on the repeat sales approach used and developed by the Central Bank.

Keywords
Hedonic index; Brownian motion, real estate market.

JEL Classification
G12, E44

1. Introduction

In Morocco, two phases characterize the evolution of the real estate market in Morocco. The first phase is the one before the 2000s, when real estate prices were stagnating with real estate supply out of phase with the needs of the population. From the 2000s and precisely during the year 2003, the real estate sector benefited from several tax and regulatory advantages that boosted the real estate market.

The method adopted in the property price index in Morocco is that of repeat sales. The index based on this approach consists in selecting all the properties that, during the period considered, gave rise to two or more transactions. The difference in price of a building between its purchase and its resale determines the growth of the price of this property during the period.

[1] Centre Interdisciplinaire De Recherche En Performance Et Competitivite

Thus, the index of repeated sales over a period requires observing the goods that are trading during a given time interval. This market-based approach is sound in the sense that it captures all transactions made on the asset in question. However, it neglects the economic aspects and the factors that favor the realization of the sale and the purchase.

To this end, we propose in this paper a new approach to the development of a real estate index. This is indeed the hedonic approach whose design is more or less based on the theoretical training of real estate prices.

Hedonistic price models have been used in housing studies since (Lancaster, 1966) and Rosen (1974) to explore the determinants of housing prices. In the last three decades, this form of modelling has been used to evaluate the value of real estate worldwide. In this design, the choice of housing confers not only the consumption of the property and the structural characteristics of the dwelling, but the consumption of all the characteristics of the property's location such as proximity to environmental benefits and utilities.

This paper aims to model the determinants of real estate prices in Morocco based on the characteristics of goods sold and bought on the Moroccan market. In addition, the paper also proposes to develop the hedonic index by referring to an approach borrowed from stochastic finance where real estate prices are supposed to follow a Brownian geometric movement.

Of those, this article will be structured as follows: in a first, we will present an empirical literature review of work on the issue. Then, a presentation of the methodology and the data used, will allow understanding the nature of the variables retained and the technique used for the development of the new index. Finally, the last part of the article will focus on the presentation of the results and possible interpretations of the new index.

2. Literature review

Early work applying the hedonic modelling approach to real estate prices started in the early 1920s, despite the fact that there is no consensus as to the actual date of their introduction. For example, Colwell and Dilmore (1999) reported that Haas' work in 1922 is the pioneering study to evaluate farmland in Minnesota (USA). Similarly, Bruce and Sundell (1977) have argued that this technique was used in real estate valuation research in 1924. In addition, Wallace (1926) adopted the HPM technique in US cropland. Ridker and Henning (1967) used HPM for the evaluation of air quality and air quality on residential property values.

Other studies have tried to explain the time required to sell a house and the reasons for this decision. Indeed, two approaches have been adopted namely: duration models and linear regression models. The use of duration models is justified by the significance of time in determining selling prices in

the real estate market. The basic assumption is that the more the good is in the market the more its value increases. Sirmans et al. (2005) found that most studies using the hedonic approach and verified the relevance of the temporal variable.

Due to the price dependence of negotiations between buyers and sellers, the use of the hedonic approach does not allow this element to be taken into account. Harding et al. (2003) estimated the effects of trading by including in the hedonic equation a vector of buyer and seller characteristics. Turnbull and Dombrow (2006) study the spatial situation of real estate to describe the impact of neighbourhood on the price of goods. Their results show that the spatial effect dominates and depends on the market trend. To account for the spatial phenomenon, many authors have begun to apply the hedonic model to spatial factors. Dubin (1998) uses the geostatistical method to evaluate the covariance structure of the model. Can (1990) proposes to use models with spatial delay.

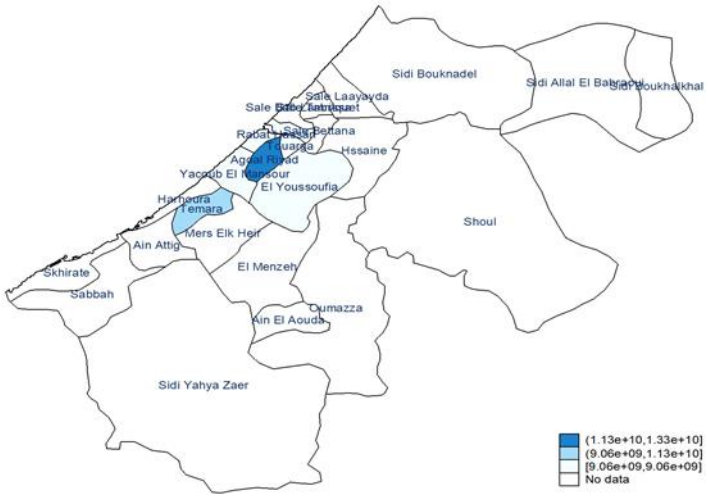
In addition, other works have indicated the importance of the sales season of real estate on sales prices. For example, prices are often higher during the summer periods because of household availability during these times and because of the possibility of hiding property failures during this season (Haurin 1988, Springer 1996, Forgey et al. Knight 2002, Harding et al., 2003).

3. Data and methodology

This paper analyzes the formation of real estate prices in the Rabat-Témara area, which is an important area in Morocco's real estate heritage. In addition to its nomination as capital of the country, the Rabat region is the cultural center of the Kingdom. All the communes of the Rabat region were analysed, only those of Kenitra were not included in the analysis because of the unavailability of the data. We have taken into account all the real estate belonging to this area, be it bare land, apartments, villas and other types of derived habitats.

According to the graph, urban areas in the Rabat region do not have the same value. The Hay Riad-Agdal area is the one that accounts for most of the property value in terms of value. For other areas, the price is almost aligned and there is not much difference, except for the Harhoura area and downtown Rabat.

Figure 1. Value of Real Assets of Rabat Region (data 2014)



The producer of real estate data in Morocco is land conservation. It holds all the information on the characteristics of the goods and the different actions of sales and purchases. We were able to access their 2014 database for the Rabat region. Indeed, our analysis will focus on the central regions of the capital, Témara and Hay Riad-Agdal. These three zones constitute the most important regions in the greater region of the capital of Morocco.

A hedonic price index is established through the estimation of the hedonic function. A hedonic function is an econometric relation between the price of a dwelling and the quantities of the characteristics of a dwelling. In general, the establishment of a hedonic function consists in estimating a statistical model in the form:

$$\ln(p_j) = \propto \sum_{i=1}^{n} \beta_n . X_{n,j} + \varepsilon_j$$

"Pi" is the price of dwelling i and "X" the quantity of characteristic n of dwelling j. in the empirical literature, the endogenous variable can be approximated either by the price of housing or by the price divided by the area (price per square meter). As for the exogenous variables to be considered in the hedonic model, they are generally controversial, often introduced; the geographical location, the size and type of housing, overall any other characteristics likely to affect the evolution of property prices. To this may be added the variables; age or housing facilities.

We start elementarily by considering that the value of a reference stock (here the year of the database, i.e. 2014). To set up the index, it is necessary to have the current year of which we have no data; this configuration will calculate the real estate price index, according to the following training:

$$\text{IPHB}^{2}=\frac{\text{Current value}}{\text{Initial value}}$$

On the basis of this formulation, we can consider that in the absence of real estate asset prices for future years and periods, it is possible to consider that asset prices follow a geometric Brownian motion, where:

$$\ln(S_t) = \ln(S_0) + \left(\mu - \frac{1}{2}\sigma^2\right) + \sigma W_t$$

S_t is the price at time t, S_0 is the initial price, μ is the mean return and $\sigma\sigma$ is the volatility. The real estate's index is:

$$\text{IPHB}=\frac{S1}{S_0} = \frac{\ln(S_0) + \left(\mu - \frac{1}{2}\sigma^2\right) + \sigma W_t}{\ln(S_0)}$$

In this sense, it is necessary to estimate the three main parameters, the price according to the hedonic model, the average of the returns and the volatility. Thus, and in order to determine the value of real estate price returns, we have resorted to the theory of prices in the financial markets, including real estate markets. We know that prices are determined by their intrinsic value, so are the economic determinants of price that can explain its evolution, according to Euler's training we can write that:

$$\mathrm{P}_t = \delta(E_t P_{t+1} + ED_{t+1}) \ (1.1)$$

With: δ = 1/(1 + x) is discount factor and x is discount rate. If one adopts the fundamental design of the Muth, (1961) rationalized asset price evaluation, and accepts that the transversally condition is satisfied, then the fundamental value is considered the only solution to the valuation problem of asset prices:

$$S_0^* = \sum_{i=1}^{n} \delta^i \, ED_{t+i} \qquad (1.2)$$

$S_0^*$is the fundamental value. Thus, we can write:

$$S_t^* = \sum_{i=1}^{n} \delta^i \, ED_{t+i} \ = \sum_{i=1}^{n} \beta_n \, . X_{n,j}$$

Then the coefficients estimated in the hedonic equation, all else being equal, describe the factors of actualization.

In general, the estimation method adopted in hedonic models is ordinary least squares, since the model is generally linear and satisfies the required conditions. However, real estate is of a specific nature where the valuation of property depends on several parameters in addition to the intrinsic

[2] IPHB: Brownian hedonic price index

characteristics. Thus, one of the important factors is the value of neighbouring properties. In fact, the higher the value of a good, the greater the probability that a neighbouring good will have such a high price (phenomenon of real estate mimicry). In this design, it is necessary to take into account the spatial autocorrelation that measures neighbourhood effects. Spatial autocorrelation is based on the observation that spatialized observations in cross sections are not independent. Spatial autocorrelation is defined as the correlation, positive or negative, of a variable with itself arising from the geographic location of the data.[3].

4. Results

To test this spatial autocorrelation we opted for the Geary index (1954) which measures the local spatial dependence between real estate. The value obtained from the Geary index is less than one, making it possible to reject the hypothesis of non-existence of spatial autocorrelation. In this respect, we can say that real estate prices in Rabat-Hay Riad and Témara are auto correlated in space. In other words, the value of goods is influenced by neighbouring prices. The existence of the spatial correlation up to the fifth neighbourhood in the region corroborates the hypothesis of the influence of neighbouring prices on the real estate sales value. Therefore, we proposed a model that takes into account this spatial dependence effect estimated using the instrumental approach to correct the low exogeneity bias. Models were estimated according to the following specification:

$$y_{i,j} = \beta y_{i-t,j} + \sum_{i=1;j=1}^{n,m} \alpha_{j+} \beta_{i.j} X_{i,j} + \varepsilon_{i,j}{}^{4}$$

Where i is the individual dimension describing real estate and j is the nature of the property. Indeed, we have distinguished in the estimation between the different categories of real estate (apartment, villa, land,... etc.). Therefore, we have 8 categories of property and more than 11,000 properties in the region of Rabat-center, 6,500 properties on Hay Riad and 20,400 on the city of Témara.

[3] Julie Le Gallo. Econométrie spatiale (1, Autocorrélation spatiale). [Rapport de recherche] Laboratoire d'analyse et de techniques économiques(LATEC). 2000, 45 p., Table, ref. bib. : 5 p. <hal-01527290>.

[4] $\varepsilon_{i,j}$ is a combination of supposedly random residues and effects specific to types of real estate.

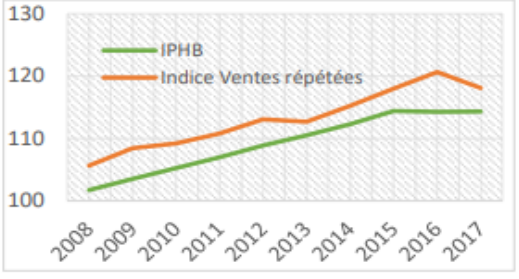
Table 1. Econometrics Result of panel data modelling (all region)

	Rabat-Centre[5]			Hay-Riad			Hay-Riad		
Variable	**Coefficient**	**t-Statistic**	**Prob.**	**Coefficient**	**t-Statistic**	**Prob.**	**Coefficient**	**t-Statistic**	**Prob.**
LOGPRICE (-1)	0.354213	28.00295	0.0000	0.474743	55.71992	0.0000	0.545158	99.77354	0.0000
LOGMETRE	**0.812602**	**37.79219**	**0.0000**	**0.770388**	**46.57118**	**0.0000**	**0.821840**	**54.41956**	**0.0000**
AGE	0.000615	7.446784	0.0000	0.000419	3.474022	0.0005	0.000850	8.214648	0.0000
BALCON	0.056413	1.652031	0.0986	0.163186	1.944212	0.0519	0.026991	1.772235	0.0764
N°ETAGE	0.042172	3.251869	0.0011						
COURS	-0.026646	-2.341135	0.0192				0.073186	4.846090	0.0000
ETAGE	-0.000396	-1.885191	0.0594				0.000843	2.458912	0.0139
D1	-0.021123	-3.804266	0.0001	0.068631	9.254140	0.0000	0.116537	19.56736	0.0000
D3	-0.060075	-12.62286	0.0000						
D4	-0.056507	-2.212520	0.0269						
Garage				0.100649	4.144788	0.0000	0.025104	0.466069	0.6412
D5							0.157300	18.85926	0.0000
D8							0.114777	7.577319	0.0000
CAVE							0.387825	3.015558	0.0026
Piscine							-0.516307	-5.683270	0.0000
R-squared	**0.548718**			**0.660739**			**0.659114**		
J-statistic	**12647.00 (0.00)**			**7140.000**			**10.07631 (0.00)**		

In the three models estimated, we took into account all the potentially explanatory characteristics of real estate prices, while taking into account the specific effects of each type of property.

We applied the standard Brownian motion approach where the price is decomposed into drift and volatility. We considered that the coefficients explaining the influence of the intrinsic and fundamental variables are the factors of actualization. Subsequently, we extracted the different discount rates for the different zones. This made it possible to estimate the drift of the function and to be able to calculate the volatility of the estimated coefficients. Thus via the equation describing IPHB presented previously, we were able to calculate the index over several periods.

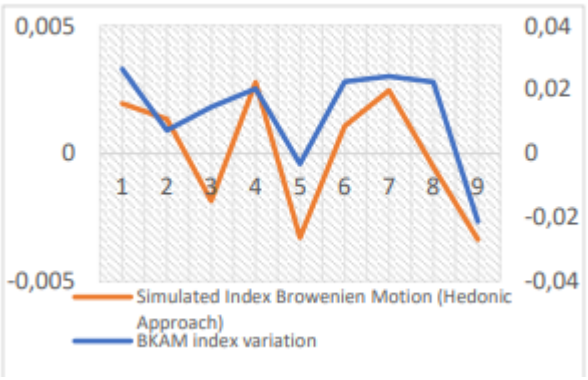
Figure 2. IPHB vs Central bank index (BKAM)



To be able to validate the results obtained via our new approach, we used the index developed by the Central Bank of Morocco (BKAM), which is based on the repeat sales method.

[5] Estimates were made using the two-stage instrumental variables approach to correct the low exogeneity bias

Figure 3. Variation of BKAM index and hedonic index for Rabat Region



The use of the hedonic approach and predictive simulation according to the stochastic model has allowed to give a clue in the behaviour is almost identical to that of the official index of Bank al Maghrib. The cycles translated by the two indices are the same. Indeed, the degree of correlation between the two variations is close to 77%, which reinforces the results obtained via the hedonic approach.

5. Conclusion

Due to the importance of the real estate market in Morocco, public action has spread over the development of a range of indicators for the monitoring of this sector, which is one of the most important sectors in Morocco. In this perspective, two elements are essential, first, the need to understand the factors determining the evolution of property prices in Morocco, and then implement a new, economic approach to the construction of real estate indices that can be complementary to that in use at the central bank.

In this paper, we have developed a hedonic model to identify factors or characteristics that can explain the formation of real estate prices. Our analysis was spread over the Rabat region where three areas were studied namely: the area of the centre of Rabat, the residential and administrative district of Hay-Riad-Agdal and the city of Témara. The results indicate that the spatial correlation factor and the footage are the two major determinants of real estate pricing in the Capital Region. Thus, the mimetic impact of the neighbourhood largely affects the price of goods in the Rabat region and the spatial character described by the square meter is decisive in the negotiations on the market.

References

1. Abraham, Jesse M., Hendershott, Patric H., 1996. Bubbles in metropolitan housing markets. Journal of Housing Research 7, 191–207.

2. Abraham, Jesse M., Schauman,William S. 1991. New evidence on home prices from Freddie Mac repeat sales. Real Estate Economics 19, 333–352.
3. Case,Karl E., Shiller, Robert J., 1987. Prices of single-family homes since 1970: new indexes for four cities. New England Economic Review, 46–56.
4. Case, Karl E., Shiller, Robert J., 1989. The efficiency of the market for single-family homes. The American Economic Review 1, 125–137.
5. Case, Karl E., Shiller, Robert J., 1990. Forecasting prices and excess returns in the housing market. Real Estate Economics 18, 253–273.
6. Cauley, S.,Pavlov,A.,Schwartz,E., 2007. Homeownership as a constraint on asset allocation. Journal of Real Estate Finance and Economics 34, 283–311.
7. Cavanagh,C., Elliott,G., Stock, J., 1995. Inference in models with nearly integrated regressors. Econometric
8. Theory 11, 1131–1147.
9. Clapham, Eric, Englund, Peter, Quigley, John M., Redfearn, Christian L., 2006. Revisiting the past and settling the score: index revision for house price derivatives. Real Estate Economics 34, 275–302. Forecasting Real Estate Prices 575
10. Webb,Cary, 1981a.A discrete random walk model for price levels in real estate.Working Paper, Department of Mathematics,Chicago State University.
11. Webb, Cary, 1981b. The expected accuracy of a real estate price index. Working Paper, Department of Mathematics. Chicago State University.
12. Webb, Cary, 1981c. Trading activity and the variance of an index number.Working Paper, Department of Mathematics,Chicago Statr University.
13. Welch, Ivo, Goyal, Amit, 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455–1508.

Combined criteria for dose optimisation in early phase clinical trials

M. Iftakhar Alam[1], D. Stephen Coad [2], Barbara Bogacka[2]
[1]Institute of Statistical Research and Training, University of Dhaka, Dhaka 1000, Bangladesh
[2]School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, U.K.

Abstract
The paper aims to investigate whether any bridge is possible between so-called best intention and D-optimum designs. It introduces combined criteria for dose optimisation in seamless phase I/II adaptive clinical trials. Each of the optimality criteria considers efficacy and toxicity as endpoints and is based on the probability of a successful outcome and on the determinant of the Fisher information matrix for estimation of the doseresponse parameters. In addition, one of the criteria incorporates penalties for choosing a toxic or inefficacious dose. Starting with the lowest dose, the adaptive design selects the dose for each subsequent cohort that maximises the respective defined criterion. The methodology is illustrated with a dose-response model that assumes trinomial responses. Simulation studies show that the method is capable of identifying the optimal dose accurately without exposing many patients to toxic doses.

Keywords
Adaptive design; Continuation ratio model; D-optimum design; Penalty function; Phase I/II trial.

1. Introduction

Different methods are being developed to increase the popularity of seamless phase I/II clinical trials. There are designs proposed by Thall and Russell (1998), Thall and Cook (2004) and Zhang et al. (2006). These designs have the intention of allocating the best dose to the cohort of patients based on the current knowledge and are known as the "best intention designs". They may lead to poor learning of the dose-response relationship. In contrast, there are methods which rely on optimal design criteria for estimation of model parameters. Heise and Myers (1996) constructed the *D*-optimal design using the Gumbel model for bivariate binary data. Fan and Chaloner (2004) described the *D*-optimal design for trinomial responses using a continuation ratio model. Dragalin and Fedorov (2006) considered binary outcomes for each endpoint and used either Gumbel bivariate binary logistic regression or the Cox bivariate binary model.

This paper investigates whether any trade-off between the best intention designs taking care of individual ethics and *D*-optimal designs focusing more on collective ethics is possible. The underlying idea is to develop a design that

exposes few cohorts in a trial to either subtherapeutic or toxic doses and that can also find the optimum dose accurately.

2. Methodology

Although the penalised *D*-criterion in Dragalin and Fedorov (2006) introduces a penalty function to improve the quality of treatment during dose escalation, we have found it not to be improved as expected. Therefore, further effort has been taken with the combined criterion defined below. Also, clinicians may be interested in achieving several objectives, such as efficient estimation of the model parameters and allocation of the most efficacious doses to the cohorts during a clinical trial. The combined criteria in (1) and (2) balance these two objectives.

The penalised combined criterion is a linear combination of the determinant of the Fisher information matrix for the dose-response model, penalised for inefficacy and toxicity, and the probability of success. On the other hand, the simple combined criterion does not penalise for inefficacy and toxicity. At each stage of the adaptive trial, we select that dose for which the criterion is maximised.

To implement the penalised criterion, we initially determine the doses $x_{k+1}^{\psi s}$ and x_{k+1}^{PD} that maximise the probability of success and the determinant of the penalised Fisher information matrix (FIM), respectively. Since the determinant and the probability of success may have quite different magnitudes, we scale them at the dose x as

$$E_{PD}(x) = \frac{\Phi_{PD}\{M(x|\boldsymbol{\xi}_k, \widehat{\boldsymbol{\vartheta}}_k)\}}{\{M(\boldsymbol{x_{k+1}^{PD}}|\boldsymbol{\xi}_k, \widehat{\boldsymbol{\vartheta}}_k)\}} \text{ and } E_{\psi s}(x) = \frac{\psi s(X, \widehat{\boldsymbol{\vartheta}}_k}{\psi s(x_{k+1}^{\psi s}, \widehat{\boldsymbol{\vartheta}}_k)}.$$

The penalised combined criterion then selects the dose x_{k+1} for the next cohort of patients so that

$$x_{k+1} = \underset{x \in \chi}{\operatorname{argmax}}\{aE_{PD}(x) + (1-a)E_{\psi s}(x)\}. \quad (1)$$

where a is some weight such that $0 \leq a \leq 1$.

If the D-optimum dose x_{k+1}^{D} and $E_D(x)$are chosen instead, then the simple combined criterion takes the form

$$x_{k+1} = \underset{x \in \chi}{\operatorname{argmax}}\{aE_D(x) + (1-a)E_{\psi s}(x)\}. \quad (2)$$

This is a special case of the penalised combined criterion if $C_S = C_T = C = 0$ in the penalty function. This is due to the fact that, when $C = 0$, the penalised D-criterion reduces to the D-criterion. Obviously, the results will depend on the choice of a . It is clear that, when $a = 1$, the combined criterion is simply the penalised D-criterion or D-criterion. Similarly, for $a = 0$, we have dose

selection based on the probability of success only. For any other choice of α, the design is expected to allocate the most efficacious doses to the cohorts and to give precise estimates of the parameters leading to the recommendation of the best dose for further study.

An Example

To explore the proposed methodology, we introduce an example which is based on the continuation ratio doseresponse model. Simulation studies, detailed in Section 2, are conducted to investigate the properties of the design.

For an experimental drug, we use a flexible continuation ratio model (Agresti, 1990), which is given by

$$\log\left\{\frac{\varphi_1(x,\boldsymbol{\vartheta})}{\varphi_0(x,\boldsymbol{\vartheta})}\right\} = \vartheta_1 + \vartheta_2 x \text{ and } \log\left\{\frac{\varphi_2(x,\boldsymbol{\vartheta})}{1-\varphi_2(x,\boldsymbol{\vartheta})}\right\} = \vartheta_3 + \vartheta_4 x$$

The above equations have the following solutions

$$\varphi_0(x,\boldsymbol{\vartheta}) = \frac{1}{\left(1+e^{\vartheta_1+\vartheta_2 x}\right)\left(1+e^{\vartheta_3+\vartheta_4 x}\right)},$$

$$\varphi_1(x,\boldsymbol{\vartheta}) = \frac{e^{\vartheta_1+\vartheta_2 x}}{\left(1+e^{\vartheta_1+\vartheta_2 x}\right)\left(1+e^{\vartheta_3+\vartheta_4 x}\right)},$$

and

$$\varphi_2(x,\boldsymbol{\vartheta}) = \frac{e^{\vartheta_3+\vartheta_4 x}}{1+e^{\vartheta_3+\vartheta_4 x}}.$$

If we are at the kth stage in a trial, then k cohorts have been treated with doses selected from the set of ordered doses χ. Let x be the $k \times 1$vector of doses with components x_1 and let R be the $k \times 3$ outcome matrix with $R_l = (R_{l0}, R_{l1}, R_{l2})$ as the lth row, $l = 1,2,\ldots,$ k. Note that $R_{l0}, R_{l1}, R_{l2} = c$, where c is the number of subjects in a cohort treated with dose x_l The successive components of R_l are the counts of neutral, successful and toxic responses for the lth cohort. Thus, the likelihood function is

$$L_k(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{R}) \propto \prod_{l=1}^{k}\{\psi_0(x_1,\boldsymbol{\vartheta})\}^{R_{10}}\{\psi_1(x_1,\boldsymbol{\vartheta})\}^{R_{11}}\{\psi_2(x_1,\boldsymbol{\vartheta})\}^{R_{12}}.$$

Since maximum likelihood estimation is unsuitable because of small sample sizes at the early stages of a trial, we employ a Bayesian approach to estimate the parameters $\boldsymbol{\vartheta}$. The posterior estimate of $\boldsymbol{\vartheta}$ at the kth stage is

$$\widehat{\boldsymbol{\vartheta}}_k = \frac{\int_\Theta \boldsymbol{\vartheta} p^{(\boldsymbol{\vartheta})} L_k(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{R}) d\boldsymbol{\vartheta}}{\int_\Theta \boldsymbol{\vartheta} p^{(\boldsymbol{\vartheta})} L_k(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{R}) d\boldsymbol{\vartheta}},$$

where p(ϑ) is the joint prior distribution of the parameters. Let us assume that $0 < \vartheta_2 < \mu_1, 0 < \vartheta_4 < \mu_2, v_1 < \vartheta_1 < v_2$, and $v_3 < \vartheta_3 < v_4$, and that the joint prior distribution is uniform. Then we obtain

$$p(\boldsymbol{\vartheta}) = \frac{2}{\mu_1\mu_2(v_2 - v_3)^2}, \boldsymbol{\vartheta}\epsilon\, \widetilde{\Theta},$$

where

$$\widetilde{\Theta} = \{\boldsymbol{\vartheta}: v_3 < \vartheta_3 \leq \vartheta_1 < v_2, 0 < \vartheta_2 \leq \mu_1, 0 < \vartheta_4 \leq \mu_2\}$$

The associated FIM $\boldsymbol{I}(x, \boldsymbol{\vartheta})$ for the model parameters can be obtained easily.

Simulation settings

Six dose-response scenarios are considered for the simulation study, as shown in Figure 1. The parameter values for these scenarios are chosen to obtain various shapes for the dose-response curves. For each scenario, it is assumed that 20 doses are available in the set χ = {0.5,1.0,...,10.0} and that the acceptable level for the probability of toxicity is Υ = 0.2. Also, the minimum success probability that an OD should have is assumed to be δ = 0.5. To check stopping for futility and/or toxicity, we set $\lambda = \delta - \gamma = 0.3$. Each trial assigns the

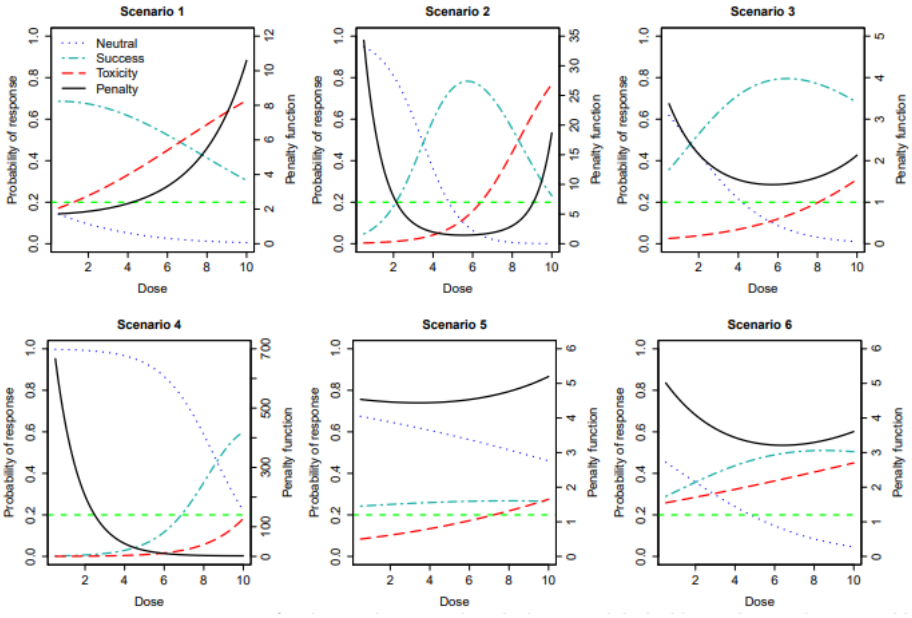


Figure 1: Dose-response scenarios for the simulation study. The horizontal dashed line indicates the acceptable level for toxicity and the penalty functions for the scenarios are obtained assuming that $C_S = C_T = 1$.

lowest available dose of 0.5 mg/kg body weight to a cohort of patients. After obtaining responses from the first cohort, the dose escalation is based on the criterion in (1) or in (2), using the updated posterior means of the dose-response parameters at each stage. The posterior means are obtained through numerical integration using the *R* package *cubature* (Johnson and Narasimhan, 2013). We use a joint uniform prior distribution for ϑ. The

parameter space $\widetilde{\Theta}$ is chosen for each scenario so that the true values of the parameters lie in the middle of the corresponding intervals. For instance, since Scenario 1 has the true parameters $\vartheta = (1.44, 0.26, -1.70, 0.25)^T$, we consider $\widetilde{\Theta} = \{\vartheta \in \Theta: 0 < \vartheta_1 < 2.88, 0 < \vartheta_2 < 0.52, -3.40 < \vartheta_3 < 0, 0 < \vartheta_4 < 0.50\}$. The same approach is followed for the other scenarios. A trial can stop early for futility and/or toxicity. Apart from that, it stops when the same dose is repeated for r cohorts or when the trial reaches the maximum number of m cohorts, whichever comes first. It is assumed that $r = 6$ and $m = 20$. The number of patients in each cohort is 3, that is, $c = 3$.

3. Results

With the control parameter values set to 1, we run the penalised combined criterion for each of the scenarios. One thousand simulated trials are generated in each case for various values of the weight a. Note that, when $a = 0$, the dose selection is based on the probability of success only. On the other hand, the criterion reduces to the penalised D-criterion for $a = 1$: see (1).

Table 1 illustrates the simulation results for the penalised combined criterion for the considered scenarios. The higher the values of %OD and %AD are, the better the design is. Similarly, %TD is expected to be as small as possible. The sampling and decision efficiency measures can be obtained once the distributions of dose allocation and optimum dose selection are available. The information obtained per observation, information obtained per cost and various risk measures are also obtained. We expect the information per observation and information per cost to be as much as possible. Similarly, the risks are expected to be as small as possible.

Table 1: Performance of the penalised combined criterion for the six scenarios. Percentage of optimum doses chosen as the true optimum one, recommended for further studies (%OD), percentage of no dose recommended (%ND), percentage of doses recommended as optimum, but carrying the probability of toxicity above the acceptable level (%TD), percentage of cohorts treated at the true optimum doses throughout the trials (%AD), decision efficiency (DE), sampling efficiency (SE), average cohorts (AC), information per observation (IPO), information per cost (IPC) and some risk measures. The value of a in bold is regarded as the best in terms of the performance measure DE.

Scenario	a	%OD	%ND	%TD	%AD	DE	SE	AC	IPO	IPC	c		
											Targeted population	Patient sample	nth patient
1	0.0	63.2	5.6	21.6	42.9	0.728	0.593	10.5	0.003	0.083	0.563	14.7	1.42
	0.2	72.6	6.5	17.8	48.2	0.757	0.547	10.3	0.009	0.125	0.455	17.9	1.73
	0.4	75.6	7.7	15.2	45.8	0.771	0.481	11.2	0.046	0.212	0.469	39.7	3.54
	0.6	86.8	6.2	6.6	50.2	0.872	0.508	11.0	0.087	0.247	0.255	51.4	4.66
	0.8	91.7	6.4	1.7	48.5	0.918	0.484	11.7	0.136	0.266	0.084	72.0	6.13
	1.0	76.1	14.1	6.3	34.0	0.796	0.339	15.6	0.188	0.272	0.133	151.5	9.47
	$b(\hat{\pi})$	28.0	14.9	37.9	-	0.472	-	20.0	-	-	1.495	-	-
2	0.0	68.8	3.8	7.7	39.4	0.872	0.665	16.8	0.011	0.093	1.423	62.9	3.74
	0.2	66.3	4.1	8.0	38.9	0.863	0.656	16.9	0.012	0.094	1.533	63.8	3.78
	0.4	70.4	2.6	7.0	39.2	0.890	0.656	17.3	0.013	0.098	1.050	64.7	3.74
	0.6	68.4	3.4	7.5	37.0	0.876	0.625	17.7	0.019	0.107	1.308	68.4	3.86
	0.8	67.4	4.9	8.8	34.7	0.847	0.603	18.3	0.025	0.114	1.795	74.1	4.05
	1.0	41.0	7.7	5.1	17.2	0.777	0.498	19.7	0.043	0.118	3.508	104.9	5.31
	$b(\hat{\pi})$	71.5	0.0	5.2	-	0.863	-	20.0	-	-	0.288	-	-
3	0.0	88.7	0.2	1.5	58.7	0.974	0.869	18.0	0.095	0.332	0.765	103.8	5.78
	0.2	87.1	0.2	2.1	57.7	0.967	0.851	18.1	0.104	0.339	0.816	105.9	5.84
	0.4	89.4	0.1	1.2	57.7	0.978	0.839	18.2	0.116	0.348	0.742	107.5	5.89
	0.6	88.3	0.2	2.2	54.7	0.967	0.802	18.6	0.141	0.364	0.868	113.7	6.11
	0.8	87.2	0.1	2.2	51.0	0.968	0.766	18.9	0.176	0.382	0.874	121.7	6.44
	1.0	60.9	0.0	1.6	26.5	0.921	0.657	20.0	0.334	0.442	3.797	205.8	10.30
	$b(\hat{\pi})$	76.3	0.0	5.2	-	0.935	-	20.0	-	-	1.113	-	-
4	0.0	71.2	15.1	0.0	28.6	0.816	0.544	17.9	0.017	0.007	16.025	355.1	19.78
	0.2	71.7	13.6	0.0	28.7	0.821	0.542	17.9	0.017	0.007	14.996	357.0	19.88
	0.4	74.4	13.6	0.0	29.7	0.829	0.541	17.8	0.017	0.007	14.404	356.5	20.06
	0.6	73.1	14.3	0.0	29.4	0.814	0.538	17.8	0.017	0.007	15.335	358.0	20.10
	0.8	76.6	11.8	0.0	31.0	0.847	0.539	17.6	0.017	0.007	12.788	356.0	20.21
	1.0	47.9	23.7	0.0	26.4	0.647	0.506	19.2	0.018	0.007	26.495	381.7	19.84
	$b(\hat{\pi})$	60.1	10.4	0.0	-	0.873	-	20.0	-	-	10.48	-	-
5	0.0	-	84.4	10.2	-	0.844	-	17.3	1.699	0.215	10.785	615.4	35.54
	0.2	-	79.4	16.9	-	0.794	-	16.6	2.069	0.228	17.115	628.5	37.82
	0.4	-	73.8	22.7	-	0.738	-	16.4	2.499	0.245	22.898	661.8	40.3
	0.6	-	66.0	28.8	-	0.660	-	16.2	2.992	0.262	28.865	660.5	40.72
	0.8	-	57.3	32.7	-	0.573	-	15.6	3.833	0.281	32.746	617.3	39.41
	1.0	-	99.9	0.0	-	0.999	-	17.5	2.858	0.270	0.000	569.2	32.49
	$b(\hat{\pi})$	-	99.9	0.0	-	0.999	-	20.0	-	-	0.015	-	-
6	0.0	-	99.1	0.9	-	0.991	-	1.9	0.097	0.037	0.337	21.4	11.23
	0.2	-	99.8	0.2	-	0.998	-	1.7	0.081	0.036	0.200	13.3	7.57
	0.4	-	99.8	0.2	-	0.998	-	1.8	0.084	0.037	0.048	15.3	8.47
	0.6	-	99.7	0.3	-	0.997	-	1.9	0.115	0.044	0.242	19.3	9.89
	0.8	-	99.5	0.5	-	0.995	-	1.8	0.102	0.037	0.500	17.5	9.70
	1.0	-	99.9	0.1	-	0.999	-	1.9	0.137	0.038	0.016	17.7	9.28
	$b(\hat{\pi})$	-	93.5	6.5	-	0.935	-	20.0	-	-	2.054	-	-

Although many measures are presented, the best a is regarded based on DE, as it reflects the quality of decision regarding the future of an experimental drug.

Consider the results for Scenario 1. The penalised *D*-criterion ($a = 1$) gives better results than maximisation of the probability of success only ($a = 0$) in terms of most of the measures. However, the SE is the highest when $a = 0$. This is expected, as the criterion based on the maximisation of the probability of success is defined so that more patients are treated during the trial with efficacious doses. The design performs the best overall when $a = 0.8$, indicating the importance of the penalised *D*-criterion. Both the information per observation and information per cost increases as a increases. The risk for targeted population is minimum when $a = 0.8$. As a increases, the risks for both patient sample and nth patient increases. This is obvious as the design puts more weights on the penalised *D*-optimality with the increase in a. It is clear from the measures that our design outperforms the nonparametric benchmark design by Cheung (2014).

The results obtained from the simulated trials utilising the simple combined criterion in (2) are presented in Table 2. Whereas the penalised combined criterion selects the true OD 91.7% of the time when $a = 0.8$, the simple combined criterion selects this dose only 88.4% of the time. On the

whole, the combined criterion with $a = 0.8$ enhances the performance for Scenario 1, whether *D*-optimality is penalised or not.

Now consider the results for Scenario 2 in Table 1. The best result in terms of DE is attained when $a = 0.4$. Of the other values, $a = 0.0$ and $a = 0.6$ also provide satisfactory results. In fact, there is little variation in the performance for the values of a between 0 and 0.6. The most noticeable difference is observed for $a = 1$. For this choice of weight, the design identifies the OD much less accurately compared with the other weights. Fewer cohorts are treated with the optimum dose as the weight increases and there is also a sharp drop when $a = 1$. The efficiency measures DE and SE are both considerably smaller in this case as well. The ratio of the probability of correct identification of the true OD at $a = 0.4$ relative to the benchmark is 70.4 ÷ 71.5 × 100 = 98%. However, the DE at $a = 0.4$ is well above that at the benchmark design. As seen in Table 2, the simple combined criterion produces very similar results for this scenario. However, the best a is observed at 0.0.

We obtain similar results for the weights ranging between 0 and 0.8 in Scenario 3 in Table 1. The best a is observed at 0.4. As in the previous scenario, we observe a decreasing trend in the measures %AD for Table 2: Performance of the simple combined criterion for the six scenarios. Percentage of optimum doses chosen as the true optimum one, recommended for further studies (%OD), percentage of no dose recommended (%ND), percentage of doses recommended as optimum, but carrying the probability of toxicity above the acceptable level (%TD), percentage of cohorts treated at the true optimum doses throughout the trials (%AD), decision efficiency (DE), sampling efficiency (SE), average cohorts (AC), information per obervation (IPO), information per cost (IPC) and some risk measures. The value of a in bold is regarded as the best in terms of the performance measure DE.

Scenario	a	%OD	%ND	%TD	%AD	DE	SE	AC	IPO	IPC	Risk		
											Targeted population	Patient sample	*n*th patient
1	0.0	64.3	6.2	20.0	44.6	0.738	0.606	10.3	0.003	-	0.464	13.4	1.31
	0.2	68.8	5.5	19.6	45.2	0.749	0.589	10.6	0.005	-	0.475	17.1	1.61
	0.4	73.0	7.7	16.1	43.6	0.762	0.469	11.6	0.043	-	0.454	41.6	3.57
	0.6	80.5	9.8	8.8	46.6	0.814	0.473	11.4	0.117	-	0.286	65.7	5.74
	0.8	88.4	7.3	3.7	46.0	0.889	0.460	12.1	0.214	-	0.143	104.4	8.62
	1.0	40.0	32.4	15.0	16.8	0.526	0.168	18.8	0.272	-	0.333	354.6	18.87
	$b(\hat{\pi})$	28.0	14.9	37.9	-	0.472	-	20.0	-	-	1.495	-	-
2	**0.0**	68.5	2.2	7.7	40.0	0.887	0.669	16.8	0.010	-	0.930	62.6	3.73
	0.2	69.6	3.8	7.9	39.9	0.871	0.658	17.0	0.011	-	1.405	63.5	3.73
	0.4	69.2	3.7	7.5	39.6	0.876	0.652	17.2	0.013	-	1.352	64.8	3.77
	0.6	68.2	4.5	7.3	36.6	0.867	0.622	17.7	0.019	-	1.665	69.1	3.90
	0.8	66.5	4.4	6.3	34.7	0.877	0.604	18.3	0.026	-	1.620	74.5	4.07
	1.0	42.7	7.4	6.8	13.1	0.747	0.387	19.8	0.058	-	3.782	141.4	7.14
	$b(\hat{\pi})$	71.5	0.0	5.2	-	0.863	-	20.0	-	-	0.288	-	-
3	0.0	86.3	0.0	2.3	57.9	0.967	0.866	17.8	0.094	-	0.776	104.3	5.8
	0.2	89.5	0.1	1.6	58.3	0.974	0.858	18.1	0.100	-	0.719	104.8	5.8
	0.4	88.9	0.1	1.3	56.8	0.977	0.836	18.2	0.117	-	0.723	108.2	5.9
	0.6	88.2	0.2	2.2	54.7	0.967	0.800	18.6	0.145	-	0.810	113.7	6.1
	0.8	88.2	0.1	2.2	50.6	0.968	0.758	18.9	0.182	-	0.883	123.0	6.5
	1.0	77.8	0.2	5.0	29.2	0.914	0.657	19.9	0.365	-	2.466	226.3	11.3
	$b(\hat{\pi})$	76.3	0.0	5.2	-	0.935	-	20.0	-	-	1.113	-	-
4	0.0	69.0	15.7	0.0	27.9	0.795	0.542	18.0	0.016	-	17.268	356.7	19.8
	0.2	71.6	13.2	0.0	28.4	0.818	0.542	17.9	0.017	-	14.680	355.0	19.8
	0.4	68.6	15.6	0.0	28.4	0.794	0.536	17.9	0.016	-	17.059	361.7	20.2
	0.6	75.5	12.1	0.0	30.2	0.836	0.540	17.7	0.017	-	13.552	356.5	20.2
	0.8	76.1	10.6	0.0	31.0	0.847	0.539	17.6	0.017	-	11.879	354.0	20.1
	1.0	32.2	30.0	0.0	24.8	0.523	0.491	19.7	0.018	-	34.431	404.0	20.5
	$b(\hat{\pi})$	60.1	10.4	0.0	-	0.873	-	20.0	-	-	10.48	-	-

	0.0	-	81.3	12.4	-	0.813	-	17.3	1.733	-	13.300	618.8	35.7
	0.2	-	78.8	16.5	-	0.788	-	16.8	2.038	-	16.691	642.5	38.2
5	0.4	-	76.7	20.7	-	0.767	-	16.7	2.392	-	20.820	659.2	39.5
	0.6	-	62.9	32.3	-	0.629	-	16.1	3.170	-	32.371	671.5	41.7
	0.8	-	53.3	39.3	-	0.533	-	15.7	4.175	-	39.342	660.1	42.0
	1.0	-	100.0	0.0	-	1.000	-	17.5	3.008	-	0.0	591.1	33.8
	$b(\hat{\pi})$	-	99.9	0.0	-	0.999	-	20.0	-	-	0.015	-	-
	0.0	-	99.5	0.5	-	0.995	-	1.8	0.087	-	0.251	14.8	8.2
	0.2	-	99.8	0.2	-	0.998	-	1.8	0.090	-	0.142	13.9	7.7
6	0.4	-	99.7	0.3	-	0.997	-	1.8	0.100	-	0.137	19.1	10.5
	0.6	-	99.4	0.6	-	0.994	-	1.8	0.105	-	0.317	17.3	9.4
	0.8	-	99.7	0.3	-	0.997	-	1.8	0.093	-	0.212	18.9	10.3
	1.0	-	99.7	0.3	-	0.997	-	1.8	0.103	-	0.115	13.0	7.3
	$b(\hat{\pi})$	-	93.5	6.5	-	0.935	-	20.0	-	-	2.054	-	-

dose allocation and SE for the sampling efficiency, with a sharp drop at $a = 1$. All of the performance values indicate that the penalised *D*-criterion on its own is performing poorly compared to the other cases. However, it might be worth combining the criteria in this case, with $a \in [0.2, 0.8]$. Also, the proposed design outperforms the benchmark design in this scenario. Similar conclusions can be drawn in the case of the simple combined criterion, as seen in Table 2.

For the penalised combined criterion in Scenario 4, it can be argued that the design is performing similarly for weights between 0.4 and 0.8. But the penalised *D*-optimum design on its own is not performing well in this scenario either. The proposed design is more efficient than the benchmark design in identifying the true OD. However, DE at the benchmark design is the maximum value in this scenario. It happens as the distrbution of OD in the benchmark design is more around the true OD than that by the other design. A good percentage of trials do not recommend any dose for further developtment in this scenario. As the most extreme dose is the true OD here, a trial has more chance to stop early for futility. The results obtained for the simple combined criterion are found to be quite competitive with those for the penalised combined criterion.

The design is most efficient when $a = 1$ in Scenario 5. The DE decreases until $a = 1$. The design is equally as efficient as the benchmark design when $a = 1$. The similar results are found when the simple combined criterion is used. In Scenario 6, the performance of the penalised combined is very similar across the values of a. Also, the DE of our design at these values is well above that at the benchmark design. The average number of cohorts utilised in each trial is very small. Also, the results are very consistent with those produced by the simple combined criterion in Table 2. Since very small number of cohorts are engaged in each trial, the penalised/simple combined has little role in the identification of the OD. It is the stopping rule for futility and/or toxicity that plays the significant role and thereby lead to the very similar results.

Interesting observations can also be made by comparing the results for penalised and non-penalised *D*-optimality, that is, when $a = 1$. Tables 1 and 2 clearly show the superiority of the penalised criterion. Most of the measures have better values for the scenarios. It is worth mentioning that the penalised *D*-optimum design ($a = 1$) always requires on average more cohorts than the designs for $a = 0$ and for the best a. The penalised combined criterion with

the best a needs very similar numbers of cohorts to the design with $a = 0$. A similar trend is found when the simple combined criterion is utilised.

4. Discussion and Conclusion

We have looked at three different approaches for dose finding. In the first approach, the intention is to allocate doses to the patients that are most efficacious according to the current knowledge while searching for the best dose, an approach known in the literature as the "best intention". The second approach targets the most effective, from the parameter estimation point of view, gathering of information and consequently should give the best estimate of the optimum dose. The third approach tries to achieve a trade-off between the two. Best intention designs are ethically attractive, as they take care of the patients, but, unlike the one based on the *D*-optimality criterion, they have limitations in terms of convergence to the optimum dose. Pronzato (2000) and Fedorov et al. (2011) report that "best intention" designs may converge to a sub-optimal dose. Their studies are based on the frequentist approach and use the least squares or the maximum likelihood estimates of the parameters. We are using Bayesian parameter estimation, and, to our knowledge, the convergence properties are not known in this case.

The gains in the combined criteria over the penalised *D*-criterion or *D*-criterion are evident from the presented results. All of the performance measures are found to be improved. Most importantly, we notice an appreciable improvement in the quality of treatment allocation, reflected through the sampling efficiency measure SE and the measure of OD allocation during the trial, %AD. The quality of optimum dose selection for the next phase, presented through the DE, is also found to be improved. The combined criteria also outperform the criterion based on the maximisation of the probability of success. In general, the combined criteria utilise a reasonable number of cohorts compared to the other two designs.

All of these results guide us to recommend the proposed combined criteria as dose-optimisation tools in early phase clinical trials. In terms of performance, the penalised combined criterion does slightly better than the simple combined criterion. The choice of values for a will solely depend on the objective. In extreme scenarios like 1, 4, 5 and 6, we have seen that high values of a perform surprisingly well. The middle values are also found to perform satisfactorily in the majority of the scenarios. Since, in reality, we may not know the shape of the dose-response relationship in advance of the trial, we suggest using the middle values. Alternatively, we can have some idea on the optimum value of a during the progress of a trial. Once a trial has come through some reasonable number of stages, we can locate where the optimum dose may lie based on the estimates of probability of success and toxicity at hand. If it is found to be either at the lower end or at the upper end of dose region, we can

use higher values of α afterwards. However, if the OD is likely to be in the middle of dose region, we can use middle values of α.

References

1. Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York
2. Cheung, Y. K. (2014). Simple benchmark for complex dose finding studies. *Biometrics* 70 (2), 389–397.
3. Dragalin, V. and V. Fedorov (2006). Adaptive designs for dose-finding based on efficacy-toxicity response. *Journal of Statistical Planning and Inference* 136 (6), 1800–1823.
4. Fan, S. K. and K. Chaloner (2004). Optimal designs and limiting optimal designs for a trinomial response. *Journal of Statistical Planning and Inference* 126 (1), 347–360.
5. Fedorov, V. V., N. Flournoy, Y. Wu, and R. Zhang (2011). Best intention designs in dose-finding studies. Preprint.
6. Heise, M. A. and R. H. Myers (1996). Optimal designs for bivariate logistic regression. Biometrics 52 (2), 613–624.
7. Johnson, S. G. and B. Narasimhan (2013). *Cubature: Adaptive Multivariate Integration Over Hypercubes*. R package version 1.1-2.
8. Pronzato, L. (2000). Adaptive optimization and D-optimum experimental design. Annals of Statistics 28 (6), 1743–1761.
9. Thall, P. F. and J. D. Cook (2004). Dose-finding based on efficacy–toxicity trade-offs. Biometrics 60 (3), 684–693.
10. Thall, P. F. and K. E. Russell (1998). A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics* 54 (1), 251–264.
11. Zhang, W., D. J. Sargent, and S. Mandrekar (2006). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine* 25 (14), 2365–2383.

Parameter estimation for misspecified diffusion processes with noisy, nonsynchronous observations

Teppei Ogihara
The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo

Abstract

Forecasting variances of stocks and covariances of stock pairs is an important task to control the loss from stock assets for many financial institutions which hold a huge amount of stocks. Statistical analysis of stock price data and data of financial statements is useful for this purpose. Nowadays, we can easily get intraday stock prices data such as all transactions of a stock in a day. Then, the study of highfrequency data becomes more important because huge information of highfrequency data enable us to forecast stock variances and covariances more accurately. However, there are two problems on statistical analysis of highfrequency data. The first one is market microstructure noise: when we model stock prices by using diffusion processes, some empirical facts suggest the existence of additional noise. The second one is nonsynchronous observations: We observe stock prices when transactions occur. So observation times must be different for different stocks.

In this talk, we study parametric inference under the existence of market microstructure noise and nonsynchronous observations. We study maximumlikelihood-type estimation for parametric diffusion processes with noisy, nonsynchronous observations, assuming that the true model is contained in the parametric family. We further study the case that this assumption is not satisfied. Such a model is called a misspecified model. Ogihara (2018) studied maximum-likelihood-type and Bayes-type estimation for a model of parametric diffusion processes with noisy, nonsynchronous observations, and showed asymptotic mixed normality of the estimators with the convergence rate $n^{-1/4}$. In this model, we assume that the true model is contained in the parametric family.

In practice for high-frequency data, to satisfy the assumption that the true model is contained in the parametric family, we need to choose the parametric family carefully so that it accurately captures microstructure of stock prices. This is a difficult task because several empirical facts of a stock market (intra-day seasonality, volatility clustering, complicated dependence structure of stocks, and so on) make it difficult to capture the stock microstructure. On the other hand, high-frequency data contains huge information and therefore machine learning methods such as neural network or support vector machine are useful to identify the structure of the diffusion coefficient. In this approach, we need to consider a theory of misspecified model.

We will study the asymptotic theory of a maximum-likelihood-type estimator for misspecified model. In this setting, the original maximum-likelihood-type estimator cannot attain the optimal convergence rate $n^{-1/4}$ due to the asymptotic bias. We construct a new estimator which attains the optimal rate by using a bias correction and show the asymptotic mixed normality.

1. Introduction

Forecasting variances of stocks and covariances of stock pairs is an important task to control the loss from stock assets for many financial institutions which hold huge amount of stocks. Statistical analysis of stock price data and data of financial statements is useful for this purpose. Nowadays, we can easily get intraday stock prices data such as all transactions of a stock in a day. Then, the study of high-frequency data becomes more important because huge information of high-frequency data enable us to forecast stock variances and covariances more accurately. However, there are two problems on statistical analysis of high-frequency data. The first one is market microstructure noise: when we model stock prices by using diffusion processes, some empirical facts suggest the existence of additional noise. The second one is nonsynchronous observations: We observe stock prices when transactions occur. So observation times must be different for different stocks. In this paper, we study parametric inference under the existence of market microstructure noise and nonsynchronous observations. We study maximum-likelihood-type estimation for parametric diffusion processes with noisy, nonsynchronous observations, assuming that the true model is contained in the parametric family. We further study the case that this assumption is not satisfied. Such model is called a misspecified model. Ogihara [3] studied a parametric statistical model that a stochastic process Y_t is given by

$$\mathrm{d}Y_t = a(t, X_t)dt + b(t, X_t, \sigma_*)dW_t \tag{1.1}$$

for some unknown value $\sigma*$ of a parameter with noisy, nonsynchronous observations of Y. Maximum-likelihood- and Bayes-type estimators were constructed by using a quasi-likelihood function, and their asymptotic normality were shown. Asymptotic efficiency of the estimators was also proved by showing local asymptotic normality when the diffusion coefficients are deterministic and noises follow normal distributions. In this model, we assume that the true model is contained in the parametric family.

In practice for high-frequency data, to satisfy the assumption that the true model is contained in the parametric family, we need to choose the parametric family carefully so that it accurately captures microstructure of stock prices. This is a difficult task because several empirical facts of a stock market (intra-day seasonality, volatility clustering, complicated dependence structure of

stocks, and so on) make it difficult to capture the stock microstructure. On the other hand, high-frequency data contains huge information and therefore machine learning methods such as neural network or support vector machine is useful to identify the structure of the diffusion coefficient. In this approach, we need to consider a theory of misspecified model.

We will study asymptotic theory of a maximum-likelihood-type estimator for misspecified model. In this setting, the original maximum-likelihood-type estimator cannot attain the optimal convergence rate $n^{-1/4}$ due to the asymptotic bias. We construct a new estimator which attains the optimal rate by using a bias correction and show the asymptotic mixed normality.

2. Parametric estimation under misspecified settings

Let (Ω, F, P) be a probability space with a filtration $\boldsymbol{F} = \{F_t\}_{0 \le t \le T}$ for some $T > 0$. We consider a γ −dimensional $\boldsymbol{F}$ −adapted process $Y = \{Yt\}_{0 \le t \le T}$ satisfying an integral equation:

$$Y_t = Y_0 + \int_0^t \mu_s ds \int_0^t b_{s,\dagger}\, dW_s, \qquad t \in [0,T] \tag{2.1}$$

Where $\{W_t\}_{0 \le t \le T}$ *is* a γ_W − dimensional standard **F**-Wiener process, $\{\mu_t\}_{0 \le t \le T}$ and $b_\dagger = \{b_{t,\dagger}\}_{0 \le t \le T}$ are $\mathbb{R}^{\gamma}$–and $\mathbb{R}^{\gamma} \otimes \mathbb{R}^{\gamma W}$ -valued **F**-progressively measurable processes, respectively.

We assume that the observations of processes occur in a nonsynchronous manner and are contaminated by market microstructure noise, that is, we observe the vectors $\{\tilde{Y}_i^k\}_{0 \le i \le J_{k,n}}, k = 1,2$, where $\{\mathbf{J}_{k,n}\} 1 \le k \le \gamma, n \in N$ are positive integer-valued random variables, $\{S_i^{n,k}\}_{i=0}^{J_{k,n}}$ are random times, $\{\in_i^{n,k}\}_{i \in \mathbb{Z}_+, 1 \le k \le 2}$is an independent identical distributed random sequence, and

$$\tilde{Y}_i^k = Y_{S_i^{n,k}}^k + \in_i^{n,k}. \tag{2.2}$$

Let T denote the transpose operator for matrices (and vectors). We consider estimation of covariance matrix $\sum_{t,\dagger} = b_{t,\dagger} b_{t,\dagger}^T$ of the diffusion coefficient by using functional $\Sigma(t, X_t \sigma)$ with a γ_X − dimensional càdlàg stochastic process X_t and a parameter σ. We observe possibly noisy data: $\tilde{X}_j^l = X_{T_j^{n,l}}^l + \eta_j^{n,l}$ for $0 \le j \le \mathbf{K}_{l,n}$ and $1 \le l \le \gamma_X$, where $\{\mathbf{K}_{l,n}\}_{1 \le l \le \gamma X}, n \in N$ are positive integer-valued random variables, $\{T_j^{n,l}\}_{j=0}^{\mathbf{K}_{l,n}}$ are random times, and $\{\eta_j^{n,l}\}_{j=0}^{\mathbf{K}_{l,n}}$ are random variables possibly equal to zero.

We can arbitrarily set the explaining variable X_t and the function $\Sigma(t,x,\sigma)$. For example, we set other stock price processes, a price process of stock index, accumulated volume of stock trade, Y_t itself or some combination of those.

Let X_t be a càdlàg stochastic process and Σ(t,x,σ): [0,T]×O×clos(Λ) → $R^{\gamma} \otimes R^{\gamma}$ be some known continuous function, where O ⊂ R^{γ} X be an open set and the parameter space $\Lambda \subset \mathbb{R}^d$ be a bounded open set with $d \in \mathbb{N}$.

Let $\Pi_n = \left(\{S_i^{n,k}\}_{k,i}\{T_j^{n,l}\}_{l,j}\right)$, then we assume that $\mathcal{F}_T, (\Pi_n)_{n\in\mathbb{N}}$ and $\{\in_i^{n,k}\}_{n,k,i}$ are mutually independent. Let

$$\mathcal{G}_t = \mathcal{F}_t \vee \mathfrak{B}(\{\Pi_n\}_n) \vee \mathfrak{B}\left(A \cap \{S_i^{n,k} \le t\}; A \in \mathfrak{B}\{\in_i^{n,k}\}, k \in \{1,2\}, i \in \mathbb{Z}_+, n \in \mathbb{N}\right),$$

where $\mathcal{H}_1 \vee \mathcal{H}_2$ denotes the minimal σ-field which contains σ-fields $\mathcal{H}_1$ and $\mathcal{H}_2$. Moreover, we assume that $\eta_j^{n,l} 1_{\{T_j^{n,l} \le t\}}$ is $\mathcal{G}_t$-measurable, $E[\in_0^{n,k}] = 0$ and $E[(\in_0^{n,k})^2] = v_{k,*}$, where 1_A is the indicator function for a set A and $v_{k,*}$ is positive constant for $1 \le k \le \gamma$.

We consider a maximum-likelihood-type estimator of σ based on a quasi-likelihood function. Construction is based on that of Ogihara [3]. Let $\{b_n\}_{n\in\mathbb{N}}$ and $\{l_n\}_{n\in\mathbb{N}}$ be sequences of positive numbers satisfying $b_n \ge 1, \ell_n \in \mathbb{N}, b_n \to \infty, \ell_n / {b_n}^{-1/3-\epsilon} \to \infty$ and $b_n^{1/2-\epsilon}/\ell_n \to \infty$ as $n \to \infty$ for some $\epsilon > 0$.

By technical issues, we construct a quasi-log-likelihood function by dividing whole observation interval $[0,T]$ into disjoint local intervals $\{s_{m-1}, s_m)\}_{m=1}^{\ell_n}$. Here the sequence $\{b_n\} n \in N$ is the order of sampling frequency, that is,

$$0 < P\text{-}\lim_{n\to\infty} (b_n^{-1}\mathbf{J}_{k,n}) < \infty$$

Almost surely for $1 \le k \le \gamma$.

We prepare several notations: $k_n = b_n {\ell_n}^{-1}, K_0^j = -1, K_m^j = \#\{i \in \mathbb{N}; S_i^{n,j} < s_m\}, k_m^j = k_m^j - k_{m-1}^j - 1, I_{i,m}^k = [S_i^{n,k} + k_{m-1}^k, S_{i+1+k_{m-1}^k}^{n,k})$ and $M(l) = \{2\delta_{i_1,i_2} - 1_{\{|i_1,i_2|=1\}}\}_{i_1,i_2=1}^{l}$, where δ_{ij} is Kronecker's delta. For an interval $J = [a,b)$, we denote $|J| = b - a$. ε_l denotes a unit matrix of size l.

Let us consider an observable approximation $\hat{X}_m$ of $X_{s_{m-1}}$ defined by

$$\hat{X}_m = \left(\#\{j; T_j^{n,k} \epsilon [s_{m-1}, s_m)\}^{-1} \sum_{\mathrm{j};T_j^{n,k}\epsilon[\mathrm{s_{m-1}},\mathrm{s_m})} \tilde{X}_j^k \right)_{1\le k\le yx}.$$

Let $\sum_m(\sigma) = \sum(s_{m-1}, \hat{X}_{m-1}, \sigma), M_{j,m} = M(k_m^j), Z_{m,l}^i = \tilde{Y}_{l+1+K_{m-1}^i}^i - \tilde{Y}_{l+K_{m-1}^i}^i$ and $Z_m = \left(\left(Z_{m,l}^1\right)_l^T, \ldots, \left(Z_{m,l}^y\right)_l^T\right)^T$.

For a matrix A, we denote its (i,j)-element by $[A]_{ij}$. Then roughly speaking, we obtain approximation:

$$\mathrm{E}\left[Z_{m,i}^k Z_{m,j}^k | F_{\mathrm{s_{m-1}}}\right] \approx \left[\sum_{\mathrm{s_{m-1}}}, \dagger\right]_{kk} |\mathrm{I}_{i,m}^k| \delta_{ij} + v_{k,*}\left[M_{k,m}\right]_{ij},$$

$$\mathrm{E}\left[Z_{m,i}^k Z_{m,j}^l | F_{\mathrm{s_{m-1}}}\right] \approx \left[\sum_{\mathrm{s_{m-1}}}, \dagger\right]_{kl} |\mathrm{I}_{i,m}^k| \cap I_{j,m}^l|$$

for $k \neq l$. Therefore, by setting

$$S_m(B,v) = \begin{pmatrix} [B]_{11}\mathrm{diag}\left(\left(|I_{i,m}^1|\right)_i\right) & \cdots & [B]_{1\Upsilon}\left\{|I_{i,m}^1 \cap I_{j,m}^\Upsilon|\right\}_{ij} \\ \vdots & \ddots & \vdots \\ [B]_{\Upsilon 1}\left\{\left|I_{i,m}^1 \cap I_{j,m}^\gamma\right|\right\}_{ji} & \cdots & [B]_{\Upsilon\Upsilon}\mathrm{diag}\left(\left(|I_{i,m}^\Upsilon|\right)_i\right) \end{pmatrix} + \begin{pmatrix} v_1 M_{1,m} & & \\ & \ddots & \\ & & v_\gamma M_{\gamma,m} \end{pmatrix}$$

for a $\gamma \times \gamma$ matrix B and $v = \left(v_1, \ldots, v_\gamma\right)$, and $S_m(\sigma, v) = S_m(\sum_m(\sigma), v)$, we define a quasi-log-likelihood function

$$H_n(\sigma, v) = -\frac{1}{2}\sum_{m=2}^{l_n}\left(Z_m^T S_m(\sigma, v)^{-1} Z_m + \log\det S_m(\sigma, v)\right).$$

The function H_n contains two parameters: the first one is σ, which is the parameter for estimator of $\sum$ and is of our interest. The second parameter is v, which is the parameter for noise variance. Though we consider a simultaneous maximization of σ and v, we accept beforehand estimation of v so that we apply our results to the case of non-Gaussian noise.

[V] There exist estimators $\{\hat{v}_n\}_{n\in\mathbb{N}}$ of v_* such that $\hat{v}_n \geq 0$ almost surely and $\left\{b_n^{1/2}(\hat{v}_n - v_*)\right\}_{n\in\mathbb{N}}$ is tight.

Such $\hat{v}_n$ can be easily obtained. For example, let $\hat{v}_n = (\hat{v}_n^1, \ldots, \hat{v}_n^\gamma$ and

$$\hat{v}_{k,n} = \left(2\mathbf{J}_{k,n}\right)^{-1}\sum_{i,m}\left(Z_{i,m}^k\right)^2$$

Then $\hat{v}_n$ satisfies [V] if $\left\{b_n \mathbf{J}_{k,n}^{-1}\right\}_{n\in\mathbb{N}}$ is tight and $\sup_{n,k,i} E\left[\left(\in_i^{n,k}\right)^4\right] < \infty$.

We fix $\hat{v}_n$ which satisfies [V]. We define a maximum-likelihood-type estimator

$$\hat{\sigma}_n = \mathrm{argmax}_\sigma H_n(\sigma, \hat{v}_n).$$

H_n is constructed based on local Gaussian approximation of $Z_{m,i}^k$. This approximation seems valid only when observation noise $\in_i^{n,k}$ follows a normal

distribution. However, we can see in the proof that this approximation is also valid and $\hat{\sigma}_n$ still works for the case of non-Gaussian noise.

3. Asymptotic theory for misspecified model of diffusion cesses

Ogihara [3] studied specified model:

$$\Sigma_{t,\dagger} \equiv \Sigma(t, X_t, \sigma_*) \text{ for some nonrandom } \sigma_* \in \Lambda \qquad (3.1)$$

and showed asymptotic mixed normality of $\hat{\sigma}_n$ and local asymptotic normality when $(\mu_t)_t$ and $(b_{t,\dagger})_t$ are nonrandom. In machine learning theory including neural network, we usually consider a model which does not necessarily satisfy (3.1) (misspecified model).

In the study of a misspecified model, we sometimes face different asymptotics with a specified model.

For example, Uchida and Yoshida [5] studied ergodic diffusion $X = (X_t)_{t\geq 0}$ with observations $X_{kh_n}{}_{k=0}^{n}$, where $h_n \to 0, nh_n \to \infty$ and $nh_n^2 \to 0$ as $n \to \infty$. They showed that convergence rate of a maximum-likelihood-type estimator for parameter in diffusion coefficients is $\sqrt{nh_n}$, which is different from the rate $\sqrt{n}$ for the specified model.

In our case, we also face different phenomenon from the specified case. The maximum-likelihood-type estimator $\hat{\sigma}_n$ cannot attain optimal rate of convergence due to the existence of asymptotic bias. However, we can construct estimator which attains the optimal rate by modifying the asymptotic bias.

a. Consistency

In this section, we study results related to consistency of $\hat{\sigma}_n$. Since convergence of $\hat{\sigma}_n$ is not ensured, we characterize convergence by means of a function $D(\Sigma, \Sigma')$.

Here, we assume some conditions on the latent stochastic process X,Y and market microstructure noise $\in_i^{n,k}$. Let $E\Pi[\boldsymbol{X}] = E[\boldsymbol{X}|\{\prod_n\}_n]$ for a random variable **X** and $|A|^2 = \sum_{i,j}|A|_{ij}^2$ for a matrix A.

[A1] 1. There exists a locally bounded function $L(x, y)$ such that
$$|\Sigma(s, x, \sigma) - \Sigma(t, y, \sigma)| \leq L(x, y)(|t - s| + |y - x|)$$
for any $s, t \in [0, T], x, y \in \mathcal{O}$ and $\sigma \in \Lambda$.

2. $\sum(t, x, \sigma)$ is positive definite for any $(t, x, \sigma) \in [0, T] \times \mathcal{O} \times clos(\Lambda)$.
3. $\mu_t$is locally bounded, that is, there exists an increasing sequence $\{Tl\}_l$ of stopping times such that $lim_{l\to\infty}T_l = T$ almost surely and $\{\mu_{t\wedge T_l}\}_{0\leq t\leq T}$is bounded for each l.
4. $\sup_{n,k,i}\mathrm{E}\left[\left(\in_i^{n,k}\right)^q\right] < \infty$ for any $q > 0$ and $\sup_{0\leq s<t\leq T}(E[|X_t - X_s|^2]/|t - s|) < \infty$.

5. $P[\min_{1\le m\le l_n}\#\{j; T_j^{n,k}\in[s_{m-1},s_m)\}\ge 1]\to 1$ as $n\to\infty$ and $\{\ell_n \max_{m,k}(\#\{j; T_j\in[s_{m-1},s_m)\})^{-1} E_\Pi\left[\left|\sum_{j;T_j^{n,k}\in[s_{m-1},s_m)}\eta_j^{n,k}\right|^2\right]\}_n$ is tight.
6. There exist progressively measurable processes $\{b_t^{(j)}\}_{0\le t\le T,0\le j\le 1}$ and $\{\hat{b}_t^{(j)}\}_{0\le t\le T,0\le j\le 1}$ such that $\sup_t E\left[|b_t^{(j)}|^q \vee |\hat{b}_t^{(j)}|^q\right]<\infty$,

$$\sup_{s<t} E[|b_t^{(j)}-b_s^{(j)}|^q\vee|\hat{b}_t^{(j)}-\hat{b}_s^{(j)}|^q]^{1/q}|t-s|^{-1/2}<\infty$$

for $0\le j\le 1$ and any $q>0$, and

$$b_{t,\dagger}=b_{0,\dagger}+\int_0^t b_s^{(0)}ds+\int_0^t b_s^{(1)}dWs,\quad b_t^{(1)}=b_0^{(1)}+\int_0^t\hat{b}_s^{(0)}ds+\int_0^t\hat{b}_s^{(1)}dWs$$

for $t\in[0,T]$.

Some of these conditions are standard conditions in this field and easy to check. Related to market microstructure noise $\eta_j^{n,k}$ for X, point 5 of (A1) is required. Roughly speaking, this condition is satisfied if the summation of $\eta_j^{n,k}$ is of an order equivalent to the square root of the number of $\eta_j^{n,k}$. This is satisfied if sampling frequency of $\{T_j^{n,k}\}$ is of order b_n and $\eta_j^{n,k}$ satisfies certain independency, martingale conditions or mixing conditions. Decomposition of X in point 6 of (A1) is used when we estimate the difference of $E_m[Z_mZ_m^T]$ and $\sum_{s_{m-1},\dagger}$ which appear in an asymptotic representation of H_n.

We further assume conditions for the sampling scheme. Let $r_n=\max_{i,k}|S_i^{n,k}-S_{i-1}^{n,k}|$ and $\underline{r_n}=\min_{i,k}|S_i^{n,k}-S_{i-1}^{n,k}|$. For $\eta\in(0,1/2)$, let S_η be the set of all sequences $\{[s'_{n,l},s''_{n,l})\}_{n\in\mathbb{N},1\le l\le L_n}$ of invervals on $[0,T]$ satisfying $L_{n\,n\in\mathbb{N}}\subset\mathbb{N}$, $[s'_{n,l_1},s''_{n,l_1})\cap[s'_{n,l_2},s''_{n,l_2})=\emptyset$ for $n,l_1\ne l_2$, $\inf_{n,l}\left(b_n^{1-\eta}(s'_{n,l}-s''_{n,l})\right)>0$ and $\sup_{n,l}\left(b_n^{1-\eta}(s'_{n,l}-s''_{n,l})\right)>\infty$.

[A2] There exist $\eta\in(0,1/2), k>0, \dot{\eta}\epsilon(0,1]$ and positive-valued stochastic processes $\{a_t^j\}t\in[0,T], j=1,2$ such that $\sup_{t\ne s}(|a_t^j-a_s^j|/|t-s|^{\dot{\eta}})<\infty$ almost surely, $b_n^{-1/2+k}k_n(b_n^{-1}k_n)^{\dot{\eta}}\to 0$ and

$$k_nb_n^{-1/2+k}\max_{1\le l\le L_n}\left|b_n^{1-\eta}(s''_{n,l}-s'_{n,l})^{-1}\#\{i;[s_{i-1}^{n,j},s_i^{n,j})\subset[s'_{n,l},s''_{n,l})\}-a^j_{s'_{n,l}}\right| \quad (3.2)$$

Converges to zero in probability as $n\to\infty$ for $\{[s'_{n,l},s''_{n,l})\}_{1\le l\le L_n,n\in\mathbb{N}}\in S_\eta$ and $j=1,2$. Moreover, $(k_nb_n^{1-\epsilon})\vee(b_n^{-1-\epsilon}\underline{r_n}^{-1})\to^p 0$ for any $\epsilon>0$.

Condition [A2] is about the law of large numbers of $s_l^{n,k}$ in each local interval $[s'_{n,l},s''_{n,l})$. This conditions ensures that the intensity of observation count converges to some intensity a_t^j with the order $(k_nb_n^{-1/2+k})^{-1}$.

Condition [A2] holds if observation times are generated by mixing processes as seen in the following example.

Example 3.1. *Let* $\{N_t^k\}_{t\geq 0}$ *be an exponential* α*-mixing point process with stationary increments for* $k = 1,2$*. Assume that* $E\left[\left|N_1^k\right|^q\right] < \infty$ *for any* $q > 0$ *and* $k = 1,2$*. Set* $S_i^{n,k} = \inf\{t \geq 0; N_{b_n t}^k \geq i\}$*. Then [A2] is satisfied with* $a_t^j \equiv [N_1^j]$ *(constants) by Rosenthal-type inequalities (Theorem 4 in [4]) and a similar argument to the proof of Proposition 6 in [4]. Also, [B2] (defined later) is satisfied if further* $k_n b_n^{-3/5+\epsilon} \to 0$ *for some* $\epsilon > 0$*.*

We study results related to consistency under Conditions [A1] and [A2]. Let $\tilde{a}_t = (\tilde{a}_t^1, \ldots, \tilde{a}_t^\gamma), \tilde{a}_t^j = a_t^j / v_{j,*}$ for $1 \leq j \leq \gamma, \sum_\dagger = (\sum_{t,\dagger})0 \leq t \leq T, \sum_t(\sigma) = \Sigma(t, X_t, \sigma), \Sigma(\sigma) = (\sum_t(\sigma))_{0\leq t\leq T}$, and $D(t, A) = ([A_t]_{ij}\sqrt{\tilde{a}_t^i \tilde{a}_t^j})_{1\leq i,j\leq\gamma}$ for $A = (A_t)_{t\geq 0}$: $\gamma \times \gamma$ matrix valued.

We define

$$D(\sum(\sigma), \sum_\dagger) = \int_0^T \left\{\frac{1}{4} tr\left((\mathcal{D}(t, \sum_\dagger) - \mathcal{D}(t, \sum(\sigma)), \mathcal{D}(t, \sum(\sigma))^{-1/2}\right) - \frac{1}{2} tr(\mathcal{D}(t, \sum_\dagger)^{1/2}) + \frac{1}{2} tr(\mathcal{D}(t, \sum(\sigma))^{1/2})\right\} dt$$

$$= \frac{1}{4}\int_0^T tr\left(\left(\mathcal{D}(t, \sum_\dagger)^{1/2} - \mathcal{D}(t, \sum(\sigma))^{1/2}\right)^2 \mathcal{D}(t, \sum(\sigma))^{-1/2}\right) dt.$$

Theorem 3.1. *Assume [A1], [A2] and [V]. Then*

$$D(\Sigma(\hat{\sigma}_n), \sum_\dagger) \to^p \min_\sigma D(\sum(\sigma), \sum_\dagger)$$

as $n \to \infty$.

Under boundedness of $a_t^j + (a_t^j)^{-1}$ and $\|\Sigma(\sigma)^{-1}\|$, there exist positive constants C_1 and C_2 such that

$$C_1 \int_0^T \left|\sum(t, X_t, \sigma) - \sum_{t,\dagger}\right|^2 dt \leq D\left(\sum(\sigma), \sum_\dagger\right) \leq C_2 \int_0^T \left|\sum(t, X_t, \sigma) - \sum_{t,\dagger}\right|^2 dt \quad (3.3)$$

where C_1 and C_2 depend only the upperbounds of $a_t^j + (a_t^j)^{-1}$ and $\|\Sigma(\sigma)^{-1}\|$. The proof is left in the appendix. Then we can say that D is equivalent to L^2 norm.

Remark 3.1. *In the specified setting of Ogihara [3]* $,\gamma_1(\sigma) = D\left(\sum(\sigma), \sum(\sigma_*)\right)$ *holds for* $\gamma_1(\sigma)$ *in Section 2.2 of [3]. The presentation 2.8 of* $,\gamma_1(\sigma)$ *is obtained by calculating elements of* $D\left(\sum(\sigma)\right)^{1/2}$ *for* $\gamma = 2$.

3.2 Optimal rate convergence

In this section, we study optimal rate convergence. Ogihara [3] showed that local asymptotic normality holds for the specified model with nonrandom diffusion coefficients, the optimal rate of convergence is equal to $b_n^{1/4}$ for

estimators of the parameter in the diffusion coefficients, and the maximum-likelihoodtype estimator $\hat{\sigma}_n$ attains the optimal rate. On the other hand, we will see that $\hat{\sigma}_n$ cannot attain the rate $b_n^{1/4}$ in the misspecified setting due to an asymptotic bias term. We can attain optimal rate if we construct a maximum-likelihood-type estimator $\hat{\sigma}_n$ by using a bias-modified quasi-log-likelihood function.

For a vector $x = (x_1, \cdots, x_k)$ we denote $\partial_x^l = \left(\frac{\partial^l}{\partial_{x_{i_1}} \cdots \partial_{x_{i_l}}}\right)_{i_1,\cdots,i_1=1}^{k}$. We assume that $\Lambda \subset \mathbb{R}^d$ satisfies Sobolev's inequality; that is, for any $p > d$, there exists $C > 0$ such that $\sup_{\sigma\in\Lambda}|\mu(\sigma)| \leq C\sum_{K=0,1}\left(\int_\Lambda |\partial_\sigma^k \mu(\sigma)|^p d\sigma\right)^{1/P}$ for any $u \in C^1(\Lambda)$. This is the case when Λ has a Lipshitz boundary. See Adams and Fournier [1] for more details.

To obtain optimal rate convergence, we need strengthened versions of the assumptions [A1] and [A2].

[B1] [A1] is satisfied, $\sup_{0\leq s<t\leq T}(E\,[E[X_t - X_s|\mathcal{F}_s]^2]/|t-s|^2) < \infty$, and $\partial_x\sum$ and $\partial_\sigma\sum$ exist and are continous on $[0,T] \times \mathcal{O} \times clos(\Lambda)$. Moreover, there exists a locally bounded function $L(x,y)$ such that

$$|\partial_x\Sigma(t,x,\sigma) - \partial_x\Sigma(t,y,\sigma)| \leq L(x,y)|x-y|.$$

for any $t \in [0,T], x, y \in \mathcal{O}$ and $\sigma \in \Lambda$.

[B2] There exists positive-valued stochastic processes $\{a_t^j\}_{t\in[0,T],1\leq j\leq\gamma}$ such that for any $q > 0$ and

$\epsilon > 0, (r_n b_n^{1-\epsilon}) \bigvee (b_n^{-1-\epsilon} r_n^{-1}) \bigvee \left(k_n b_n^{-3/5+\epsilon}\right) \to^p 0$,

$E\left[\sup_{t\neq s}\left(\left|a_t^j - a_s^j\right|^q/|t-s|^q\right)\right] < \infty, E\left[\sup_{j,t}\left(\left|a_t^j\right| + 1/\left|a_t^j\right|^q\right)\right] < \infty$

and

$$\sup_n \sup_{[s_n',s_n'')} E\left[\left(\sqrt{b_n(s_n''-s_n')}\left(\frac{\#\{i;[S_{i-1}^{n,j}, S_i^{n,j}) \subset [s_n', s_n'')\}}{b_n(s_n''-s_n')} - a_{s_n'}^j\right)\right)^q\right]$$

is finite for any $1 \leq j \leq \gamma$, where the second supremum is taken over all sequences $\{s_n'\}_n, \{s_n''\}_n \subset [0,T]$ such that $s_n' < s_n''$ and $\sup_n(\ell_n(s_n'' - s_n')) < \infty$.

We can see that [A2] holds under [B2].

Here, we see the bias of the quasi-log-likelihood function H_n.

Let $A(a) = diag\left(\left(a_j^{1/2}\right)_j\right)$. For a 2 × 2 positive definite matrix $B = (B_{ij})_{ij}$ and a 2 × 2 symmetric matrix $C = (C_{ij})_{ij}$, set

$$E_m(a,B,C,v) = tr\left(S_m(B,v)^{-1}S_m(C,v)\right) - (1/2)Tb_n^{1/2}\ell_n^{-1}tr\left(A(a)(C-B)A(a)(A(a)BA(a))^{-1/2}\right),$$

$$F_m(a,B,v) = \log\det S_m(B,v) - Tb_n^{1/2}\ell_n^{-1}tr\left(\left(A(a)BA(a)\right)^{-1/2}\right),$$
$$G_m(a_1,a_2,B,C,v) = E_m(a_1,a_2,B,C,v) + E_m(a_1,a_2,B,v).$$

Let $\widetilde{\sum}_m = \widetilde{\sum}_m(\sigma) = \sum(s_{m-1}, X_{s_{m-1}}, \sigma), \widetilde{\sum}_{m,\dagger} = \sum_{s_{m-1},\dagger}$ and $\tilde{S}_{m,\dagger} = S_m(\widetilde{\sum}_{m,\dagger}, v_*)$. We denote

$$\triangle_n = (\sigma_1,\sigma_2) := -\frac{1}{2}\sum_m \mathrm{tr}\left(\left(\tilde{S}_m^{-1}(\sigma_1) - \tilde{S}_m^{-1}(\sigma_2)\right)\left(\tilde{Z}_m\tilde{Z}_m^T - \tilde{S}_{m,\dagger}\right)\right)$$

for $\sigma_1, \sigma_2 \in \mathrm{clos}\,(\Lambda)$.

Proposition 3.1. *Assume [B1], [B2] and [V]. Let* $\{\sigma_{j,n}\}_j = 1,2, n \in \mathbb{N}$ *be* clos(Λ)*-valued random variables. Then*

$$\begin{aligned} b_n^{-1/4}&\left(H_n(\sigma_{1,n},\hat{v}_n) - H_n(\sigma_{2,n},\hat{v}_n)\right) \\ &= b_n^{-1/4}\,\triangle_n\left(\sigma_{1,n},\sigma_{2,n}\right) - b_n^{1/4}\left(D\left(\sum(\sigma_{1,n}),\sum_\dagger\right) - D\left(\sum(\sigma_{2,n}),\sum_\dagger\right)\right) \\ &+ \frac{1}{2}b_n^{-1/4}\sum_{j=1}^{2}(-1)^j\sum_m G_m\left(\tilde{a}_{s_{m-1}}, \widetilde{\sum}_m(\sigma_{j,n}), \widetilde{\sum}_{m,\dagger}, v_*\right) + o_p(1). \end{aligned} \quad (3.4)$$

The third term in the right-hand side of (3.4) is bias term which does not appear in the specified model. Due to the third term, we cannot ensure that $b_n^{-1/4}(\hat{\sigma}_n - \sigma_*)$ is $o_p(1)$.

In the following, we consider removing the bias. First, we consider an estimator $\left(B_{m,n}\right)$ of $\sum_{s_{m-1},\dagger}$ by using the function g appearing in Jacod et al. [2], that is, $g : [0,1] \to \mathbb{R}$ is continuous, piecewise $C^1, g(0) = g(1) = 0$, and $\int_0^1 g(x)dx > 0$. Let $g_l^j = g\left(l/\left(k_m^j + 1\right)\right), \Psi_1 = \int_0^1 g(x)^2dx, \Psi_1 = \int_0^1 g'(x)^2dx$. For example, let $g(x) = x\Lambda(1-x)$ on$0 \le x \le 1$, then $\Psi_1 = 1/12$ and $\Psi_2 = 1$. Let $\hat{a}_m = \left(\hat{a}_m^1, \cdots, \hat{a}_m^\gamma\right), \hat{a}_m^i = k_m^i\hat{v}_{i,n}^{-1}(Tk_n)^{-1}1_{\{\hat{v}_{i,n}>0\}}$, and let $B_{m,n}$ be a $\gamma \times \gamma$ matrix satisfying.

$$\left[B_{m,n}\right]_{ij} = \frac{\ell_n}{\mathrm{T}\Psi_1}\left\{\left(\sum_{l=1}^{k_m^i} g_l^i Z_{m,l}^i\right)\left(\sum_{l=1}^{k_m^j} g_l^j Z_{m,l}^j\right) - \frac{\hat{v}_{i,n}}{k_m^i}\Psi_2 1_{\{i=j\}}\right\}.$$

Then we define a bias-corrected quasi-likelihood function $\breve{H}_n(\sigma)$ by

$$\breve{H}_n(\sigma) = H_n(\sigma,\hat{v}_n) + \frac{1}{2}\sum_m G_m\left(\hat{a}_m, \sum_m(\sigma), B_{m,n}, \hat{v}_n\right).$$

By setting $\hat{D}_{m,n,\dagger} = \left(\left(\hat{a}_m^i\hat{a}_m^j\right)^{1/2}\left[B_{m,n}\right]_{ij}\right)_{ij}$ and $\hat{D}_{m,n} = \left(\left(\hat{a}_m^i\hat{a}_m^j\right)^{1/2}\left[\sum_m(\sigma)\right]_{ij}\right)_{ij}$, $\breve{H}_n(\sigma)$ is simplified as

$$\begin{aligned} \breve{H}_n(\sigma) = -\sum_m\Bigg\{&\frac{1}{2}tr\left(S_m^{-1}(\sigma)\left(Z_mZ_m^T - B_{m,n}\right)\right) \\ &+ \frac{\sqrt{T}}{4}\ell_n^{-1/2}tr\left(\left(\hat{D}_{m,n}(\sigma) + \hat{D}_{m,n},\dagger\right)\hat{D}_{m,n}(\sigma)^{-1/2}\right)\Bigg\}. \end{aligned} \quad (3.5)$$

Let $\breve{\sigma}_n = \mathrm{argmax}_\sigma\breve{H}_n(\sigma)$. Then we obtain optimal rate convergence for $\breve{\sigma}_n$.

Theorem 3.2. *Assume [B1], [B2] and [V].Then* $\left\{\mathrm{b}_n^{1/4}\left(D\left(\sum(\check{\sigma}_n),\sum_{\dagger}\right)-\min_{\sigma}D\left(\sum(\sigma),\sum_{\dagger}\right)\right)\right\}_{n\in\mathbb{N}}$ *is tight.*

References

1. R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.
2. J. Jacod, Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter. Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic Process. Appl.*, 119(7):2249{2276, 2009.
3. T. Ogihara. Parametric inference for nonsynchronously observed diffusion processes in the presence of market microstructure noise. *Bernoulli*, 24(4B):3318{3383, 2018.
4. T. Ogihara and N. Yoshida. Quasi-likelihood analysis for nonsynchronously observed diffusion processes. *Stochastic Process. Appl.*, 124(9):2954{3008, 2014.
5. M. Uchida and N. Yoshida. Estimation for misspeci_ed ergodic diffusion processes from discrete observations. *ESAIM Probab. Stat.*, 15:270{290, 2011.

The servicification of manufacturing in Asia: Redefining the sources of labor productivity using time

Valerie Mercer-Blackman[1], Christine Ablaza[2]
[1]Asian Development Bank
[2]University of Queensland

Abstract

Current measures of productivity using national accounts do not properly assess the indirect contribution of services to other sectors. For example, the additional value generated by services in the production of manufactured goods can be substantial, but it is not properly accounted for because of the indivisible and intangible nature of services. We propose a conceptual framework of servicification, offer some preliminary evidence of this phenomenon in Asia using an input-output framework and propose the adoption of time-use surveys to measure productivity when the distinction between labour and leisure disappears in the digital age.

Keywords

Servicification; Premature Deindustrialization; Labor Productivity; Input - output tables; Time-use surveys

1. Introduction, definitions and the measurement problem

Services are becoming increasingly prominent in terms of both output and employment in developing Asian countries. The fact that this shift is occurring even as manufacturing has yet to fully develop has prompted some to call the deindustrialization "premature" (Rodrik 2016). This hypothesis, based on the premise that the manufacturing sector is more productive than services and thus the driver of growth, has incorrectly raised concerns about the role of services in development. This paper shows that the sectoral approach to measuring output ignores the increasing fragmentation of production wherein tasks may be outsourced to other sectors domestically or internationally, so the contribution of many services to the manufacturing process are not properly captured. We show that current productivity measures suffer from biases in definition and measurement. as it is difficult to measure their contribution given the indivisible, intangible nature of services. At the same time, the contribution of services is becoming even more important in a knowledge-based, digital economy. This paper measures the extent of "servicification" in manufacturing in Asia and globally; shows why measures using national accounts do not capture this phenomenon and proposes a method using the principle of time-use.

Services encompass a wide range of activities that fall outside of agriculture, manufacturing, or other industries (Andersen and Corley 2003).

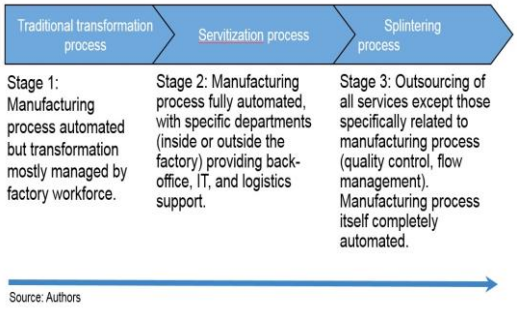
Difficulty in measuring them implies that they are treated as a residual after subtracting the value of agriculture and industry from GDP. Given their heterogeneous nature, services have been classified by literature in different ways. One of the most commonly used classification systems is based on the primary product of a firm or an enterprise: traditional or "stagnant" services and modern, hi-tech, or "progressive" services. Another classification is based on how services are used or consumed. In terms of their role in manufacturing, services can be considered horizontal or vertical, and can be supplied either for domestic or foreign use

Current productivity estimates of services may be biased due to measurement issues. Simply defined, productivity is the amount of real output produced by a given set of real inputs. This implies that the quantity of output and inputs, as well as the prices used to deflate both components, must be captured accurately. This is difficult to do in practice given the intangibility, indivisibility and value-creation nature inherent in services. As a result, the output of manufacturing would appear larger and its productivity higher relative to services. Given the increasingly important role played by services in the manufacturing process, the bias could be significant. Moreover, intersectoral comparisons overlook the indirect contribution of services to the productivity of other sectors. When specialization occurs, the resulting economies of scale not only translate to greater output for manufacturing firms, but also to lower prices for services that are used as inputs into production, and this is missed because the linkages are not caputured.

Some definitions are in order. Servicification can come in two forms: when services are bundled into a manufacturing good (servitization, as in Miroudot and Cadestin, 2017), or through splintering (servification). The splintering of production could manifest itself as manufacturing firms closing their services departments and outsourcing. Essentially, this allows businesses to subcontract part of their operations to independent suppliers located either domestically or abroad (e.g., offshore outsourcing). The gradual transformation of the manufacturing process to a service-oriented one is what creates the symbiotic relationship between goods and services. Figure 1 shows the various stages of servicification.

There are variants of this relationship. An example is the movement from producing songs on compact disks to making them available digitally: The music industry is still alive but is reclassified from producing a"good" to producing a "service." Another example is when an auto company separates its auto maintenance and leasing business. Each unit can act separately, but the efficiency and survival of the service and leasing units depend on the extent of sales of that type of car (subordination of the service to the manufacturing process

Figure 1: The Servification of Manufacturing



When services are bundled with goods sold by manufacturing firms, the measure of their contribution to value added gets complicated. In theory, national accounts should reflect the division between in-house goods and services of servicification. For example, the total output of a manufacturing firm that offers financing should be recorded as two separate transactions: the first as a good, and the second as a service. In practice, there are likely to be differences across countries and industries regarding how output is measured and recorded by national statistical offices. It is especially difficult to disaggregate output when the sale is conducted as a single transaction or when the service is not "consumed" simultaneously with the good (e.g., maintenance or repair). As a result, the servitization of manufacturing output is likely to be understated in national accounts. Crozet and Milet (2017) thus refer to this phenomenon as "hidden deindustrialization." Box 1 discusses the main challenges for national accounts.

Box 1: CAN NATIONAL ACCOUNTS ADEQUATELY MEASURE TODAY'S PRODUCTIVITY?

National accounts were conceived during a very different time to today. Colin Clark, Simon Kuznets, and Richard Stone of the UK began to conceive the national accounts in the 1930s. This was a time when manufacturing and construction were the engines of growth: The production of goods was a clearly tangible process of man "working" with machines or tools to transform mostly physical goods into consumable outputs. Services were sometimes supportive (for example, transportation utilities, etc.). But services were mostly consumable or publicly provided; they were considered marginal to production or a leisure activity. To draw a line as to what should be value added in a given year, Kuznets chose to consider only "productive"

activities in the new economic statistics, defined as those that produced goods or services that could be bought or sold in the market economy. Thus, their unit value was the price. It was also important to be able to measure the activities, and the industrial classification developed thus "treat[ed] services as 'immaterial' (i.e., everything that is not manufacturing or agriculture), while ignoring that the activity of services in the economy, as well as the corporate structure of firms, transcend such classification schemes at any level of aggregation" (Andersen and Corley 2003). However, as the definition of a "productive activity" or the definition of a "unit" of service became increasingly blurred in an age of hi-tech manufacturing, artificial intelligence, and apps that run on cellphones, the compilation of national accounts as originally conceived is experiencing serious challenges measuring intangibles

The example of the national accounting of the Korean automobile company Hyundai illustrates the difficulty of measuring every process. Its factory in Montgomery, Alabama, in the US, can produce almost 400,000 cars and trucks per year with 3,000 employees for distribution across North America, but the company also leases them and finances their purchase, and its 800 dealerships provide the servicing. Almost all the parts, including sophisticated electronic components and sensors, are produced elsewhere around the world. How will national accountants put together Hyundai's economic contribution? They will request revenues and costs for all the operations of Hyundai, and will divide the main activities (primary, secondary, tertiary activities, etc.). At some reasonable cutoff for the number of "principal" activities of Hyundai US, they will assign the value added (revenues minus costs) to the different subsector categories under the NAICS classification. Leasing, repair, engineering services, logistics, accounting, etc., that are services contracted out to third parties will be considered a cost for Hyundai and a revenue for the service providers, so the input-output links will clearly show that these activities are linked (although presentation of national accounts on the production side will not show the links). Labor compensation will be classified depending on the worker's place of affiliation. Overall, the value added will be in aggregate fully accounted for in one sector or another. There are four important problems in this measurement that lead to undermining the contribution of services to productivity.

- **First, the labor productivity of that plant (number of cars per worker/hour, 400,000 autos per year/300 plant workers in man-hours) will be solely attributed to the auto manufacturing sector in the national accounts, and not to the myriad of services that contributed**. Due to bundling of services, batching of computer programming, robotics installed in earlier years, etc., most of the unit average costs in that period will be components, parts and utilities, consumption of fixed capital defined under statutory depreciation

rules, etc. The contribution of services in that period will be small. Nonetheless, the output would be impossible without the provision of "indivisible" services with huge economies of scale. Their contribution cannot be accounted for as a share of the final output if it is included elsewhere as a stand-alone "service."

- **Second, workers involved in services within the Hyundai plant are unlikely to even appear as separate services employees in the accounting**, and thus cannot contribute to an increase in services value added in the national accounts, particularly if they comprise a small or ancillary cost of production. The extent of servitization mismeasurement is greater when the service is provided in-house (Crozet and Milet 2017).
- **Third, services are typically priced through bundling, cognizant of their indivisibility property**. It is common for insurance, accounting services, and TV and phone services to be priced as "monthly services," which means that two users of the same plan may use vastly different amounts of the "bundled" service. In practice, the difference in productivity derived from the service by each user may be huge. For example, national accounts will show the "phone services" of the customer service desk as being equal to the "phone services" of the staff lounge room and attach to it the bundle price, erroneously attributing the same value added to these two users. When deflated, services with different usage rates are assumed to be equally productive. This is compounded by the lack of homogeneity of the service unit once used up. Goods, on the other hand, are tangible and clearly divisible, so their unit value can be more easily measured.
- **Finally, there are many services that are becoming almost free because they rely on a repetitive code that has already been designed from before, as we rely on an accumulation of knowledge by others**: for example, an algorithm designed to optimize the shipping routes for Hyundai cars ready for delivery. This has a fixed cost (charged by the programmers), but no marginal cost. Again, national accounts may attribute the efficiency of the distribution process to the manufacturing process itself, when in fact it was the infinite economies of scale of the network specialist's algorithm that enabled this shipping efficiency

Other issues discussed in the literature exacerbate these problems. However, various studies (for example, IMF 2018) argue that the size of the estimated effects is insufficient to explain the fall in labor productivity over the last two decades. Going forward, these issues will lead to large measurement biases. There are perhaps five main issues that arise: (i) deflators of new goods or hi-tech goods do not reflect goods' "unit value" when calculating real GDPs,

and not all statistical offices adjust appropriately; (ii) "free" goods, such as Facebook, Wikipedia, pictures from a phone, etc., are not included in national accounts (because their price is zero), thereby underestimating the value they contribute to GDP. If these platforms are used for e-commerce, for example (which is very common in developing Asia), their contribution to efficient distribution is not properly accounted for; (iii) goods or services produced but not remunerated (unpaid household work, family help) are also not included because they are free; (iv) when corporations splinter production offshore, the valuation of each of the stages of production sometimes relies on inaccurate pricing by multinational companies, who declare their ownership of each stage of production in the locality that minimizes their tax liability (transfer pricing). Even if all production stages could be accurately valued, it would require all countries providing full, accurate reporting and sharing their data on companies with other national accounts statistics offices, which is beyond the capacity of most countries' institutions (Moulton and van de Ven 2018); and (v) the spillover effects from agglomeration economies of a talented team working together to produce new knowledge is crucial to productivity and generally not accounted for. The human capital of a university scientist in the team, for example, is classified as an "education" service. Such a service is valued at cost—sometimes subsidized if provided by the public sector—because there is no tangible output.

2. Methodology and evidence using input-output data

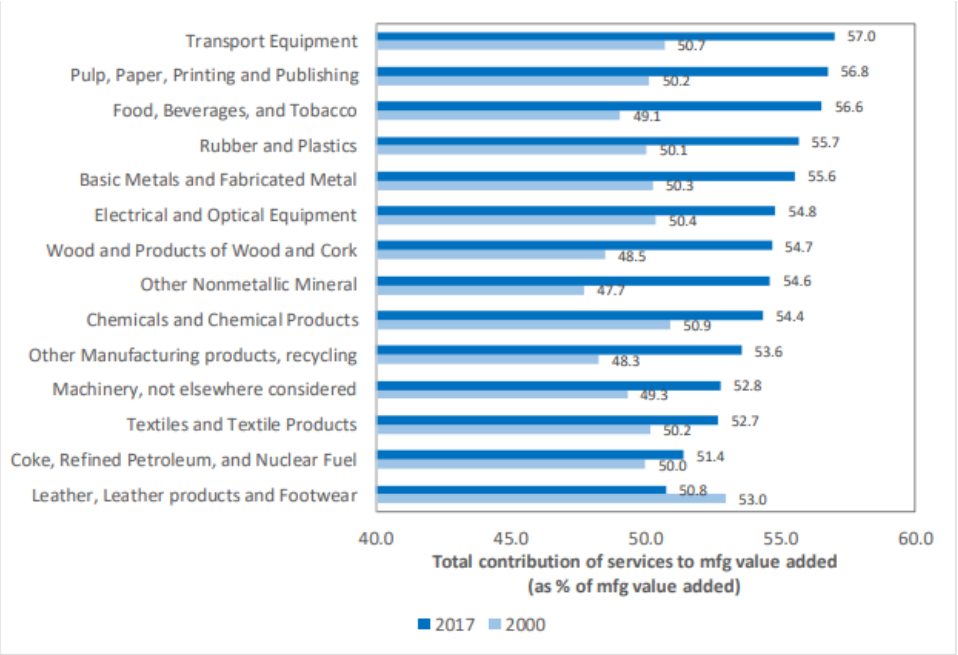
The ADB Multi-Regional Input-Output Table (MRIOT) allow us to measure sector-level components of servicification using some refinements on the well-known direct and Leontief coefficients. Using the technical coefficient matrix, we can quantify the number of services that are directly used as inputs in manufacturing sectors for arm's-length transactions. By subtracting this matrix from the Leontief matrix, we also obtain an estimate of services that are indirectly used by a particular sector (see ADB (2018) for a detailed description of the decomposition). The Leontief coefficients themselves give us the total number of services used in manufacturing. In other words, they represent the sum of what we denote as direct and indirect components. To illustrate these concepts, consider the case of an automobile manufacturer. To produce one vehicle, it uses equipment leased by another company. The rent paid for the equipment is an example of a direct service used as an input by the automobile manufacturer. However, this does not account for all the equipment rentals that are paid in the production of one vehicle. For instance, the automobile manufacturer may require basic metals as part of its raw materials. Assuming these metals are also produced using leased equipment, then the rent serves as an indirect input to the manufacture of a vehicle. Figure 8 shows that the direct contribution of services to manufacturing's value added between 2000

and 2017 stayed broadly constant: On average, a dollar of demand for manufacturing production generates nearly $0.20 of services globally. However, the indirect component is not only about twice as large but has grown by more than 15%: the total (direct and indirect) contribution of services to a $1 dollar value added in manufacturing increased from $0.55 in 2000 to $0.62 in 2017.

The contribution of services to exports is also high. We find that although services exports as a share of total exports have stayed around 17% between 2000 and 2017 in Asia, the contribution of services to exports has grown by 25% during the same period, and constitute about 34% of total exports I value added terms. Moreover, most of the services are domestically-procured, and come from sectors that are not exported such as finance, wholesale trade and transportation. All countries show a substantial contribution: for every $1 of manufacturing value added demand in developing Asia, about $0.16 comes from the direct contribution of services and $.26 comes from services' indirect contribution. This compares to $0.21 and $0.38 in OECD countries, respectively.

Globally, all manufacturing sectors show that services contribute to between 50% and 60% of their value added, and the phenomenon is not only limited to high-tech manufacturing sectors. Figure 2 shows the direct and indirect contribution of services to the value added of each manufacturing sector globally. Transportation equipment, which is deeply embedded in global value chains, is not only the most servified manufacturing sector (at 57% in 2017), but this contribution grew the most of all manufacturing sectors between 2000 and 2017. This is not surprising: transport equipment, particularly autos, is also one of the most automated sectors (using robots). Other sectors, such as paper printing and publishing, as well as food, beverages, and tobacco, tend to be mostly nontraded, and are directly linked to services such as publishing and restaurants, respectively. Only one sector, leather and footwear, became less servified between 2000 and 2017, although services still contribute to 50% of their value added.

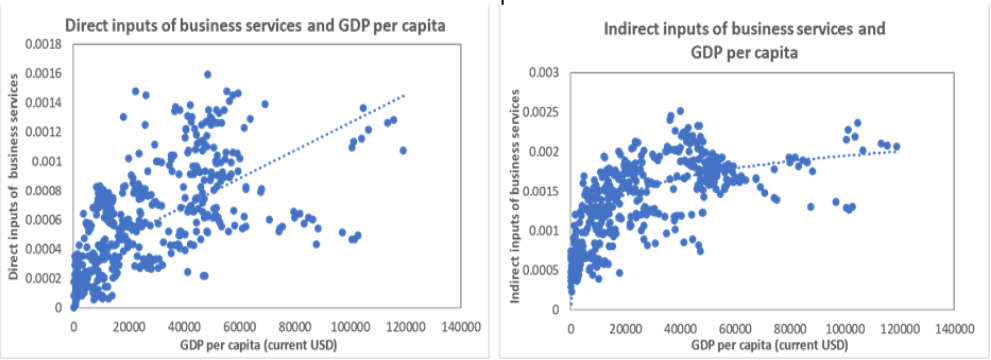
Figure 2: Total (direct and indirect) contribution of services to manufacturing value added by sector (2000 and 2017)



Source: Authors' calculations based on ADB MRIOTs. Note: Figures represent the average input coefficient of services for all 62 countries. The original data is expressed in terms of one dollar of manufacturing output.

We also examined "servification of services" and found that business services tend to be a key player in development for high-income economies, despite being barely traded internationally. Since the majority of services are not directly exported, it is easy to undervalue their importance in the growth of manufacturing and an export-led development strategy. When high-tech manufacturing products are produced, this tends to stimulate business services (which include legal and professional services). Indeed, the greater the direct and indirect linkages (servification) of business services in manufacturing value added, the more developed the economy is (Figure 3). This ratio is generally low for most of Asia except for Singapore and Hong Kong, China. Interestingly, both direct and indirect linkages increase quickly in the early stages of development. Indirect linkages are highly correlated with development, particularly for advanced economies (Figure 3, right panel). Contrary to interpretations by Rodrik (2016), it implies that the barometer for the speed of economic development may no longer be to increase the share of employment in manufacturing, but instead the degree of links (servicification) between hi-tech services such as business services and manufacturing value added.

Figure 3: Impact of services sectors on manufacturing and business services and economic development



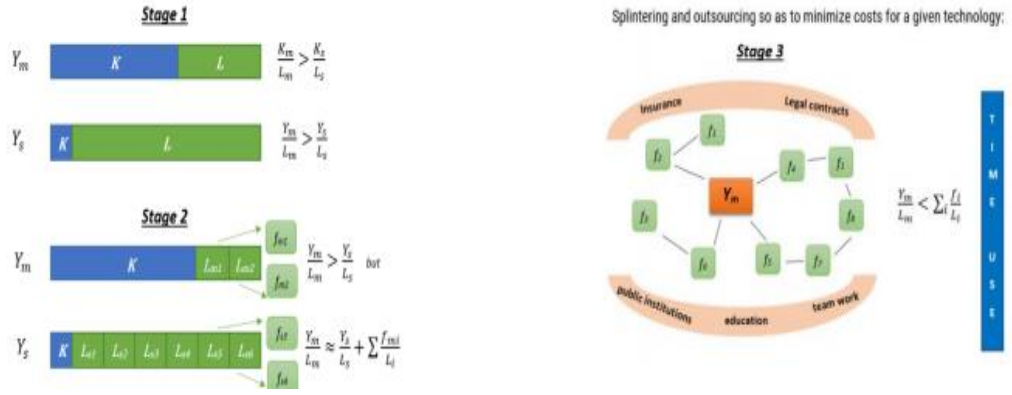
3. Results, implications for measurement and the value of time-use measures

If there is servicification, how does that improve productivity? The empirical evidence linking servicification of the economy to productivity growth and economic development is still quite limited. Traditional Solow-model approaches estimate a distinct decline in labor productivity in OECD countries using national accounts. A recent paper notes that "fully 28 of 29 other countries for which the OECD has compiled productivity growth data saw a deceleration in labor productivity growth over the last few decades. The unweighted average annual labor productivity growth rate across these countries was 2.3% from 1995 to 2004 but only 1.1% from 2005 to 2015" (Brynjolfsson, Rock, and Syverson 2017, p. 6). This is a robust result. Part of the problem is that traditional Solow-based models characterize the aggregate production function as a function of in-house factors of production. It implies that the greater productivity emanates from inside the firm or production unit. This means that total factor productivity also incorporates the productivity that should be attributed to outsourced services, but its value is biased upwards because Solow models erroneously attribute to it only improved efficiency.

Using this new characterization, the distinction between what is a manufactured good and what is a service becomes more blurred. This gradual transformation of manufacturing production characterized as a situation from a single production function of a firm in the 1950s (stage 1) to specialization of labor (stage 2) to splintering of production units (stage 3) is illustrated in Figure 4. YM is the output of manufacturing, whereas Ys is the output of services. In the past manufacturing was more capital intensive, and a simple measure of labor productivity would always yield a greater number in the manufacturing sector than in the service sector (YM/LM > YS/LS). Once the output of manufacturing is characterized as the result of a value chain of production units (stage 3), it becomes less clear that output per worker is higher for manufacturing than for services, because the distinction is blurred.

The suggested setup divides inputs into production units, not a good or a service.

Figure 4: Accounting for labor productivity within the stages of servicification



Source: authors

This leads to understanding the activities as labor effort within the production unit in time and space. Time use, through activity and technical competence, becomes a more relevant unit of measurement. This would entail classifying activities by degree of effort and valuing them according to difficulty or technical competence as is already done for time use surveys.

An example illustrates how the mismeasurement of productivity of services has as much to do with servicification as it does with the concept of what constitutes "work." Although the context of the activity should be reported, as in national accounts, it becomes irrelevant from the point of view of the valuation. The productivity of the activity is the labor effort within the production unit in time and space. The unit value could be the price, but should reflect a commonly defined social cost (as with public goods). Take the activity of roof repair of a 2x2 hole, which takes 3 hours with standard tools. It is performed by three men: Mr. A (who repairs the hole of his house's roof); Mr. B (a Church community volunteer) and Mr. C (a roofing repair worker at a large shingles manufacturing company, RM)). Using national accounts valuation, the marginal product of labor of Mr. C is higher, and equals his wage wRM, which means manufacturing sector is measured as most productive:

$$\frac{\Delta Y_C^{RM}}{\Delta L_C} \equiv W_{RM} > \frac{Y_B^{CH}}{L_B} \equiv \; > \frac{Y_A^{HH}}{L_A} = 0, \quad (1)$$

where YJi measures the real output or activity in a unit of time performed by individual i ∈ {A,B.C} in sector j ∈ {RM, CH, HH}. In the example, RM is the roofing manufacturing and installation sector, CH is the charity and community sector, and HH is the household sector. Li is the labor of person i. If we value with time use survey--sectors CH and HH are in the services

sectors--then twice the amount of value per worker was produced in the services sector compared to manufacturing sector (YHH+YCH>YRM). Since $YHH = YCH = YRM$ and $LA = LB = LC$

$$\frac{\Delta Y_A^{HH} + Y_B^{CH}}{\Delta(L_A + L_B)} > \frac{\Delta Y_C^{RM}}{\Delta L_C} \cong wRM \quad (2)$$

In conclusion, the productivity of services should be reassessed. The premature deindustrialization hypothesis assumes there is something inherently "special" about the traditional organization of manufacturing production, but the arguments and evidence for Asia here presented suggest otherwise. Part of the problem is that services are being erroneously measured and valued using the same national accounting tools we use to measure tangible manufactured goods. With the introduction of disruptive technologies in all spheres of life this measurement error is likely to grow. Policymakers and statistical offices need to adopt alternative measures of labor productivity sooner rather than later based on time use.

References

1. Andersen, B., & Corley, M. (2003). The theoretical, conceptual and empirical impact of the service economy (Vol. UNU-WIDER Discussion Paper No. 2003/22). Helsinki: UNU World Institute for Development Economics Research.
2. Asian Development Bank (2018b). *Key Indicators of Asia and the Pacific,* 49th edition, Asian Development Bank, Manila.
3. Brynjolfsson, E., Rock, D., & Syverson, C. (2017). *Artificial Intelligence and the Modern Productivity Paradox: A Clash Of Expectations and Statistics*. National Bureau of Economic Research Working Paper 24001, Cambridge, MA, November.
4. International Montary Fund (2018b). Measuring the Digital Economy, a Report by the Staff of the IMF. Washington, D.C.: International Monetary Fund, February 28.
5. Miroudot, S., & Cadestin, C. (2017). Services In Global Value Chains: From Inputs to Value-Creating Activities (Vol. OECD Trade Policy Papers No. 197): OECD Publishing.
6. Moulton, B. and van de ven, P., 2018, Addressing the Challenges of Globalization in National Accounts, NBER Conference, http://papers.nber.org/conf_papers/f100570/f100570.pdf
7. Rodrik, D. (2016). Premature deindustrialization. *Journal of Economic Growth, 21*(1), 1-33. doi:10.1007/s10887-015-9122-3.

The measurement of inequality of opportunity in China

Yu Jin, Ni Wendong
School of Statistics, Dongbei University of Finance and Economics,Dalian ,China

Abstract
Based on the data of China General Social Survey from 2003 to 2013 and a review of the existing methods of inequality of quantitative measurement opportunities, this paper uses parametric and no-parametric methods to quantitatively measure the degree of opportunity inequality and its changing trend of income inequality in China. The results reveal that the inequality of opportunity which the China's residents are suffering in the process of obtaining income is increasingly serious, and the inequality of opportunity is rising faster than income inequality, at the same time, the influence of individual' efforts on the income is becoming increasingly weak.

Keywords
Inequality of Opportunity; parameter method; no-parameter method; Changing Trend

1. Introduction

With the continuous development of China's economy and society, the residents' income has achieved steady growth. But the income gap is still large and unable to be solved, the Gini coefficient of the income remains high. According to the data released by the National Bureau of statistics, the Gini coefficient has been above 0.44 of the global average since 2003. In 2016, the Gini coefficient of China's resident income was 0.465, and 0.4 is usually taken as a warning line in the world. More than 0.4 indicate that the income gap is large. Thus, there has been a very serious problem of unequal income distribution in our country. Income inequality causes residents to "hate rich" mentality, which leads to rising crime rate and social unrest. At the same time, excessive income inequality will affect economic growth, and even more serious will lead to political instability.

The Central Committee of the party and the State Council attaches great importance to the issue of income distribution, and has repeatedly proposed to focus on solving the problem of excessive income gap, thus forming a reasonable distribution pattern of income, which will benefit the people all over the country more fairly and truly realize the goal of common prosperity for the whole people. Based on this "big environment", income distribution has always been a hot topic for domestic scholars. In recent years, the trend

of related research has gradually shifted from the direct study of income inequality tilt to the study of opportunity inequality in the process of income acquisition. Equality includes the starting point, the opportunity and the result equality, and the equal opportunity among the three is the most important. The "hatred of the rich" is often due to the lack of fair access to high income. If the income level is determined entirely by individual efforts, the unfairness of the income results will not cause dissatisfaction. Therefore, it is a new research perspective to study income inequality through inequality of opportunity.

2. Methodology

2.1 Data sources

This paper uses the China Comprehensive Social Survey (CGSS) data provided by the China Survey and Data Center of Renmin University of China. Since 2003, large-scale sampling surveys (excluding 2007) have been conducted annually, and what is available now is the survey data for a total of eight years from 2003 to 2013 (excluding 2004, 2007 and 2009). CGSS data has a large sample size, wide coverage and sufficient information, which is the authoritative micro-data for the study of social problems in China

2.2 Variable selection

In the process of choosing "environment" factors, firstly, for urban and rural factors, because the object of the study of opportunity inequality is not family information, the income of rural residents is recorded by the total family income in the survey process. But the total family income is determined by the objective "environment" factors and "efforts" of all members of the family, which will make it difficult to match the "environment" information of the investigated rural individuals with their personal income. However, considering that the urban and rural household registration system greatly affects the income gap, this paper uses the equivalent per capita income formula to estimate the personal income of rural residents, and then matches the "environment" factors of the investigated individuals. Secondly, based on previous studies, it is found that parents' educational level and occupation have a significant impact on their children's future income after entering the labor market. Therefore, this paper incorporates parents' educational level and occupation into the "environment" factors. However, due to the problem of questionnaire design, the different years of the questionnaire on career and education have different classification. At the same time, because of the wide range of occupational variables, it is necessary to re-classify and re-code the occupations before analyzing. According to the existing literature on the re-classification of occupations (Gao Yong, 2009 [9]), the occupation is divided into three categories: high, middle and low. Higher-ranking professions include state power organs, judicial organs, government agencies and

departments at all levels, heads of enterprises and institutions, and heads of scientific, educational, cultural and health institutions. The middle class occupations include service personnel, production and processing personnel, maintenance equipment personnel, planting and feeding personnel, informal employees and other employment personnel difficult to classify; the lower class occupations are farmers. For the level of parental education, in the process of data collation, we found that most of the parents of the surveyed individuals have a low level of education, and very few have received high school or higher education. Based on this, the education level of parents is re-classified. Firstly, the information-deficient individuals are deleted, and then the education level is divided into three categories: high, middle and low by analogy with the classification of occupation. The upper classifications include secondary and higher education, the middle classifications are primary education, and the lower classifications are non-formal education and illiteracy. In addition to household registration, parental occupation and parental education, gender isalso included as an important "environmental" factor in the analysis.

a. Measurement methods

The tools used in the existing research on measure inequality of opportunity may be different, but the ultimate aim is to work on the inequality index of the Computer Association (Bourguigonon, 2003; Ferreria and Gignoux, 2008; Checchi and Peragine, 2010), thus comparing the inequality of opportunities in different regions or different exogenic conditions. According to different tools, measurement methods can be divided into parameter method and no-parameter method. However, both parametric and no-parametric methods first need to separate the functions of "environment" and "effort" by constructing a virtual counterfactual distribution, and then calculate the inequality index of opportunity.

Bourguigonon (2007),Ferreira and Gignous (2011) put forward the parameter method. They suggest that the logarithm of income be set in linear form and the least squares method be used to estimate (1) and (2). The construction of y^c is as follows:

$$y^c = \{\exp(C^1\psi)1_{N_1}\} \dots\dots, \{\exp(C^T\psi)1_{N_T}\} \quad (1)$$

Where ψ is the fitted value of the parameter, $N_t$represents the number of individuals in the t -th class "environment", and 1_{N_T} represents the unit row vector of the N_t dimension. This method of constructing the counterfactual distribution is called the parameter method.

Checchi and Peragine（2010）proposed a nonparametric method for estimating y^c :

$$y^c = \{\mu(y^1)1_{N_1}, \dots, \mu(y^T)1_{N_T}\} \quad (2)$$

Where $\mu(y^t)$ means the mean of y^t . Since it has been assumed that the fitting results are entirely determined by objective "environmental" factors, $I(y^c)$ is a measure of opportunity inequality.

Jiang Qiuchuan and Zhang Kezhong（**2015**） introduced the concept of equitable distribution equivalence income proposed by Atkinson[1] (1970). Its magnitude depends on the unequal degree and average income level of real income distribution. Therefore, the introduction of this concept can reflect both expected income and income fluctuation. The equal distribution equivalent income of y^{ts} is:

$$y_{EDE}^{ts} = \begin{cases} \left(\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_{tk}}\left(y_{ki}^{tS}\right)^{1-\lambda}\right)^{\frac{1}{1-\lambda}} & \lambda \neq 1 \\ \left(\prod_{k=1}^{k}\prod_{i=1}^{N_{tk}} y_{ki}^{tS}\right)^{1/N_t} & \lambda = 1 \end{cases} \quad (3)$$

Among them, $\lambda \geq 0$ is the unequal aversion coefficient proposed by Atkinson (1970), thus a new nonparametric estimation is obtained:

$$y^c = \left\{y_{EDE}^{1S}1_{N_1}, \ldots, y_{EDE}^{tS}1_{N_t}, \ldots, y_{EDE}^{TS}1_{N_t}\right\} \quad (4)$$

By comparing the newly established counterfactual distribution, the expected income and income risk of different "environments" are fully considered. Therefore, this paper uses this counterfactual distribution to measure the degree of inequality of opportunity by no-parametric method.

3. Results

The construction of counterfactual distribution is to measure the degree of inequality of opportunity. When we measure inequality of income and inequality of opportunity, the unequal index we choose is Gini coefficient. However, in order to ensure the accuracy of the results, all calculations are repeated using the Theil index. It is found that inequality of opportunity estimated by Theil index is smaller, but the overall trends and differences between the two indicators are basically the same. So the results calculated by Gini coefficient will only be presented in the subsequent reports. At the same time, this paper will analyze the absolute value $I(y^c)$ and the relative value $I(y\hat{}c)/I(y)$ of the opportunity inequality.

Firstly, we analyze the general situation and the trend of change year by year of inequality of opportunity in the income of our residents from 2003-2013, and the results are shown in Figure 3-1. From the calculated Gini coefficient of residents' income in each year, we can see that the inequality of residents' income in China has a rising trend in 2003-2013. In this paper, two points need to be emphasized: first, the income data in this paper do not adjust inflation. Considering that the calculation result of income Gini

coefficient is essentially a relative value, this paper argues that there is no need to adjust inflation; secondly, the Gini coefficient of household income is calculated by personal income data, so the results may be different from those calculated by per capita household income in most studies. However, the conclusion is the same that the income Gini coefficient has a tendency to increase, which shows that the measurement of income Gini coefficient in this paper is scientific and reasonable.

Table 1 Measurement of inequality of opportunity

Year	2003	2005	2006	2008	2010	2011	2012	2013
Gini coefficient	0.427	0.450	0.440	0.457	0.612	0.570	0.546	0.529
Part A absolute value								
Method1	0.129	0.153	0.126	0.169	0.383	0.337	0.317	0.307
Method2	0.112	0.150	0.135	0.162	0.375	0.362	0.365	0.369
Method3	0.129	0.154	0.128	0.167	0.373	0.332	0.314	0.306
Part B relative value								
Method1	0.302	0.341	0.286	0.369	0.626	0.592	0.581	0.581
Method2	0.263	0.334	0.306	0.356	0.612	0.635	0.668	0.697
Method3	0.302	0.341	0.292	0.365	0.609	0.583	0.575	0.577
Sample size	4151	4796	3501	2922	6415	3357	7213	6021

Explanation: Part A is the absolute value of calculating inequality of opportunity, Part B is the relative value of calculating inequality of opportunity. At the same time, Method1、Method2、 Method3 represent three kinds of counterfactual distribution construction techniques respectively: nonparametric methods of C □ecc □i and Peragine(2010), the combination of nonparametric and parametric methods of Ferreira and Gignoux（2011） and new technical methods proposed by Jiang Qichuan and Zhang Kezhong (2015）.In addition, the results of the 200 bootstrap tests for all the estimates in this paper are significant.

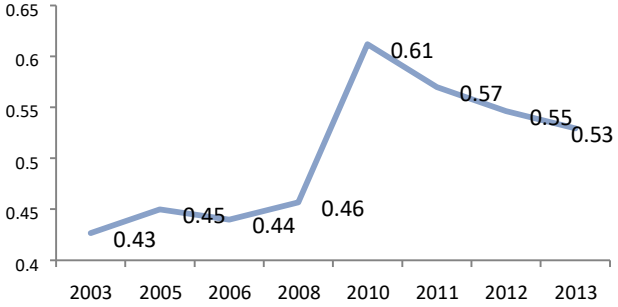


Figure 1 The trend of change of Gini coefficient of residents' income from 2003 to 2013

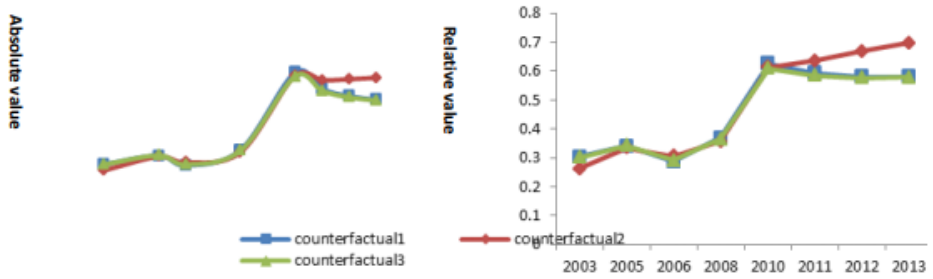


Figure 2 The trend of change of inequality of opportunity

This paper focuses on the proportion of opportunity in inequality of income, that is, to what extent does inequality of opportunity create inequality of income. Part A is the absolute value of inequality of opportunity. The results obtained by using three different counterfactual distribution methods are similar and indicate that inequality of opportunity faced by our residents in gaining income has shown a trend of rising first and then decreasing slightly in the past decade. In recent years, the trend has tended to be gentle and staid at a higher value. The results show that the "environment" factors determine the level of personal income to a large extent. Having a better environment means that people with better opportunities are more likely to earn higher incomes. Part B is the relative value of inequality of opportunity. It is the ratio of inequality of opportunity to income inequality. The relative value is determined by the growth rate of inequality of opportunity and inequality of income. The trend of the relative value is the same as the absolute value of inequality of inequality. All in all, it can be seen that the degree of inequality of opportunity faced by our residents in gaining income is higher and its growth rate is faster than income inequality's. In other words, the degree of effort in individual work is weakening in determining the level of income. Excessive degree of inequality of opportunity will reduce people's enthusiasm and passionfor work, and thus hamper people's incentive to accumulate their human capital.

4. Discussion and Conclusion

Compared with income inequality, the problem of inequality of opportunity is more practical and important. The more practical reason is that the old saying goes, "Inequality rather than want is the cause of trouble". The essence of people's complaints about income inequality is dissatisfaction with the unequal opportunities they encounter in obtaining income. If the income gap is caused entirely by different degrees of individual efforts, then there is no doubt about it. More importantly, the impact of inequality of opportunity on income inequality is rising. Secondly, inequality of opportunity will

undermine people's incentives to accumulate human capital, bury potential talents, and make people lose enthusiasm for hard work, thus hindering sustained economic growth. Finally, too high degree of inequality of opportunity will intensify people's " enrichment " mentality, which is not conducive to social harmony and stability. Therefore, it is imperative to explore the problem of inequality of opportunity.

This paper uses the existing methods of measuring inequality opportunity, selects several different measurement indicators, and uses the data of China General Social Survey from 2003 to 2013(except 2004, 2007 and 2009) to measure the opportunity inequality in income inequality. At the same time, it studies the difference of inequality of opportunity in different areas, different gender, different household registration and different family background, and the bias effect of the above "environment" factors on inequality opportunity. The results show that inequality of opportunity faced by our residents in the process of earning income shows an upward trend. And the growth rate of inequality of opportunity is significantly faster than income inequality's, with the relative value rising rapidly from nearly 30% in 2003 to nearly 60% in 2013.It can be seen that the effect of individual effort on the level of income is weakening.

Solving the problem of inequality of opportunity in the process of earning income is the key to narrowing the income gap of our residents. To eliminate the objective environmental factors is the essence of ensuring inequality of opportunity. The unfair distribution of resources led by power has resulted in inequality of opportunity in the process of employment. The social and economic status of parents and their social relations play a decisive role in their children's work and income. Individuals with a good family background tend to have a great competitive advantage, and this influence is often passed on from generation to generation, resulting in a pattern of "good always good, but bad hard to get ahead". In order to ensure inequality of opportunity, we must break this pattern, avoid the "family background"、"family relations" and other factors to determine personal development, and stop the "family background" and "family relations" in the intergenerational unreasonable transfer. The solutions are as follows: Firstly, state-owned enterprises and institutions as "disaster areas" should improve the open recruitment system to ensure that the recruitment process is transparent; Secondly, the supervision mechanism should be introduced to ensure the healthy competition of the labor force in the job market.

References

1. Chen Dong, Huang Xufeng,. To what extent does inequality of opportunity affect income inequality? - Based on the perspective of intergenerational transfer [J]. Economic Review.2015(1):317.

2. Chen Zhao, Lu Ming, Sato Hiroshi. Who has entered the high-income industry? - The role of relationship, household registration and productivity [J]. Economic Research.2009 (10): 121-133.
3. Chen Lin, Yuan Zhigang. The trend and internal transmission mechanism of intergenerational income mobility in China[J]. Quantitative and technical economics.2011 (1): 130-140.
4. Jiang Qiuchuan,Ren Jie,Zhang Kezhong. Study on inequality of opportunity of urban residents in China[J].World economy.2014(4): 111-140.
5. Francois Bourguignon, Francisco H.G. Ferreria Marta Menendez. Inequality of outcomes and inequality of opportunities in Brazil[J].Reviews of income and wealth.2003(12):79-102.
6. Nuneza J Tartakowsky A. The relationship between income inequality and inequality of opportunities in a high-inequality country: the case of Chile[J].Applied Economics Letters. 2011(4):359-369.

Predicting the number of newly found rare species

Tsung-Jen Shen[1],Youhua Chen[2]

[1]Institute of Statistics & Department of Applied Mathematics, National Chung Hsing University, Taiwan

[2]CAS Key Laboratory of Mountain Ecological Restoration and Bioresource Utilization & Ecological Restoration and Biodiversity Conservation Key Laboratory of Sichuan Province, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, China

Abstract

In natural ecological communities, most species are rare and thus susceptible to extinction. Consequently, the prediction and identification of rare species are of enormous value for conservation purposes. How many newly found species will be rare in the next field survey? From a Bayesian viewpoint, by using observed species abundance information in an ecological sample, we developed an accurate estimator for estimating the number of new rare species (e.g., singletons, doubletons, and tripletons) that will be found in an additional unknown sample. A semi-numerical test showed that the proposed Bayesian-weight estimator accurately predicted the number of rare new species with low relative bias and relative root mean squared error and accordingly, high accuracy.

Keywords

species rarity; biodiversity survey; Bayesian statistics; sampling theory; diversity estimation

1. Introduction

Species abundance distribution, or rank-abundance distribution, is one of the most important community patterns with wide applications in ecology (Fisher et al. 1943; Preston 1948; Chen & Shen 2017). One of its key applications is to predict species richness and diversity in ecological communities, particularly when additional ecological surveys are needed (Shen et al. 2003). However, almost all previous studies have utilized sample relative abundance derived directly from sampled ecological communities to generate rank-abundance distribution curves (Magurran 2004). This practice tends to overestimate the true relative abundance of species (Chao et al. 2015), with the overestimation being magnified for rare species or when the sample size of the studied community is small.

Because rare species are highly vulnerable and prone to extinction when exposed to climate change and habitat loss, the identification and protection of rare species is always a top research priority in conservation biology and

community ecology (Kunin & Gaston 1993). In this context, the development of robust statistical methods for accurately predicting the occurrence of new rare species in additional samples is urgent and necessary.

2. Methodology

Assume that an initial sample of n individuals is collected from a metacommunity of S species in which the species relative abundance distribution is given by $p_1, p_2, \ldots, p_S$. Let the binary random variable $Z_{i,j} = 1$, signifying that the jth selected individual is identified as species i; otherwise $Z_{i,j} = 0$, where $j = 1, 2, \ldots, n$. $(X_1, X_2, \ldots, X_S)$ represents species counts with $X_i = \sum_{j=1}^{n} Z_{i,j}$ in the sample and follows a multinomial distribution with a grand total of n and occurrence probabilities of $p_1, p_2, \ldots, p_S$. Let $F_k(m)$ be the expected number of newly found species absent in the first sample but have exactly k individuals detected in an additional sample of size m. Similarly, the binary variable $Z_{i,j}$ ($j = n+1, n+2, \ldots, n+m$) can be used to describe the sampling outcome for the additional sample. Mathematically, we can express $F_k(m)$ by

$$\begin{aligned} F_k(m) &= E\left[\sum_{i=1}^{S} I(X_i = 0)\binom{m}{k} p_i^k (1 - p_i)^{m-k}\right] \\ &= \binom{m}{k} \sum_{i=1}^{S} P\left(Z_{i,1}, Z_{i,2}, \ldots, Z_{i,n+m} \middle| \sum_{j=1}^{n+m} Z_{i,j} = k\right) P\left(\sum_{j=1}^{n+m} Z_{i,j} = k\right)' \end{aligned} \quad (1)$$

where I(A) is an indicator function such that I(A) = 1 if statement A is true and I(A) = 0 if untrue. Note that the term $p_i^k (1 - p_i)^{n+m-k}$ in $F_k(m)$ can be equivalently interpreted as exactly k individuals of species i coming out of $n+m$ individuals in the combined original and additional samples. The derivation from the second equality to the last equality in Eq. 1 is based on the fact that $Z_{i,j}$ entities can be regarded as independent random Bernoulli variates with success rates p_i while the sampling outcome is that there are k successful outcomes out of $n+m$ trials.

Incorporating Bayesian weights, we proposed an estimator of $F_k(m)$ as (Shen & Chen 2018)

$$\hat{F}_k(m)=\frac{\binom{m}{k}}{\binom{m+n}{k}-\binom{m}{k}}\sum_{q=1}^{k}f_q\binom{m}{k-q}\hat{d}_q^{k-q}(1-\hat{d}_q)^{m-(k-q)},$$

where $f_q=\sum_{i=1}^{s}I(X_i=q),\hat{\delta}_q=\left(1-\hat{\lambda}e^{-\hat{\theta}q}\right)\frac{q}{n}$, (Chao et al. 2015) is the estimated relative abundance of an observed species with $X_i=q>0$. The two estimates $\hat{I}$ and $\hat{q}$ are numerical solutions of the following system of nonlinear equations

$$\begin{cases}\sum_{q=1}^{n}f_q\frac{q}{n}\left(1-I e^{-qq}\right)=1-\frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1+2f_2}\right]\\ \sum_{q=1}^{n}f_q\left[\frac{q}{n}\left(1-I e^{-qq}\right)\right]^2=\frac{\sum_{q=1}^{n}q(q-1)f_q}{n(n-1)}-\frac{2f_2}{n(n-1)}\left[\frac{(n-2)f_2}{(n-2)f_2+3f_3}\right]^2\end{cases}.$$

Additionally, we also compare the proposed method with one unweighted estimator (a naïve method) given as follows:

$$\begin{aligned}\tilde{F}_k(m)&=\binom{m}{k}\sum_{i=1}^{S}\left(\frac{X_i}{n}\right)^k\left(1-\frac{X_i}{n}\right)^{n+m-k}\\&=\binom{m}{k}\sum_{j=1}^{n}f_j\left(\frac{j}{n}\right)^k\left(1-\frac{j}{n}\right)^{n+m-k}\end{aligned}$$

3. Results

One empirical species abundance data set was used and normalized (i.e., the abundance of each species was divided by the total number of individuals in a given dataset) so as to create a sampling relative abundance distribution ($p_1,p_2,\cdots,p_S$) for each data set. The used data set is that the Malayan butterfly data provided in Fisher et al. (1943) contained 620 species identified from 9031 individuals.

For the data set, we designated two initial sample sizes, *n* = 100 and 200. We randomly sampled each of these *n* individuals from the entire data set to represent observed species abundance information. Each additional (but unsurveyed) sample was designated a size of m = $0.5\times n, 1\times n, 1.5\times n$ and $2\times n$. For each combination of *n* and *m*, we independently simulated 2000 replicates. We then estimated the expected number of newly discovered rare species (i.e., singletons, doubletons, and tripletons) in the additional sample using the proposed Bayesian-weight estimator.

Our main findings can be concisely enumerated as follows.

a) The proposed Bayesian-weight estimator predicted the number of singletons, doubletons, and tripletons accurately, as supported by the direct comparison of true and estimated values in the two empirical data sets (Tables 1).
b) The proposed Bayesian-weight estimator had lower relative bias and lower relative root mean squared error (thus indicating higher accuracy) in comparison to the naïve estimator (Tables 1).

4. Discussion and Conclusion

Natural species abundance distributions observed in the real world show a pattern of only a few species being common while most species are rare (Magurran & Henderson 2003). Our studies showed that this was also true for newly discovered species in our simulated new samples (Tables 1). Expected and true values for doubleton new species were much smaller than for singleton new species, while the values for tripleton new species were much smaller than for doubleton new species.

In conclusion, in this study we developed a novel Bayesian-weight estimator for predicting rare new species in potential additional ecological samples. The method we developed was nonparametric and accurate. In addition to rare species, our statistical model was equally powerful for predicting the number of common species, which are important drivers of many ecosystem functions (Gaston 2011; Houadria & Menzel 2017)

References

1. Chao, A., T. Hsieh, R. Chazdon, R. Colwell, and N. Gotelli. 2015. Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. Ecology **96**:1189–1201.
2. Chen, Y., and T. Shen. 2017. A general framework for predicting delayed responses of ecological communities to habitat loss. Scientific Reports **7**:998.
3. Fisher, R., A. Corbet, and C. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology **12**:42–58.
4. Gaston, K. 2011. Common ecology. BioScience **61**:354–362.
5. Houadria, M., and F. Menzel. 2017. What determines the importance of a species for ecosystem processes? Insights from tropical ant assemblages. Oecologia **184**:885–899.
6. Kunin, W., and K. Gaston. 1993. The bioogy of rarity: patterns, causes and consequences. Trends in Ecology and Evolution **8**:298–301.
7. Magurran, A. 2004. Measuring biological diversity. Blackwell, Oxford.

8. Magurran, A., and P. Henderson. 2003. Explaining the excees of rare species in natural species abundance distributions. Nature **422**:714–716.
9. Preston, F. 1948. The commonness and rarity of species. Ecology **29**:254–283.
10. Shen, T., A. Chao, and J. Lin. 2003. Predicting the number of new species in further taxonomic sampling. Ecology **84**:798–804.
11. Shen, T. & Chen, Y. (2018) A Bayesian weighted approach to predicting the number of newly discovered rare species. *Conservation Biology*, 10.1111/cobi.13253.

Table 1. Estimate of newly found rare species (singletons, doubletons, and tripletons) in the Malaysian butterfly data set. $F_k(m)$ is the true value for rare species with abundance k in an addition sample size m. Avg denotes the average of estimates using 2000 replicates. Performance of each estimator was evaluated using relative bias (Rbias) and relative root mean square error (Rrmse).

n	m	k	$F_k(m)$	Proposed: $\hat{F}_k(m)$			Naive: $\tilde{F}_k(m)$		
				Avg	Rbias	Rrmse	Avg	Rbias	Rrmse
100	50	1	27.950	28.868	0.033	0.131	8.026	−0.713	0.713
		2	2.164	1.928	−0.109	0.225	2.154	−0.005	0.056
		3	0.191	0.179	−0.061	0.237	0.407	1.131	1.134
	100	1	48.159	50.519	0.049	0.171	9.420	−0.804	0.805
		2	6.750	6.246	−0.075	0.192	4.942	−0.268	0.273
		3	1.080	0.936	−0.134	0.276	1.792	0.658	0.664
	150	1	63.150	65.778	0.042	0.200	8.358	−0.868	0.868
		2	12.084	11.905	−0.015	0.152	6.475	−0.464	0.466
		3	2.6411	2.257	−0.145	0.290	3.418	0.294	0.306
	200	1	74.487	76.307	0.024	0.241	6.654	−0.911	0.911
		2	17.423	18.222	0.046	0.139	6.806	−0.609	0.610
		3	4.662	4.092	−0.122	0.291	4.699	0.008	0.076
200	100	1	37.244	39.138	0.051	0.128	12.314	−0.669	0.670
		2	4.334	4.021	−0.072	0.132	3.483	−0.196	0.201
		3	0.574	0.516	−0.101	0.167	0.727	0.267	0.272
	200	1	59.976	62.743	0.046	0.161	14.092	−0.765	0.765
		2	11.983	11.969	−0.001	0.087	7.614	−0.365	0.367
		3	2.755	2.435	−0.116	0.186	2.930	0.064	0.084
	300	1	74.624	76.033	0.019	0.185	12.386	−0.834	0.834
		2	19.505	21.090	0.081	0.111	9.748	−0.500	0.502
		3	5.924	5.403	−0.088	0.170	5.328	−0.101	0.115
	400	1	84.446	82.425	−0.024	0.207	9.849	−0.883	0.883
		2	26.026	30.213	0.161	0.180	10.153	−0.610	0.611
		3	9.396	9.209	−0.020	0.139	7.152	−0.239	0.244

Improving the quality of official statistics in Iran's manufacturing establishments statistics

Saeed Fayyaz[1], Gholamhossein Mosalmani Nooshabadi[2]
[1]Statistician on Transportations and ICT statistics
[2]Statistician on Manufacturing establishment statistics

Abstract

Official statistics refer to public information which is produced for the benefit of the society and is funded by the state budget under the official or statistical programme. Also, in GSBPM[1], quality of official statistics producing by statistical agencies has been highlighted significantly. Despite the fact that data quality framework has different dimensions comprising relevancy, timeliness and punctuality, coherence and comparability, accessibility and clarity, this paper focused on accuracy and reliability aspect. Additionally, a specific case study on manufacturing establishments' statistics in Iran was expanded. Due to its importance, providing consistent, reliable and high accuracy data has been one of the prominent targets of these sector. In order to improve this data quality aspect in these case, a practical software was proposed to help statisticians and experts in NSOs[2]. Although there are a range of causes and factors that are significant in improving the reliability and accuracy of data from data gathering to publishing, the proposed software focus on the Editing and imputation stage in this process, specially macro editing. Presumably, quality of the gathered data from the surveys or establishments directly are just affected by technical editing in the process and other pre-editing stages have trivial impact and not completely controllable and improvable based on experts' opinions. Additionally, this software can obviate the data non-consistency by providing multi-dimensional insights and comparable data of each establishment. In this comprehensive overview different dimensions of growth e.g. ISIC[3] activity field, provincial and sub-provincial cutting level etc. was reflected. Remarkably, financial statements as secondary register-based data source in this process either to make comparison or replacement with survey data will promote the accuracy was considered is this software. These registered data source are available free in stock market websites or received physically from Audit organizations. This study also can help other statistical agencies to produce higher quality statistics indifference the kind of activity or where they are located.

[1] Generic Statistical Business Process Model (GSBPM)
[2] National Statistics Organizations(NSOs)
[3] International Standard of Industrial Classification (ISIC)

Keywords
Quality of official Statistics; Manufacturing Establishments; Editing and imputation; Analytical Software; Secondary data source

1. Introduction

Official statistics provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honor citizens' entitlement to public information.

2. Reliability : Problems in data collection and processing

The reliability of statistical information criteria is present in various models of statistical quality assessment. It implies the absence of non-sampling errors in the case of operations from administrative records such as design, register or sub-register, misclassification, partial responses, etc. It also implies the reduction of non-response and data reliability by applying rules of internal and external validation in the case of sample surveys. Reliability rests largely on the work of collecting the data and performing the analysis in its consistency. If there are failures in these processes, reliability is affected directly, so that measures should take place to prevent this.

In other side, new global developments of the past two decades were taken into consideration bearing in mind the critical role of high-quality statistical information, the necessity to ensure public's trust in official statistics and stressing that legal and institutional frameworks need to guarantee fundamental statistical values and principles. Regarding to the official role and constitutions' law of Iran, Statistical Center of Iran has appointed as the main responsible to design, implement, develop and support the integrated statistical register systems in Iran. Also, providing high quality and on-time statistics has been one of the key goal of SCI and manufacturing statistics is not an exception.

3. Discerption of manufacturing Statistical dissemination process in SCI

There are different steps in dissemination of official statistics in SCI. All data gathered from both censuses and surveys should be pass different process's steps to be prepared for final users. As the all data in SCI, manufacturing statistics is not an exception and like other surveys it provides very practical and informative data for top managers and planner at national level to final users and university students. Manufacturing establishments' surveys hold annually and it also disseminate after finalized the data in

technical group at SCI. In general model of statistical process which is common in many different counties there are 9 steps.

Fig1.Generic Statistical Business Process Model

Manufacturing establishment's survey has been conducted on paper-based form and its statistics comprise of different steps and sub-steps as shown below:

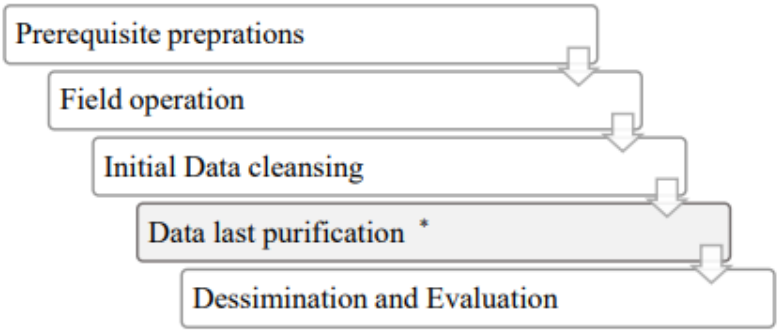


Fig2. Different steps of manufacturing establishment's statistics process

Holding surveys in such a big scale in Iran with 31 cities and approximately 1400 sub-city needs a considerable source of finance and educated workforce and it need to use hierarchy method, integrated management system, intelligent computer- aid systems and enormously precise planning. Also proving an up-to-date statistical frame is problematic issue in manufacturing establishments' statistics. With approximately 47000 establishments, SCI needs to design flexible process to provide high quality statistics and updated its statistical frame. Generally manufacturing establishments divided into 3 main categories:

- Between 10 and 49 worker
- With 50+ worker
- Confirmed financial statement with audit organization (Always 300 + worker)

It is worth mentioning that manufacturing establishments with less than 10 workers haven't been considered in the above target population and a separate survey is conducted periodically. (Every 10 years)

All the steps in the process shown in the fig1 have different steps and sub-divisions. In this study a quality- oriented perception will be discussed on theses sub-divisions which affect the final results. Although there are different causes and factors which play a role in the quality of results, just some of them can be controlled and measured. In following, all divisions and they roles will be described in tables 1 to 5. In these tables, although all the items have a significant effect (both direct and non-direct) on the quality, some of them are controllable causes and have underlying affects than others. In tables, controllable cause has been determined based on experts' opinions for every subdivision.

Table1. Process description- step 1

Level	Title	Sub-divisions	Improvable
1	Prerequisite preparations	Questionnaire design	No
		Manuals and guidelines	Yes
		Frame and sampling	Yes
		Educational course (Experts & enumerators)	Yes
		Survey planning and organization	No

As classified in table 1, there are 3 subdivisions which are controllable and affect the data quality. In order to enhance the quality based on manuals and guidelines, experts should try to clarify questions as much as possible and they should obey the international recommendations, standards and role to provide clear concepts for both enumerators and respondents. In addition, educational courses for all categories can be improved the quality if a comprehensive reference will provide. Also, Education should be vibrant and it proportional to new changes in target population.

Meanwhile, the section which need to be consider particularly is the statistical frame and sampling. It is necessary to control the frame full coverage and try to minimize the under coverage. In parallel, selecting an effective sampling method with minimum sampling error and considering all fund limitations and workforce is an undeniable issue to find optimum way. Regarding to the annual conduction of manufacturing establishments' survey, all of sub-divisions in level 1 (prerequisite preparations) should be revised and updated in a very optimum way. The next step in the process is shown in table2.

Table2. Process description- step 2

Level	Title	Sub-divisions	Improvable
2	Field operation	Face to Face interview	Yes
		Data receiving by E-mail, Fax or telephone	No
		Progress continues reporting	No
		Planning for non-cooperation units	Yes
		Inspection the process	No
		Data entry to systematic software	Yes

In the level 2 of the process, operational enumerations and data gathering in paper questionnaire in pre-scheduled time will be completed. Although, there are different data gathering methods (face-to-face, telephone, email, fax etc.), answers from respondents are considered as accurate and basic data (even the data have conformity with financial statements or not). In this level, all necessary data have been gathered by enumerator or by manufacturing establishment itself and in data entry process all of the data gathered in paper forms or other means registered in the database. Based on provided information, improving the data quality is not controlled directly and related office in the SCI just can inspect the overall process. Subsequent step contains initial data purifying describe in table3.

Table3. Process description- step 3

Level	Title	Sub-divisions	Improvable
3	Initial Data cleansing	Verification and revision	Yes
		Primary technical Editing	Yes
		Case controlling	Yes
		Completion of non-cooperation units	Yes

In this level, verification of data entry level will be examined to diminish the data entry's errors. Generally, a kind of weighted sampling and critical level of 5% and less for accepting the errors will be exerted. In this activity, two independent persons should register the data to prevent the errors again. After the data entry and verification, at the higher level, preliminary technical editing shall be done. This task includes the first questionnaire revision to be insure of completion, mathematical errors, dependency and mathematical-related errors. In the next step, accuracy and precision of questionnaire will be reviewed and finally, it will confirm by educated and experienced experts who are holding the surveys in provincials. Further, they control the practical errors, inspect the overall survey holding process, revisers and the enumerators,

provide enumerators' needs of transportation, pre-cooperation with establishment's owner, providing the necessary documents etc.

They also control and confirm the accuracy of data by technical editing and revising the gathered data from previous steps. If there is any necessity, experts control the data of establishments case by case and compare the gathered data with financial statement or register-based information, if there are available. Here, thought all the tasks have high importance in the quality of data, improving in the quality of this phase needs to consider supplementary acts as below:

1. Proving guidelines and manuals with more highlighting on quality (Quality tips)
2. Improving the quality of educational courses
3. Employing the more talented and experienced enumerators
4. More emphasis on cases especially those establishments revealed their audited financial statements before the survey
5. Applying persuasive policies for full completion, better cooperation and less non-responded units

However, there are different kinds of data accuracy controls and applying editing methods, this data base needs more cleansings and purifications to be get ready for dissemination process. This target will be cover at next level which is somehow the basic issue of this paper.

Table4. Process description- step 4

Level	Title	Sub-divisions	Improvable
4	purification	Data verification	Yes
		Micro editing	Yes
		Macro editing	Yes
		Tabulation	Yes
		Confidentiality check	No

At this level, there are extremely important factors which affect the data quality. After creation of database, verification of real data and confirmed data from experts in the provincials are very important. As far as all provincials have not the same of samples and volumes, there are some differences should be considered perfectly. It is assuming that all the data had been finalized in the last level and editing shall be done on the confirmed data by provincials' experts. Micro editing includes considering the control and checking the data accuracy from general information of establishments to financial data.

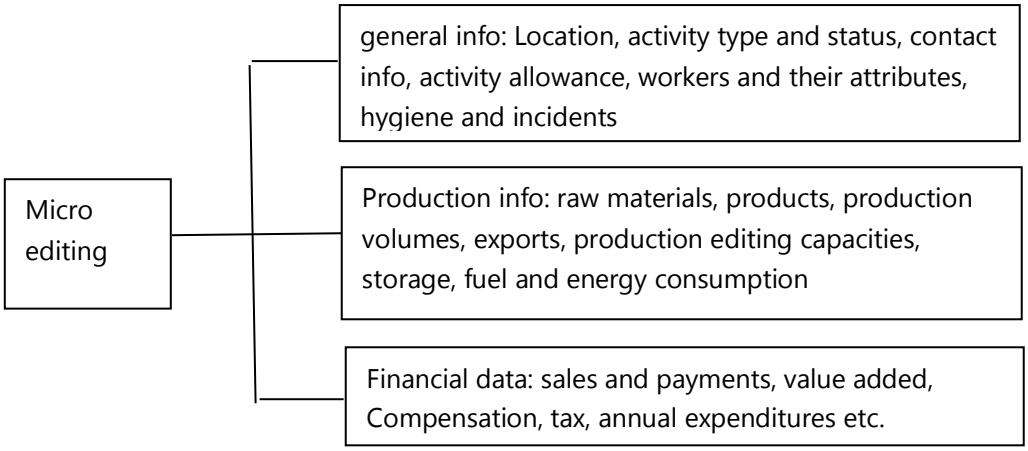


Fig3. Micro editing implications

In this paper, a software will be proposed in order to construct systematic editing and harmonize editing process. It can also prevent parallel controlling and checking and it can cause to improving the data quality.

4. Proposing Analytical software

As mention before, there are two types of editing in the level 4 including micro and macro editing. Proposed software can cover both of editing methods simultaneously. Presumably, all of the establishments have Unique Identification Number (UIN) in the frame and the frame update regularly. In addition, UIN are the same in other register-based systems in other organizations' registered systems and all of the transferred data from this database that show in register-based tab will be connected to UIN. In the software, all tabs and menus will be activated by entry the UIN and all the tools will be connected to this establishment. Supplementary, proposed software will be focus on bellow criteria simultaneously:

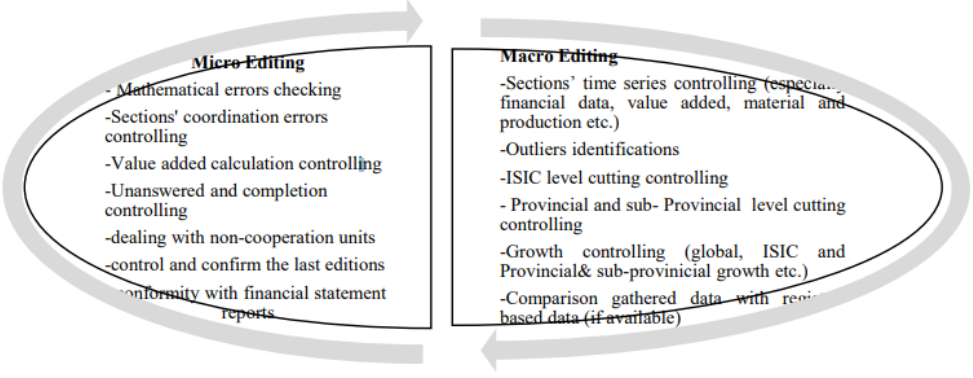


Fig4. Purification of data in level 4

All of cases in the table4 will be available in a separate tab in the software. In the software, it will be necessary to entry UIN at first and using tabs and tools to analysis the establishment status. Controlling and errors' removal is the major goal of the proposed software. In addition to the mentioned criteria there are also other measurable and comparable criteria to be considered significantly. These kind of changes are in:

1. Input-Output, value added and they proportions (especially industrial value added)
2. Raw materials, products volume, inventory during construction and they proportions
3. Workforce, men and female, permanent or contemporary combination and salary trend
4. Inventory, fuel consumption and energy with production and their proportion
5. Investments, export and production capacities
6. Manufacturing and non-operating expenses and their proportion

The software schematic will be shown in fig5-7 as below. In the software, all outliers and data inconsistency will be recognized and experts should resolve the problem or find a reasonable and defendable illustrations.

Fig5. Software schematic for Macro editing

As shown in the fig 5, the first screen of the software includes different characteristics information of statistical unit, contact information of staffs who are engage in survey in all levels & establishment and table of the checking panel. This table also provide a short description of errors and warning which may experts will face during the purification process. In other tabs of the software there are other different analysis from ISIC trend analysis and provincial cutting level statistics to cross checking and conformity with register-based data (if it is available).

Subsequently, the final level of manufacturing establishment's statistics is the dissemination. In reality, dissemination is what final users will receive online. After finalizing the data editing and tabulation process, all the tables should change to the dissemination's format and standard which had been determined before. This process will be done electronically without any interfere or individual involvements and as a result data quality won't be affected and maintain as much as latest level. In the other word, the data interface will be transformed into the dissemination standards and finally they will be published online or in paper books.

Confidentiality is the latest controlling activity in the dissemination process. In order to maintain the trust of respondents it is the utmost concern of official statistics, to secure the privacy of data providers (like households or enterprises) by assuring that no data is published that might be related to an identifiable person or business. At the same time this guarantees quality by avoiding loss of accurate data. Confidentiality protection is supposed to be implemented on each level of the statistical process from the preparation of surveys up to the dissemination of statistical products.

5. Benefit analysis of proposed software

Generally, any changes in the process and methods that cause to any improvement in time-consuming, costs and expenditures and quality of data can be drawn the orientation of managers and decision makers. Process reengineering and analysis of different sections can result to determine the affection portion of each process' level.

In the practical point of view, proposed software can bring a wide range of benefits for both data producers and final users. In the table 6 these improvement and benefits will be classified.

- Improving the data quality and customers' satisfaction
- Considering multi-dimensional data controlling
- Prevent parallel tasks and time wasting
- Proving user-friendly application and more operators' satisfaction
- Shortening the dissemination by improving purification level

6. Conclusion

It is target goal for all statistical offices to provide accurate, reliable, comprehensive, and onetime and with statistical reports all over the world and statistical center of Iran is not as exception. Process engineering and analysis with focusing on the quality improvement resulted to determine crucial sectors. In order to improve the data quality and minimize the data inconsistency, a software proposed and its tools has been illustrated. This software package tries to increase data quality by helping operators and experts to consider all of the effective micro and macro causes simultaneously

and multi-dimensional analysis to better data analysis and efficient interpretation. Also this software can be customized for other statistical surveys and it can be valuable for all experts who are working on manufacturing establishments' statistics in other NSOs.

References

1. Survey documents of manufacturing establishments in SCI, (2015)
2. United Nations Statistics Division, Fundamental Principles of Official statistic, Implementation guidelines, January (2015).
3. Production process of official statistics, Estonian Statistics" (2014)

Use of GIS and statistics in measuring land consumption rate to population growth rate in Egypt

Amal El Sharnoby
CAPMAS, Cairo, Egypt

Abstract

Like many countries around the world Egypt is facing challenges in managing rapid urbanization from ensuring adequate housing and infrastructure to support growing populations, to confronting the environmental impact of urban sprawl, With the adoption of the 2030 Agenda for Sustainable Development (New-York, 25 September 2015), the country members are invited to measure, follow-up and achieve the 169 targets associated with the 17 Sustainable Development Goals (SDGs), One of these Goals: Make Cities and Human Settlements Inclusive, Safe, Resilient and Sustainable (Goal 11). In particular, the monitoring of SDG 11.3 target. This paper explains how to apply GIS (geographic information system) techniques, remote sensing processing and census data to measure one of this target's indicators: 11.3.1; ratio between the land use consumption rate and population growth rate (LCRPGR). The methodology of computing this indicator was applied on four different cities in Egypt as a case study through satellite images and population data for the last two Egypt' census: years 2006 and 2017 to know and understand: How these cities are growing? Where is the majority of urban growth happening, in small towns with populations of less than 500,000 people or big towns? How we track and manage urban growth? Finally does city growth affect regional development, policies, and the infrastructure?

Keywords

SDG; remote sensing; indicator 11.3.1; Settlements; census data

1. Introduction

As the world becomes increasingly urbanized, Egypt like many countries faces growing numbers of slum dwellers, agricultural land infringement, worsening air quality and insufficient basic urban services, transportation and infrastructure.

The Sustainable Development Strategy (SDS): Egypt Vision 2030 follows the sustainable development goals as a general framework for improving the quality of lives and welfare, dealing with three main dimensions; economic, social, and environmental dimensions. One of these Goals: Make Cities and Human Settlements Inclusive, Safe, Resilient and Sustainable (Goal 11).

With the rapid increasing of population in Egypt (the annual population growth is 2.56 %) the monitoring of SDG 11.3 target which aims to «enhance

inclusive and sustainable urbanization and capacity for participatory, integrated and sustainable human settlement planning and management in all countries» is essential.

This paper explains how to apply GIS techniques, remote sensing and census data to measure one of the indicators of this target which is the indicator is 11.3.1 "United Nations, 2015" : ratio between the land use consumption rate and population growth rate (LCRPGR) on three cities with different criteria in Egypt as a case study, Cairo city as an urban city , Zakazek as a rural city in lower Egypt, and New Cairo city as a new developed settlement. This indicator monitors the relation between land consumption and population growth to inform and enable decision-makers to track and manage urban growth and enhances their ability to promote land use efficiency.

2. Methodology

2.1. Concepts and Definitions

- TARGET 11.3: By 2030, enhance inclusive and sustainable urbanization and capacity for participatory, integrated and sustainable human settlement planning and management in all countries.
- Indicator 11.3.1: RATIO OF LAND CONSUMPTION RATE TO POPULATION GROWTH RATE, Indicator category: Tier II
- Tier II: meaning the indicator is conceptually clear and an established methodology exists but data on many countries is not yet available.
- Land consumption Rate (LCR): the annual rate at which cities uptake land for urbanized uses (both built-up and open space demands)
- Population growth rate (PGR): the increase of a population in a given area during a period, usually, 5 to 10 year intervals.

2.2. Study Area

According to metadata information provided by the UN System for 11.3.1 indicator; use the area of reference at the level of the built-up area of the urban agglomeration and should not be as a conurbation of two or more urban areas that are in close proximity.

Egypt has 27 Governorates, four governorates are urban and the rest are rural governorates, In this study I choose the capital of Egypt; Cairo city as an urban, Zakazek city in lower Egypt, Menia city in upper Egypt and New Cairo city as a new development one, see figure (1).

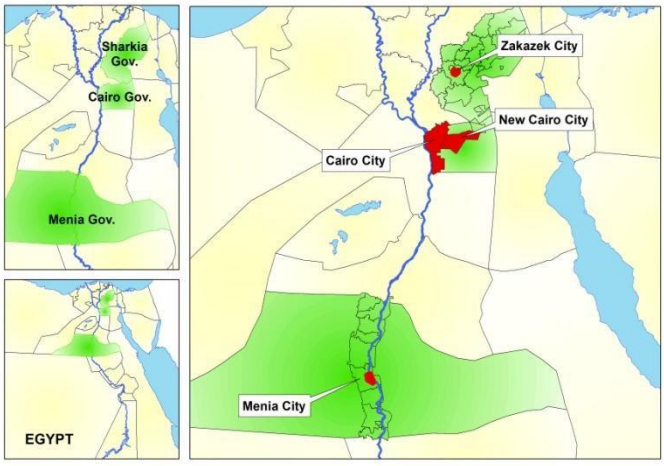


Figure1. Cities of the case study

2.3. Data Sources

Sattelite imagery

To calculate the rate of land consumption, Appropriate downloaded Images were chosen according to the following criteria for years 2006 and 2017:

1- Availability.
2- Adequate resolution.
3- Same periodic time for each year.
4- Similar spectral bands.
5- Full coverage of the study area to avoid image mosaics.

Table 1 shows some metadata of the images used in the study.

Image Info / Year	2017	2006
Landsat Scene Identifier	LC81760392017178LGN00	LT51760392006212MTI00
Acquisition Date	27/06/2017	31/07/2006
WRS Path	176	176
WRS Row	39	39
Ground Control Points Model	417	167
Image Quality	9	9
Map Projection	UTM_Wgs84_Zone36	UTM_Wgs84_Zone36

Table1. Landsat images included in the case study

And figure 2 shows the satellite images of the cities which used in the study.

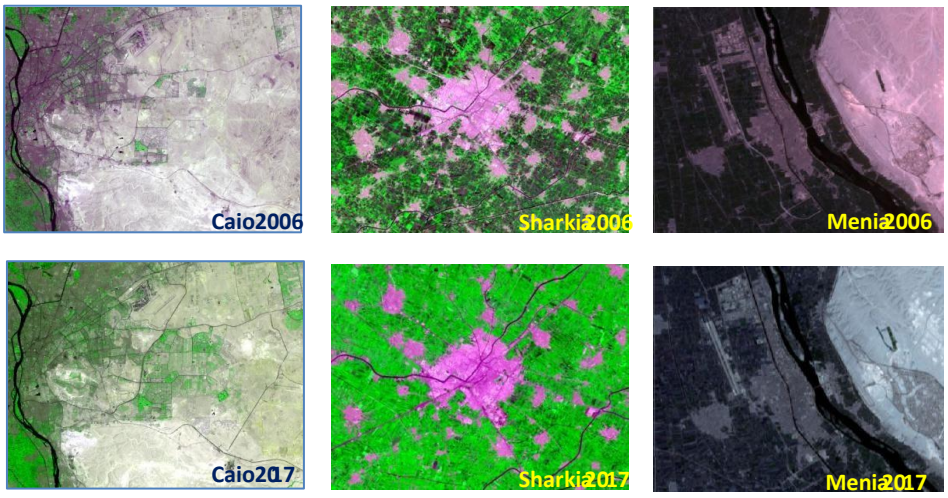


Figure2. Sattelite images of the cities in the case study

Population Data

Egypt population census 2006 and 2017 were used to compute the population growth for the four cities in the level of sub sections (Shiakha and Village).

2.4. Data Processing

The following diagram shows the detailed steps of data processing work flow; these steps are repeated for each city of the case study in the two years 2006 and 2017.

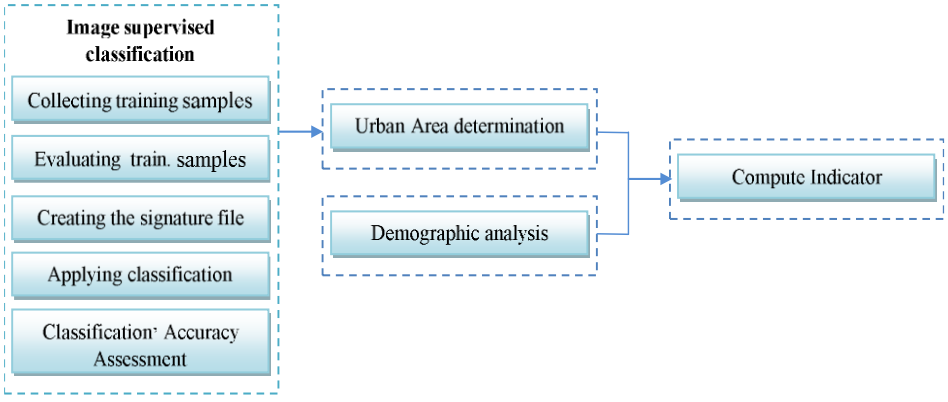
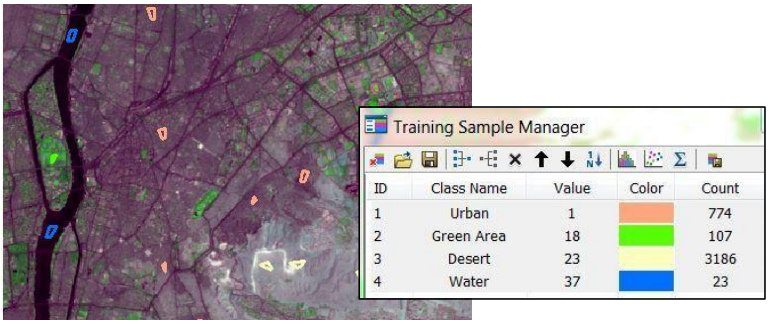


Image classification is the process of sorting pixels into finite number of individual classes or categories of data, based on their pixel values. (B. Bhatta 2010). Classification of images is used to extract information classes from a multiband raster image. The resulting raster can be used to create thematic maps; there are two types of classification: supervised and unsupervised.

In the study I used supervised classification which is based on the training samples collected directly on the image, and then the software classifies the rest of the pixels in the image according to theses samples.

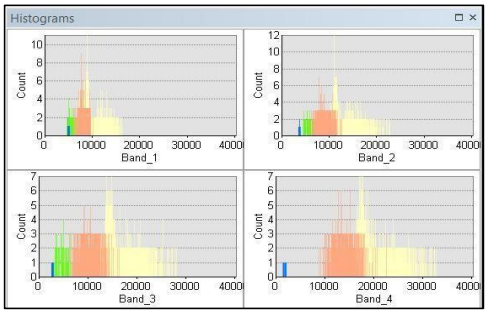
2.4.1 Collecting training samples

In supervised classification, training samples are created to identify classes by draw on the image for each class and calculate their signatures as shown in figure 3.



2.4.2 Evaluating training samples

Examine the Histograms for each class on all the bands to make sure that the classes represented by the training samples are distinguishable and not overlap as shown in figure 4.



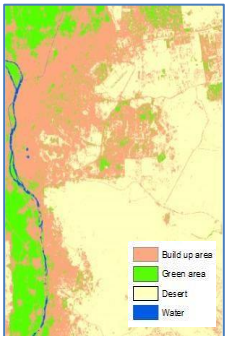
2.4.3 Creating the signature file

After determine the training sample which represent each class and examine the Histogram, a signature file was created to classify the image.

2.4.4 Applying classification

The Maximum Likelihood Classification method used to classify the image; it assigns each pixel to one of the different classes based on the means and variances of the class signatures (stored in a signature file).

Caio & New Cairo 2017 before classification

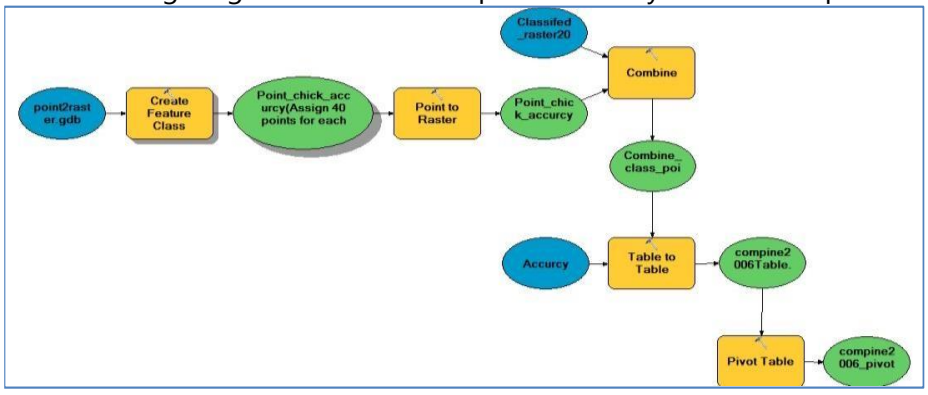


Caio & New Cairo 2017 after classification

2.4.5 Classification accuracy assessment

In the context of information extraction by image analysis, accuracy "measures the agreement between a standard assumed to be correct and a classified image of unknown quality." (Campbell, 2007). Accuracy assessment is an important part of any classification project. It compares the classified image to ground truth data which are derived from interpreting high-resolution imagery. The most common way to assess the accuracy of a classified map is to create a set of random points from the ground truth data and compare that to the classified data in a confusion matrix.

The following diagram shows the steps of accuracy assessment process.



And below is the confusion matrix of one of the cities in the study, this matrix shows that the accuracy rates range from 96.4 to 100 percent accuracy for each class.

Class	N.CAIRO_Urban	N.CAIRO_GreenArea	N.CAIRO_Desert	N.CAIRO_Water
Urban	100	0	3.571428571	0
Green Area	0	100	0	0
Desert	0	0	96.42857143	0
Water	0	0	0	100

Population Computing

Based on the Egypt Census data of the years 2006 and 2017 on the level of household, the population was aggregated on different administrative boundaries levels, in the study I used the aggregated population on level of Shiakha (city) which are the same boundary for the output classified images.

3. **Indicator estimation** (ratio of Land consumption rate to population Growth rate)

Population growth rate

The rate of population growth can be calculated using the following formula:

Population Growth rate (PGR)=LN(Popt_(t+n)/ Popt_t)/((y))

Where

Popt Total population within the city in the past/initial year

Popt+n Total population within the city in the current/final year

y The number of years between the two measurement periods

Land Consumption Rate

The rate of land consumption was calculated from the resulting classified images according to the Following formula:

Land consumption rate LCR=LN(Urb_(t+n)/Urb_t)/((y))

Where

Urb_t Total areal extenoft of the urban in km2 for past/initial year

Urb_(t+n) Total areal extent of the urban in km2 for current year

y The number of years between the two measurement periods

Indicator estimation (ratio of Land consumption rate to population Growth rate)

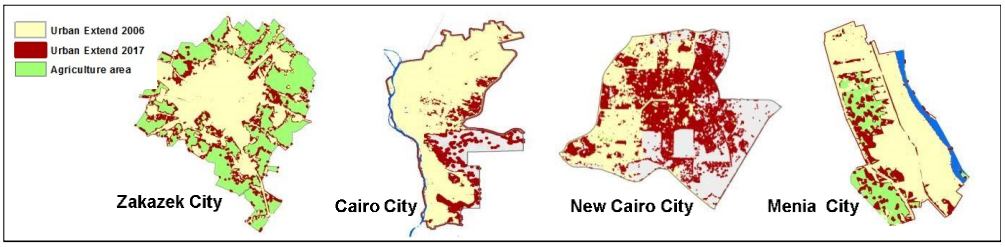
LCRPGR = Land consumption rate / population Growth rate

Results

After the image classification process, and population calculations for years 2006 and 2017. the following table shows the areas and population belonging to the cities in the case study for each year:

Year	Zakazek City		Cairo City		New Cairo City		Menia City	
	Urban (km^2)	Population	Urban (km^2)	Population	Urban (km^2)	Population	Urban (km^2)	Population
2006	21.00	485078	347.79	7740018	23.21	122339	9.24	236043
2017	25.07	609847	392.80	9123702	94.02	273377	10.17	244101

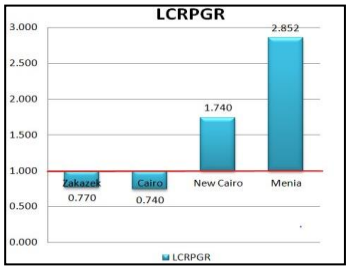
Table2. Urban and population calculations of the cities through 2006, 2017



Urban boundaries of the cities through 2006, 2017

According to the computed values in table2 the following table shows the indicator computedrates

Parameter	Equation	Zakazek City	Cairo City	New Cairo City	Menia City
LCR	$\frac{LN(Urb_{t+n}/Urb_t)}{y}$	0.016	0.011	0.127	0.009
PGR	$\frac{LN(Pop_{t+n}/Pop_t)}{y}$	0.021	0.015	0.073	0.003
LCRPGR	**LCR/PGR**	**0.772**	**0.740**	**1.740**	**2.852**



The results shows that if the indicator (LCRPGR) is greater than one, it means consumption of land is faster than the population growth and if the indicator is less than one revealing Faster population growth than city land uptake. Cairo and Zakazek have indicator below 1, New Cairo and Menia have indicator above 1. According to these results it could be observed that the new development cities as New Cairo have faster growth of land consumption than population, the same for Menia city although it is not a new development city but its population less than 500,00. Also it is observed that the old and big cities with population greater than 500,000 as Zakazek and Cairo, they have faster growth of population than land consumption and this leads to make these cities enlarge vertically causing problems of insufficient basic urban services, transportation and infrastructure.

4. Conclusion

With the rapid increasing of population in Egypt (the annual population growth is 2.56 %) the monitoring of SDG 11.3 target which aims to «enhance inclusive and sustainable urbanization and capacity for participatory, integrated and sustainable human settlement planning and management in all countries» is essential. The paper explained how to apply GIS techniques, remote sensing and census data to measure the indicators: ratio between the land use consumption rate and population growth rate, Indicator 11.3.1 on different cities with different criteria in Egypt as a case study, Cairo city as an urban big city, Zakazek as a city in lower Egypt , Menya as city in upper Egypt and New Cairo city as a new developed settlement. This indicator monitors the relation between land consumption and population growth to inform and

enable decision-makers to track and manage urban growth and enhances their ability to promote land use efficiency. The calculations appear that the old big cities which have more than 500,000 population (Cairo and Zakazek) have indicator less than 1, it means faster population growth than city land uptake, this can lead to suffocated population and insufficient served utilities, while the cities which have population less than 500,000 (Menia) or cities in a new development area (New Cairo) have indicator greater than1 which means the land consumption is faster than population growth.

References

1. CAPMAS, Handbook for users of Census 2017, Census 2006
2. SDSN. (20 de 01 de 2016). Indicators report. http://indicators.report/indicators/i-68/
3. United Nations. (2015). Sustainable Development Goals – The 2030 Agenda for Sustainable Development and the Sustainable Development Goals.
4. UN-Habitat (2012) State of the World's Cities Report: Bridging the Urban Divide, 2012. Nairobi [5]

An application of the generalized Lavallee-Hidiroglou algorithm in business surveys: A sampling methodology for the census of the Philippine Business and Industry

Desiree Robles[1], Mae Abigail O. Miralles[1], Abubakar S. Asaad[1,2]
[1]Statistical Methodology Unit, Philippine Statistics Authority
[2]College of Public Health, Department of Epidemiology and Biostatistics, University of the Philippines Manila

Abstract

This study proposes a sampling methodology for the Census of Philippine Business and Industry (CPBI), an activity undertaken by the Philippine Statistics Authority (PSA), which aims to collect information to constitute bases upon which the government and the private sector can formulate policies and build economic development plans. The proposed sampling methodology uses a stratified single stage design followed by systematic sampling of establishments within strata using total employment as the measure of size for stratification. The sampling methodology was made as such to address the skewness of the population data, an innate property of establishments and the issue on reliability of the survey estimates as indicated by the high values of their corresponding coefficients of variation (CV). The proposed sampling design applies cut-off sampling together with the Lavallee-Hidiroglou (LH) Algorithm in determining the stratum boundaries and optimal sample size at the target CV equal to 5%. The method was applied to the latest available establishment-survey data – Integrated Survey of Labor and Employment and was simulated for 200 times. The reliability of the samples using the proposed methodology was analyzed using the CV of the total employment for each industry classification. Results show that the samples obtained using the proposed methodology was able to meet the specified target CV for each industry classification.

Keywords

Skewed population; Cut-off sampling; Systematic sampling; Simulation

1. Introduction

The CPBI is one of the designated activities undertaken by the PSA. It collects data and generates statistical information such as structure and trends of economic activities from the formal sector of the entire country. From this, the government as well as the private sectors can formulate policies and build economic development plans. The CPBI is conducted every five years of which the last published census was done in 2012. Meanwhile, the latest CPBI,

together with the updating of the list of establishments, commenced on the 16th of April 2018.

The scope of the CPBI is confined only to all existing establishments and enterprises in the formal sector which are engaged in economic activities, as classified in the 2009 Philippine Standard Industrial Classification (PSIC). The rationale for only including the formal sector was based on the share of the group to the value added on the 18 major sections of the Philippine Economy covered by the CPBI.

The results using the current methodology produced high values of CVs at the establishment level. In addition, the distribution of the target variables such as the total employment from the last CPBI, is highly skewed. Hence it is important to develop a sampling methodology which would be able to produce more precise and accurate estimates as well to address the skewness of the data.

2. Methodology

In order to efficiently sample highly skewed population, there is a need to stratify the population into a take-all stratum and a number of take-some strata1. This method was adopted in a business survey where a generalized stratification using LH was proposed2. The proposed methodology accounts for the difference between the discrepancy of the stratification and survey variables. The results of his study showed that using the generalized LH algorithm, the CV of the estimators were close to the target CVs. In relation to this a study was conducted which discussed two approaches in estimating sample size for survey of enterprises in the Philippines3. One of the methods was used is the LH method. In this study, it was mentioned that the CV's of the samples obtained using the LH at the domain level is expected to be small as these are based on the aggregation of the primary strata for which the CV's were controlled. It also added that the main advantage of using LH is that it determines the boundaries as well as the sample size for each stratum.

Aside from stratification of the population, it is important to address the establishments which contribute very small amount to the total of the target variables. An example was the study which used a "cut-off" sampling in order to create a framework for business survey4. This method eliminates the sampling units whose contribution is very small to the total of the target variables.

The previous sampling design of the CPBI involves two method of sampling, one for the establishment and another for the enterprise level, respectively. For the proposed methodology, the sampling design would adapt the cut-off sampling and generalized LH algorithm at the establishment level only.

The proposed sampling design uses a stratified single-stage design with the establishments treated as sampling units. The data used for this study is the list of establishments from the 2018 Integrated Survey of Labor and Employment (ISLE). Meanwhile, the 18 PSIC Sections were considered as the primary strata or domains of the survey. For each domain, four strata were formed which are defined in detail in Table 1.

Table 1. Characteristics of Strata and Sampling Strategies

Stratum	Characteristic	Sampling Strategy
0	Very small establishments whose aggregate TE contributes at most 5% of TE of all establishments.	Take none
1	Smallest stratum formed by the LH algorithm	Take none
2	Medium stratum formed by the LH algorithm	Take none
3	Largest stratum formed by the LH algorithm	Take all

For stratum 0, no establishments will be taken as these only contributes very small amount to the overall total employment (TE) of the establishments. Meanwhile, for strata 1, 2, and 3 the stratum boundaries will be obtained using the LH algorithm.

The optimal sample size will also be obtained using LH algorithm for strata 2 and 3 while all establishments falling within the boundaries of stratum 3 will be taken as samples.

As adapted from the 2011 Survey of Enterprises in the Philippines5, the establishments were sorted according to their region and was followed by systematic selection of samples in order to ensure that samples were distributed across each region. This method would also induce implicit stratification among the regions.

As the dataset used is a survey data, the method of obtaining samples using LH was simulated for 200 times. This is to show that the CV for each of the samples obtained using LH would approach the expected target CV.

3. Results

In order to obtain the optimal sample size and boundaries for each stratum, cutoff sampling together with LH algorithm was applied. The resulting sample size excludes the establishments whose aggregate TE contributes at most 5% of TE for all establishments or those belonging to stratum 0. Hence, the samples consist of establishments belonging to strata 1, 2 and 3 only. The resulting sample size, the percentage of samples and the corresponding CV of TE for each section is shown in Table 2. Meanwhile to further inspect the CV of TE for the samples obtained, the CV for each subsection was obtained and plotted per section as shown in Figure 1.

4. Discussion and Conclusion

Results show that for each industry section, the CV of TE for the samples obtained using the proposed methodology do not exceed the target value of 5% (Table 2). Specifically, the maximum CV for the samples obtained using the proposed methodology is 0.0504 observed at section I or Accommodation and Food Service Activities and 20.3618 observed at section M or Professional, Scientific, and Technical Activities for the original samples. Further, the results show that for the original samples, seven (7) out of the eighteen (18) PSIC sections do not fall within the target CV of 5% while none does not fall above 5% for the samples obtained using the proposed methodology.

The distribution of samples in Table 2 also shows that the samples were distributed with respect to the size of the PSIC section. It is known that section C or Manufacturing section has the highest number of division and subsections among the other sections. Correspondingly, the computed percentage of samples obtained from the LH algorithm for the said section was the highest (12.38%). This is followed by section G or Wholesale and Retail Trade (9.85%) and section N or Administrative and Support Services (9.54%).

The distribution of the CV of TE for each subsection was plotted by section as shown in Figure 1. The results show that for all sections, the corresponding CV of TE per subsection does not exceed the target value of 5%. Although there are outliers present, their corresponding values still reside within the target CV.

Table 2. Distribution of sample size and CV of TE by PSIC Section

PSIC Section		Sample Size(Stratum 1, 2, 3)	Percentange (%)	CV of TV	
				Samples of obtained using Proposed Methodology	Original Samples
A	Agriculture, Forestry and Fishing	74	5.69 %	0.0488	1.2197
B	Mining and Quarrying	28	2.15 %	0.0468	3.8571
C	Manufacturing	161	12.38 %	0.0492	4.4589
D	Electricity, Gas, Steam, and Air Conditioning Supply	39	3.00 %	0.0462	1.2266
E	Water Supply, Sewerage, Waste Management and Remediation Activities	44	3.38 %	0.0492	6.4004
F	Construction	65	5.00 %	0.0464	5.6506
G	Wholesale and Retail Trade	128	9.85 %	0.0496	5.8642

H	Transportation and Storage	73	5.62%	0.0481	0.9994
I	Accommodation and Food Service Activities	79	6.08%	0.0504	3.7540
J	Information and Communications	67	5.15%	0.0475	1.5871
K	Financial and Insurance Activities	67	5.15%	0.0492	4.3055
L	Real Estate Activities	52	4.00%	0.0481	14.3633
M	Professional, Scientific and Technical Activities	58	4.46%	0.0467	20.3618
N	Administrative and Support Service Activities	124	9.54%	0.0474	5.9087
P	Education	106	8.15%	0.0494	3.5339
Q	Human Health and Social Work Activities	76	5.85%	0.0486	1.9080
R	Arts, Entertainment and Recreation	32	2.46%	0.0472	2.9389
S	Other Service Activities	27	2.08%	0.0497	7.7143
Total		1300	100%		

Using the generalized LH algorithm, the target coefficient of variation equal to 5% for each of the PSIC sections was met. Comparing the CV of the samples obtained with that of the original, the obtained CV of the proposed is lower than the original for all sections. Further, none of the samples using the proposed methodology exceeded the target CV of 5% while eight (8) out of eighteen (18) PSIC sections exceeded the target for the original samples. Meanwhile, as the results for this study was computed using a survey dataset, the actual sample size can be larger than estimated from this study. It is then recommended to use the updated List of Establishments for the methodology, when it becomes available.

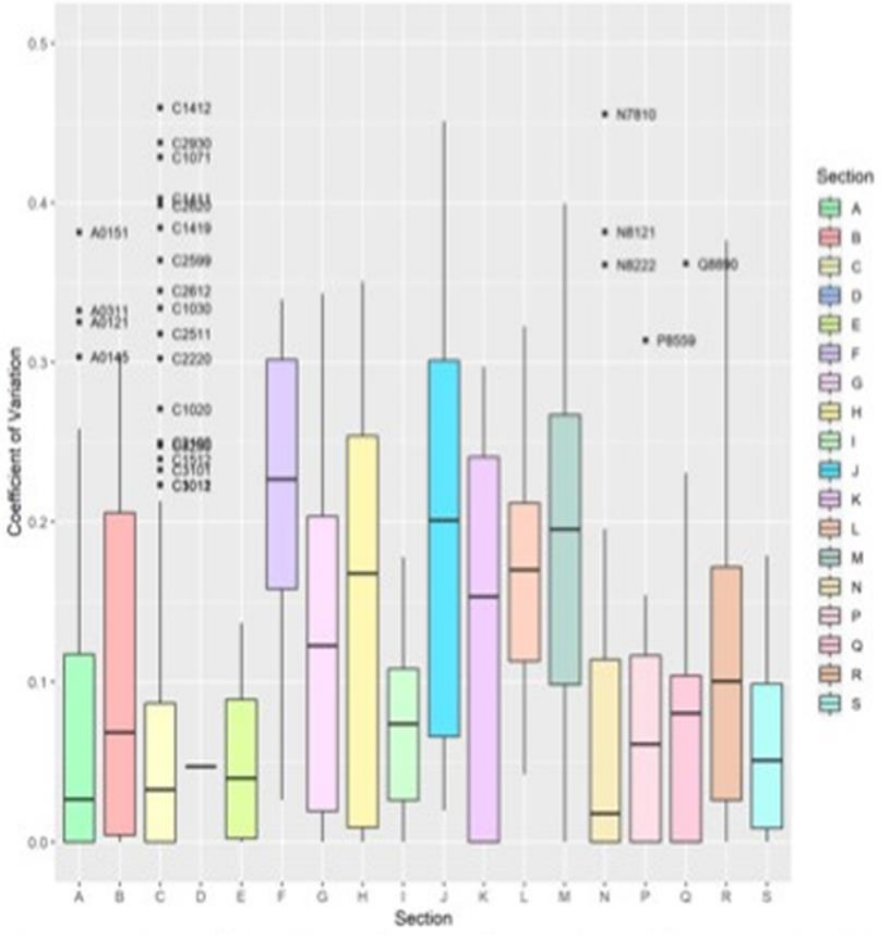


Figure 1. Distribution of the coefficients of variation of total employment for each subsection of the 18 PSIC Section.

References

1. Lavallee, P. & Hidiroglou, M. (1988). On the stratification of skewed populations. Survey Methodology, Vol. 14, 33-43.
2. Rivest, L., (2002). A generalization of the Lavallee and Hidiroglou algorithm for stratification in business surveys. Techniques d'enquete, Vol. 28, 207 - 214.
3. Chouhdhry, G.H. (2018). Sampling design for the survey of enterprises in Philippines. Statistics and Applications, Vol. 16, 45-56.
4. Benedetti, R., Bee, M., & Espa, G. (2010). A framework for cut-o↵ sampling in business survey design, Journal of O cial Statistics, Vol. 26, 651-671.
5. Barrios, E.B. (2012). Sampling from a skewed population: 2011 Survey of enterprises in the Philippines. The Philippine Statistician, Vol. 60, 129-132

Spatial analysis of winter rainfall variability of a selected region in South Africa

Willard Zvarevashe[1], Symala Krishnannair[1], Venkataraman Sivakumar[2]
[1]University of Zululand, KwaDlangezwa, South Africa
[2]University of Kwazulu-Natal, Durban, South Africa

Abstract
Quantitative and qualitative understanding of the multi-scale shifts in the rainfall is very crucial for better preparedness and management. The South Western region of Western Cape, South Africa depends on winter rainfall for its water requirements. However, in recent years there has been water challenges in the region with it being mainly attributed to draught and reduced rainfall. In this study Mann-Kendell and kriging is used to carry out analysis of spatial variability for the winter rainfall. The eleven stations used in the study shows that there is generally monotonic decreasing trend for the months of May and August and increasing trend for June for the period 1980 to 2017. There was also high rainfall variability for May and August and the month of June recorded the least variability.

Keywords
Rainfall; Kriging; Trend Analysis

1. Introduction

South Africa experiences different types of rainfall seasons with most of the country receiving austral summer rainfall and south western parts of the country receiving austral winter rainfall (Dieppois et al., 2016). In recent years the country has been experiencing draughts and sometimes floods, all this adding to the unpredictability of the rainfall patterns. The rainfall trends in South Africa are less uniform with large spatial variability over different regions (Jury, 2012). These extreme rainfall patterns have adverse effects on agriculture, water resources and economy (Niang et al., 2014). The Western Cape province has been facing water challenges recently mainly attributed to decrease of rainfall (Nxumalo, 2017). A study of the rainfall patterns is very essential for future planning.

A study by MacKeller et al., (2014) indicated that there were drier conditions along the southern coastal regions and an increase of rain days in the west coast of Western Cape province. The study was based on the period 1960 to 2010. However, another study for a period from 1921 to 2015 by Kruger and Nxumalo (2017) showed that there is rainfall increase of more than 2.5mm per decade for the June- July-August season for the regions located in the province.

In this paper, an analysis of winter rainfall variability for a selected region in Western Cape, for the period 1980 to 2017 is presented. This is done by firstly using Mann-Kendall Test (MK) to do a monthly analysis of winter rainfall trend for the months May to August. Secondly, calculating the coefficient of variability based on these months and then use kriging to find the spatial spread of the rainfall variability.

2. Methodology

2.1 Mann-Kendell and Coefficient of Variation

Mann-Kendell trend Test (MK) is used to find any trend in the rainfall for the given period. The advantage of MK is that it does not have any underlying assumptions about the distribution of the data except that it must be independent and identically distributed. It can indicate the temporal patterns in a time series (Kendall, 1980). A correlation coefficient, τ where $-1 \leq \tau \leq 1$, denotes relative strength of the trend of the time series is computed. The probability of this trend occurring by chance is also estimated, from which a measure of statistical significance can be assigned.

A 5% level of significance is used such that if the probability estimate was less than 0.05, the trend was deemed to be significant. The trend estimates were calculated for the months May to August for all the 11 stations.

Given the time series $x_1, x_2, x_3, \ldots, x_n$ then the MK statistic, S, is defined as:

$$S = \sum_{i=1}^{n-1} \sum_{j=1+1}^{n} \operatorname{sgn}\left(x_j - x_i\right) \tag{1}$$

where x_j and x_i are the data values in years j and i, respectively, with $j>i$,n is the total number of years and sgn() is the sign function. Let $x_j - x_i = \theta$ then

$$\operatorname{sgn}(\theta) = \begin{cases} 1 \; if \; \theta > 0 \\ 0 \; if \; \theta = 0 \\ -1 \; if \; \theta < 0 \end{cases} \tag{2}$$

The statistics S is approximately normally distributed, with mean zero and variance given by:

$$Var(S) = N(N-1)(2N+5)/18.$$

The standard normal variable Z is then formulated as

$$z = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} if \; S > 0 \\ 0 \; if \; \theta = 0 \\ \frac{S+1}{\sqrt{Var(S)}} if \; S < 0 \end{cases} \tag{3}$$

The coefficient of variation (CV%) is then calculated. CV%, also known as "relative variability", equals the standard deviation divided by the mean. It measures how a value is varied around the mean (Label et al., 1987). Kriging interpolation method is then used to find the spatial distribution of the variation.

2.2 Kriging

Kriging is a stochastic interpolation method which is used to map unsampled locations using the available data sets. It is also known as Wiener-Kolmogorov prediction (Robinson & Metternicht, 2003). Given a point $\mathbf{X}_0$, the ordinary kriging estimator at $\mathbf{X}_0$ based on the data $Z(\mathbf{X}_i) i = 1, \ldots, N$ is defined as the linear unbiased estimator.

$$\hat{Z}(\boldsymbol{X}_o) = \sum_{i=1}^{N} \lambda_i Z(\boldsymbol{X}_i) \quad (4)$$

of $Z(\boldsymbol{X}_o)$with minimum mean square prediction error. Where $\lambda_i \in \mathbb{R}$ is the unknown weights corresponding with the influence of the variable $Z(\boldsymbol{X}_i)$ in the computation of $Z(\boldsymbol{X}_o)$ (Bonaventura & Castruccio, 2005).

3. Results

The average monthly rainfall data was calculated based on the daily rainfall data collected by South African weather service (SAWS) for the period 1980 to 2017 for 11 stations located near city of Cape Town as shown in the figure 1 below. The analysis is based on the months May, June, July and August (MJJA) which is a winter period in South Africa and that is when most of Western Cape receives its rainfall. Rainfall data from weather stations has challenge of having missing data. The imputation method, multiple imputation by chained equations (MICE) also known as multiple sequential regression imputation was used, which assumes missing at random. The R package 'mice' is used for imputation (Van Buuren & Groothuis-Oudshoorn, 2011).

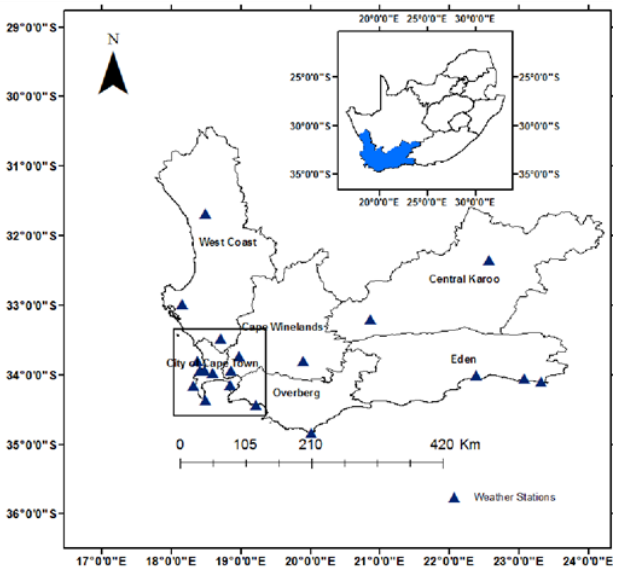


Figure 1. The location of Western Cape province in South Africa, its districts, the location of weather stations and the square area under study.

MK Test results for the stations are shown in table 1 below for the winter rainfall period for the selected stations in Western Cape. The table shows the z-statistic values for each month from May to August and the last column (MJJA) has values for the whole season. Negative z-statistic shows that there

is a monotonic downward trend in the rainfall, whilst a positive z-statistic shows that there is a monotonic upward trend. There is generally monotonic downward trend in all the stations for May with significant trend for Paarl station (p=0.041). However, in June there is generally monotonic increasing trend with significant trend for Hermanus (p=0.025) and Strand stations (p=0.005). There were no significant change in the trend for July and August. The overall seasonal trend (MJJA) for all the stations was not uniform. Cape Town WO has a significant monotonic downward trend (p=0.040), whilst Strand station has a significant monotonic upward trend (p=0.037). The results from MK Test were generally not uniform for the region. In order to find out if the variability is consistent for the region for the months under study, spatial analysis was done.

Table 1. Z-statistic values of seasonal rainfall using MK for Stations in Western Cape Province

Station	May	June	July	August	MJJA
Cape Point	-0.100	0.352	1.018	1.634	0.327
Cape Town Slangkop	-0.451	0.056	0.394	0.480	0
Cape Town WO	-1.938	-0.537	-0.397	-0.350	-1.985*
Hermanus	-1.902	2.264*	1.268	0.604	1.479
Malmesbury	-0.545	0.967	0.074	0.670	0.174
Molteno Reservoir	-1.265	0.099	-0.421	-0.272	-0.868
Paarl	-1.881*	-0.600	-0.664	-0.049	1.508
Robbeneiland	-1.542	0.537	-0.140	0.724	-0.397
S. A. Astronomical Observatory	-1.106	-0.377	-0.653	-0.050	-1.257
Stellenbosch	-1.220	1.132	-0.754	0.565	-0.352
Strand	-0.620	2.454*	1.410	0.338	2.087*

*z-statistic is significant at 0.05 level

CV% was calculated for different stations and kriging was used to find the spatial patterns of the rainfall. The spatial analysis of the Western Cape map is presented in figure 2 below. June is the least variable with CV% of between 40% and 59% and May has CV% of up to 68%. The average CV% for all the months is approximately 50%. May and August are more variable than June and July which are the wettest months. In May most part of the region selected has CV% above 57% whilst the CV% for the greater part of the region is less than 50%.

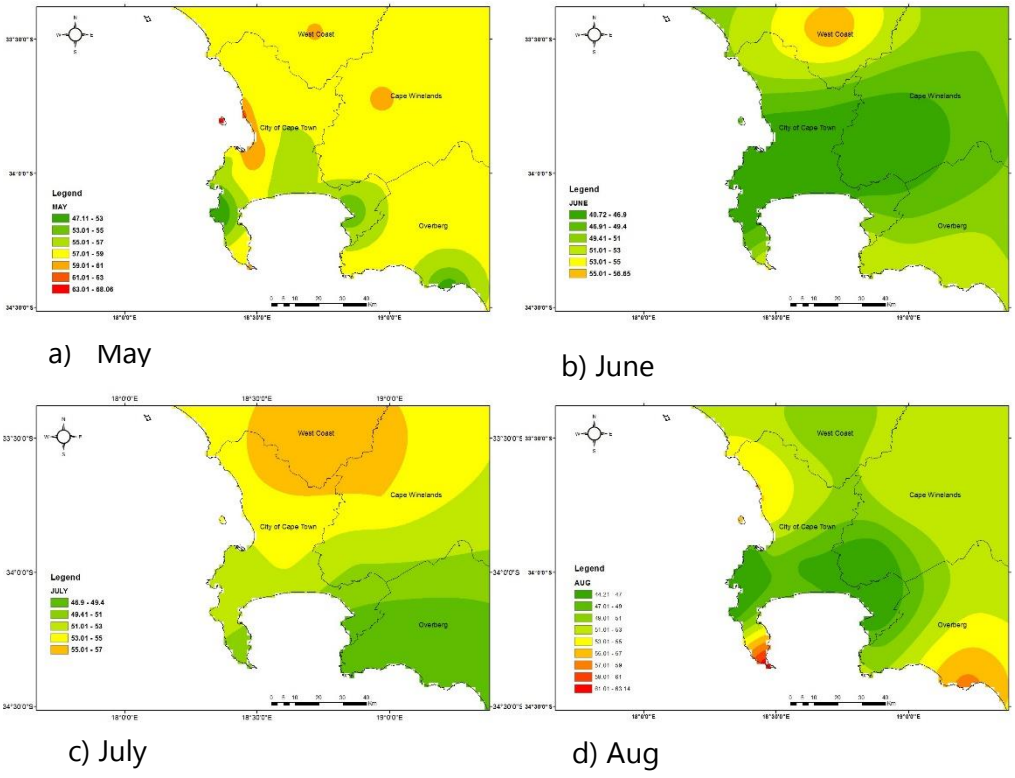


Figure 2. Monthly winter rainfall variability (CV%) for the region in Western Cape for the months a) May b) June c) July and d) August.

4. Discussion and Conclusion

The MK Test revealed that there has been few significant change in the amount of rainfall received for the stations under study from 1980 to 2017. The rainfall for the month of June has increased over the years whilst the months at the beginning and the end of the season have recorded monotonic decreasing trend. It is also interesting to note that June is least variable and the months May and August have higher variability as compared to June. The Hermanus and Strand stations located in Overberg and City of Cape Town respectively, but near to each other, are generally showing a positive trend for June and July. They also have CV% of less than 50% for June and July.

The results from this study have shown that the water challenges that City of Cape Town and surrounding areas are facing in recent times cannot be mainly attributed to reduced rainfall. Some certain months have shown an increasing and less variable rainfall quantity.

Acknowledgement

The author would like to acknowledge South African Weather Service for the data that was used in this study and National Research Fund (Grant number 112979) for the funding of the work.

References

1. Bonaventura, L., & Castruccio, S. (2005). Random notes on kriging: an introduction to geostatistical interpolation for environmental applications.
2. Dieppois, B., Pohl, B., Rouault, M., New, M., Lawler, D., & Keenlyside, N. (2016). Interannual to interdecadal variability of winter and summer southern African rainfall, and their teleconnections. *Journal of Geophysical Research: Atmospheres*, *121*(11), 6215-6239.
3. Easterling, D. R., Evans, J. L., Groisman, P. Y., Karl, T. R., Kunkel, K. E., & Ambenje, P. (2000). Observed variability and trends in extreme climate events: a brief review. *Bulletin of the American Meteorological Society*, *81*(3), 417-426.
4. Hughes, W. S., & Balling Jr, R. C. (1996). Urban influences on South African temperature trends. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *16*(8), 935-940.
5. Kendall, M. (1980). Multivariate Analysis, (London, Charles Griffin).
6. Kruger, A. C., & Nxumalo, M. P. (2017). Historical rainfall trends in South Africa: 1921–2015. *Water SA*, *43*(2), 285-297.
7. Lebel, T., Bastin, G., Obled, C., & Creutin, J. D. (1987). On the accuracy of areal rainfall estimation: a case study. Water Resources Research, 23(11), 2123-2134.
8. Mackellar, N., New, M., & Jack, C. (2014). Observed and modelled trends in rainfall and temperature for South Africa: 1960-2010. South African Journal of Science, 110(7-8), 1-13.
9. Niang, I., Ruppel, O., Abdrabo, M., Essel, A., Lennard, C., Padgham, J., & Urquhart, P. (2014). Africa. Climate Change 2014:. In V. R. Barros, C. Field, D. Dokke, M. Mastrandre, K. Mach, T. Bilir, M. Chatterjee, K. Ebi, Y. Estrada, R. Genova, B. Girma, E. K. A. Levy, S. MacCracken, P. Mastrandrea, & L. White (Eds.), *Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1199–1265).: Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
10. Robinson, T. P., & Metternicht, G. (2003). A comparison of inverse distance weighting and ordinary kriging for characterising within-paddock spatial variability of soil properties in Western Australia. *Cartography*, *32*(1), 11-24.

11. Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M., & Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 45 (3): 1–67.
12. Van Buuren S & Groothuis-Oudshoorn K (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. doi:10.18637/jss.v045.i03

Spatial distribution characteristics of lightning disaster risk and analysis of factors based on GIS--The case of Hohhot

Liu Jia[1,2] , Du Jin Zhu[2], Wang Chun Zhi[2]
[1]Tian Jin University of Finance and Economics
[2]Inner Mongolia University of Finance and Economics

Abstract

Based on the theory of risk analysis, this paper begins with the danger of disaster-causing factors, the exposure of disaster-bearing bodies and the vulnerability of disaster-bearing bodies. The lightning data, geographic information and socio-economic data obtained by the Hohhot lightning locator are used. Taking into account the factors such as ground flash density and ground flash intensity, elevation, terrain fluctuation, land use, population density, etc. the lightning disaster risk assessment index system of Hohhot is constructed, and The index is weighted by the combination of AHP and entropy method. The spatial distribution characteristics of lightning disaster risk in Hohhot are obtained by using ArcGIS spatial analysis technology. The results show that the extremely high-risk area of lightning disasters in Hohhot is the northern part of Xin cheng District, the northern part of Hui min District, and the southwest of Yu quan District. The high-risk areas are Sai han District, Wu chuan County, Tumd Left Banne , Tuo ke tuo County, and Horinger County and Qing shui he County; the densely populated areas such as the townships and towns are all general risk areas. The risk of lightning in Hohhot is quite different. The results of this paper can provide scientific decision-making basis for relevant departments to defend and mitigate lightning disasters and industrial planning layout.

Keywords

Lightning disaster risk; Analysis of Factors; Lightning locator; Spatial Distribution Characteristics

1. Introduction

According to the calculation by the field observation station, the average number of thunderstorm days per year in Hohhot is 36.4d, belonging to the middle thunder area [1]. In order to ensure the normal operation of all industrial departments in Hohhot, we must pay attention to the work of lightning protection and disaster reduction. The meteorological department of Hohhot city has been devoted to research on lightning mechanism and forecasting methods for a long time, which provides strong support for lightning forecasting and disaster prevention and reduction. As far as the current forecast level is concerned, it is still impossible to achieve the accurate

prediction of lightning. When lightning disasters occur, it will inevitably cause damage to various industries and personnel in Hohhot. Therefore, the analysis of lightning disaster factors and the study of spatial distribution characteristics are of great significance to improve the pertinence and efficiency of lightning protection and disaster reduction strategies in Hohhot

2. Analysis of lighting risk factors in Hohhot

2.1.Materials and methods

This paper selects data from three aspects: meteorology, geography and social economy. Based on the risk system theory, this paper divides the lightning disaster risk evaluation system into three levels, as shown in figure 1:

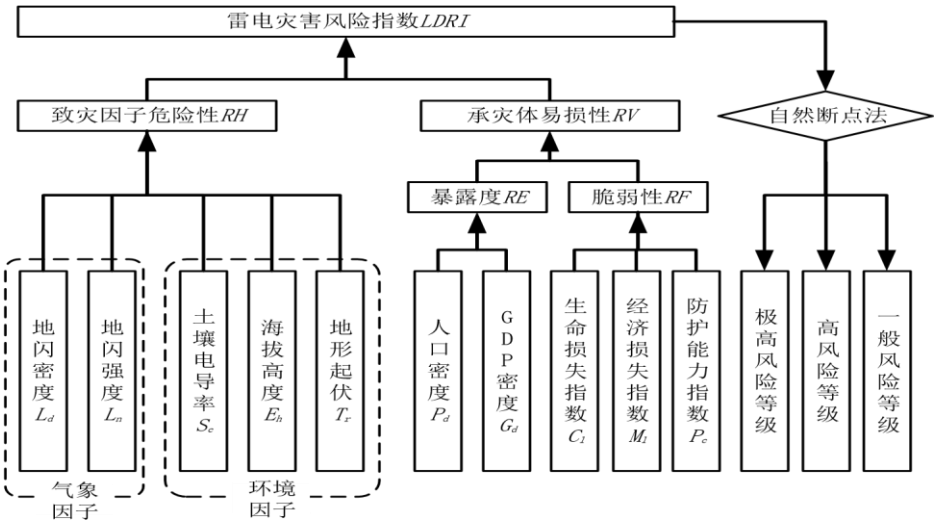


FIG. 1 structure diagram of lightning disaster risk evaluation index system

It can be seen from FIG. 1 that, when the lightning disaster risk index (LDRI) is determined according to the lightning disaster risk evaluation index system, the natural breakpoint analysis function in GIS software is used to classify the regions under the jurisdiction of Hohhot according to the extremely high, high and general three lightning disaster risk levels.

2.2 analysis of lightning risk factors in Hohhot

2.2.1 risk analysis of disaster-causing factors

(1) ground flash density

①Divide the administrative area of Hohhot city, create grid elements according to the "data management" function of GIS software, and divide the whole area according to 3000m units, as shown in figure 2.

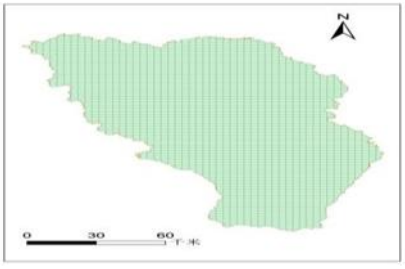


Fig. 2 3KM * 3KM grid diagram

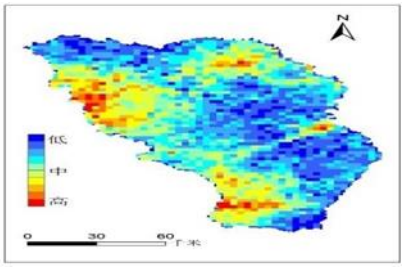


Fig. 3 lightning density vector grid

②The number of ground flashes in each grid can be calculated by superimposing the data of grid layer and lightning strike point, and then by zoning statistics. As shown in Figure 3, the highest density of red part decreases to the lowest density in blue, and the higher the value, the greater the risk.

(2) amplitude of lightning current

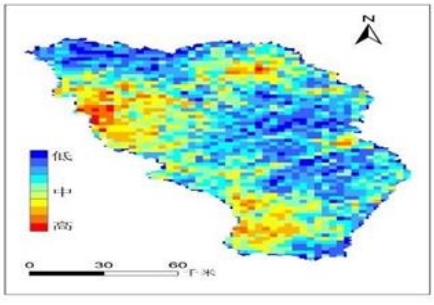
①According to table 1, the magnitude of lightning current is classified.

Table 1 amplitude magnitude of lightning current

Percentile interval	$P \leq 60\%$	$60\% < P \leq 80\%$	$80\% < P \leq 90\%$	$90\% < P \leq 95\%$	$P > 95\%$
Lightening current amplitude	$F \leq 27.90$	$27.90 < F \leq 36.30$	$36.30 < F \leq 47.00$	$47.00 < f \leq 60.70$	$P > 60.7$
Level	Level 1	Level 2	Level 3	Level 4	Level 5

(3) The soil conductivity

Here we take the reciprocal of the soil resistivity soil electrical conductivity for calculation, as shown in figure 5, dark blue is low conductivity area gradually increase to import the green area to red high conductivity area, its value is directly proportional to the size and risk.



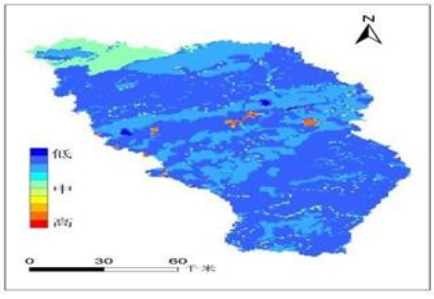


Figure 4 lightning current amplitude vector lattice Figure 5 soil conductivity vector lattice

(4) The DEM data

To compute the statistical data of area known as the processing area, processing area of value and identify the elevation values in all areas of the neighboring domain will be included in the neighborhood statistics

calculation, as shown in figure 6. After the sample size uniform method for classification, calculation results can be divided into level 3, the higher its value, on behalf of the greater the risk.

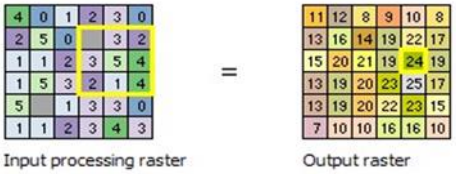


Figure 6 Sketch neighborhood analysis algorithm

In addition, altitude and lightning is inseparable, the digital elevation model (DEM) data extraction can be obtained after region above sea level, as shown in figure 7, as shown in figure 8, the greater the terrain changes, the greater the value of the higher elevation after its weighted, on behalf of the greater the risk.

The figure 3 and figure 4 shows that affected by the ground flash itself stronger area mainly concentrated in the central WuChuan County, tumed left flag in the west, and county and qingshuihe southwest of Middle East. Influence by the flag county center gradually spread to the surrounding.

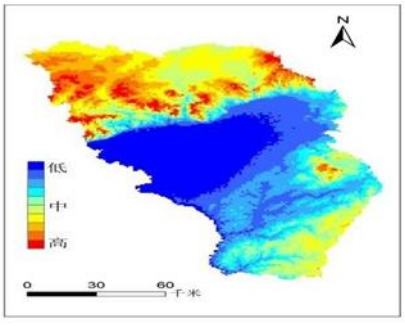


Figure 7 elevation raster data

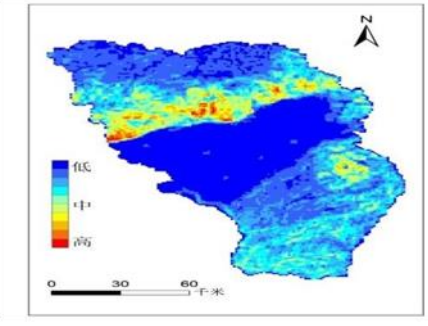


Figure 8 terrain grid data

Figure 5 shows the Hohhot regional soil conductivity high and low, high conductivity region mainly for Hohhot city, tumed left flag, togtoh county and qingshuihe larger region urbanization degree; In addition, pastoral areas and agricultural soil electrical conductivity is higher. Hohhot city, tumed left flag and togtoh county areas less affected by topography and elevation, and other areas are greatly influenced by its, Hohhot especially WuChuan County affected southern and northern Hohhot city is great, as shown in figure 7, as shown in figure 8.

3.3.2 Exposure of hazard-affected bodies The figure 9, figure 10 shows that Hohhot city, togtoh county because of the high population density is bigger,

to both GDP and so its exposure degree is higher, other areas are relatively low exposure.

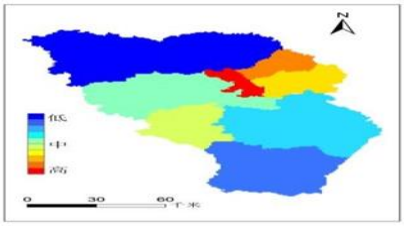


Figure 9 GDP density grid graph

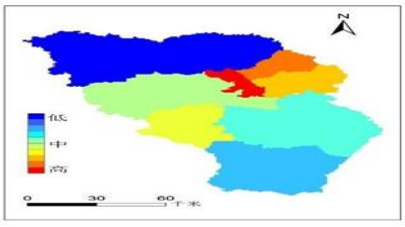


Figure 10 The population density of grid map

3.3.3 Vulnerability analysis of hazard-affected bodies

Figure 11 and figure 12 respectively Hohhot in recent 10 years with thunder and lightning disasters caused economic losses and casualties, the red represents the economic losses and casualties, blue represents the economic losses and casualties.

With reference to the relevant state land use classification standard, Hohhot to the situation of the land use shall be carried out in accordance with the table 2 assignment, raster data form Hohhot regional protection ability index.

Table 2 Protective ability index assignment standard

Land use type	Construction land	Agricultural land	Unused land
Corresponding land Raster data classification	Urban land use, the other Construction land	Cultivated land, forest land, grassland Rural residential areas	Other types other than the construction land, agricultural land use
Protective capability	1.0	0.6	0.5

1:100000 land use data analysis according to the Hohhot, Hohhot belongs to large parts of the cultivated land, forest land, grassland and rural residential type or the orange part of the figure 13, protection value of 0.6. A few parts to the top of the mountain, river and other regions the blue part in figure 13, its protection ability is weak, assignment of 0.5.

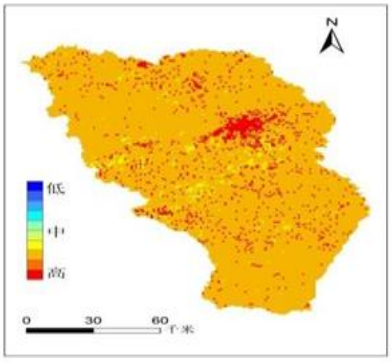


Figure 13 Land use grid graph

4. Hohhot lightning disaster risk evaluation index quantitative and empirical analysis

Through the above steps to Hohhot lightning disaster risk evaluation index quantification and secondary weighted by using entropy value method, invited 18 long engaged in lightning protection device detection, early warning and forecast of thunder and lightning, lightning disaster analysis after the expert level weighted by analytic hierarchy process (ahp), the results are shown in table 5 below:

Table5 Primary weight and secondary weight

	Level 1 weight	Level 2 weight
Lightning disaster risk index	risk 0.4182	Lightning density 0.0181
		The intensity of lightning 0.0140
		Thesoil electrical conductivity 0.1891
		The altitude 0.0394
		The terrain ups and downs 0.0582
	exposure 0.3082	The population density 0.2085
		The density of GDP 0.2135
	vulnerability 0.2736	Index of the loss of life 0.1714
		Index of economic loss 0.0547
		Protective capability indices 0.0331

Combination weights of indicators and secondary indicators in table 5, using the method of natural breakpoint, Hohhot lightning disaster risk index for classification,as shown in table 6:

Table 6 Hohhot spatial distribution of lightning disaster risk rating table

Level	LDRI lightning disaster risk index
High risk level (level I)	2.88<LDRI≤3.08
High risk level (class II)	3.08<LDRI≤3.30
General risk level (level III)	3.30<LDRI≤3.44

Hohhot spatial distribution characteristics of lightning disaster risk results as shown in figure 14. in figure 14, Hohhot north new district, hui nationality area in northern, yuquan area southwest of lightning disaster is extremely high risk area; Abraham area, tumed left flag, WuChuan County, togtoh county, and

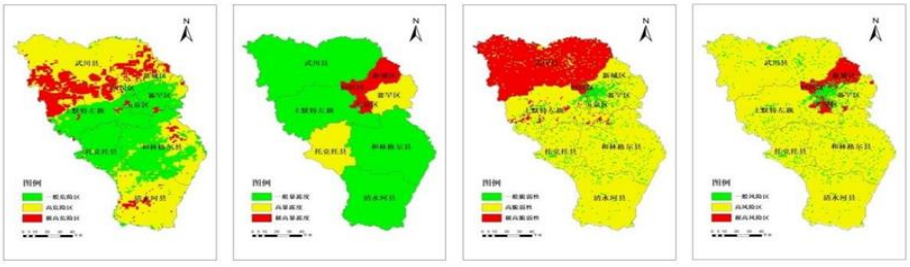


Figure 14 Hohhot characteristics of spatial distribution of lightning disasters figure

Linda lingle county, qingshuihe for high risk area; Township is located around the dense areas of housing, such as the general risk area.

5. Conclusion

With the help of the United Nations humanitarian affairs recommended formula of natural disaster risk in combination with the practical situation of lightning itself and the disaster, according to the actual situation of Hohhot, the spatial distribution characteristics of lightning disaster risk model is set up, it is concluded that the Hohhot north new district, hui nationality in southwest area north, yu quan area for high risk area of lightning disasters conclusion, made clear the key area of Hohhot thunder defence, help to Hohhot for lightning protection and disaster mitigation planning layout, improve the defense capability of the Hohhot thunder and lightning protection and disaster mitigation of management level, to Hohhot to build a new type of intelligent city to provide a strong security.

References

1. Wang Lele Song Haoze,Hou Hao , Liu Xuyang. Hohhot lightning early warning methods in key areas for [J]. Journal of Inner Mongolia meteorological, 2018 (02): 40-43.
2. BLAIKEI P, CANNON T, DAVISI. Risk: Natural hazard,people vulnerability and Disasters[M]. London:Routledge,1994:147-167.
3. Petak W J, Atkisson A A. Natural hazard risk assessment and public policy. Anticipating the Unexpected. New York: Springer, 1982: 18-27.
4. Gerardo Benito Michel Lang, Mariano Barriendos et als. Use of Systematic Palaeoflood and Historical Data for the improvement of Flood Risk.Estimation Review of Scientific Methods.Natural Hazards,2004(31):623-643.
5. Wu Ankun. Risk evaluation and zoning of lightning disasters in guizhou province [J].journal of Chinese agricultural resources and gegional, 2018, 33 (02) 6:88-93.
6. GaoYi, MengXiaoLiang, LaoXiaoQing. Based on the clustering analysis of lightning disaster vulnerability degree of risk zoning hainan island [J]. Journal of natural disasters, 2013, 22 (01): 175-182.
7. Xu Ruili, xiu-hong zhu. Based on GIS lu mountains southeast of lightning disaster risk assessment and regionalization of technology [J]. China agriculture bulletin, 2014, 30 (5): 292-296.

Application of generalised linear models during the phased liberalisation of the Malaysian motor and fire tariffs

Kelvin Hii Chee Yun, Tiffany Tan Shi En
MSIG Insurance (Malaysia) Bhd

Abstract

Bank Negara Malaysia's Phased Liberalisation of Motor and Fire Tariffs policy document, which came into effect on 1st July 2016, implements Bank Negara Malaysia's initiative to deregulate the pricing of Motor and Fire products through the gradual disapplication of requirements under the Motor and Fire Tariff. Under a phased liberalisation environment, general insurers were allowed to determine their premiums by adopting preferred rating factors in the application of statistical and predictive models. The policy document specifies that the board and senior management are expected to leverage on the advice of a professionally qualified actuary in fulfilling their governance requirements and the introduction of new products. Generalised Linear Models is widely accepted as the industry standard for pricing Motor and Fire insurance in many developed markets. In line with these developed markets, the Malaysian actuaries widely adopted Generalised Linear Models as their individual risk Motor model and their individual Fire pricing model during the phased liberalisation period.

Keywords

Phased Liberalisation; Malaysian Motor Tariff; Malaysian Fire Tariff; Actuary; Generalised Linear Models

1. Introduction

Background

Prior to 1st July 2016, general insurance companies in Malaysia determined their Motor and Fire insurance premium based on the Malaysian Motor and Fire Tariffs. Bank Negara Malaysia's Phased Liberalisation of Motor and Fire Tariffs policy document, which came into effect on 1st July 2016, implements Bank Negara Malaysia's initiative to deregulate the pricing of Motor and Fire products through the gradual disapplication of requirements under the Motor and Fire Tariff. The Motor and Fire Tariffs, which has been in place for more than three decades, will no longer be applicable to certain types of Motor and Fire products. This policy document sets out requirements for the gradual disapplication of the Motor and Fire Tariffs with the aim of promoting orderly conditions under a market-based pricing approach.

The policy document requirements are intended to ensure that:

i. the transition to market-based pricing is gradual, and is supported by prudent risk management and practices;
ii. access to basic protection or compulsory lines remain affordable, and any adjustment to the prices of such products is implemented gradually;
iii. consumers' interest remain protected through enhanced transparency and improved sales and marketing practices; and
iv. overall stability in the general insurance market is preserved.

Motor

With effect from 1 July 2016, premium rates for new products will not be subject to the Motor Tariff and shall be determined by the general insurers according to its individual risk pricing model. With effect from 1 July 2017, premium for Comprehensive and Third Party Fire and Theft may be determined by general insurers according to its individual risk pricing model. However, if the premium rates vary from the rates that are applicable on 30 June 2016 by more than 10%, general insurers need to seek Bank Negara Malaysia's approval prior to the implementation of the new rates. Premium rates for Third Party products shall continue to be determined in accordance with the Motor Tariff.

Under a tariff-based pricing regime, general insurers referred to the Malaysian Motor Tariff to determine premiums based on type of cover, geographical location, cubic capacity, and sum insured. There are no explicit driver-related risk factors allowed. The Tariff allowed for further risk-based loadings, however, these are usually not applied.

Fire

With effect from 1 July 2016, premium rates for new products will not be subject to the Fire Tariff and shall be determined by the general insurers according to its individual risk pricing model.

However, with effect from 1 July 2017, premium rates for Fire products which are defined under the Fire Tariff shall continue to be determined in accordance with the Fire Tariff with any variations as may be specified by Bank Negara Malaysia.

Actuaries

Under a phased liberalisation environment, general insurers were allowed to determine their premiums by adopting preferred rating factors in the application of statistical and predictive models. The policy document specifies that the board and senior management are expected to leverage on the advice of a professionally qualified actuary in fulfilling their governance requirements and the introduction of new products.

The responsibilities of the board and senior management includes, inter alia–

(i) the new products and features introduced are appropriately priced using suitable techniques and reliable data;

(ii) the risks associated with the development and pricing of new products are well understood and properly mitigated, which includes robust assessments of the potential impact of product and pricing decisions on profitability and capital adequacy; and

(iii) adequate systems and processes are in place to support risk analysis, pricing, and reserving at a sufficiently granular level.

Generalised Linear Models is widely accepted as the industry standard for pricing private Motor insurance in many developed markets. In line with these developed markets, the Malaysian actuaries widely adopted Generalised Linear Models as their individual risk Motor pricing model during the phased liberalisation period.

2. Methodology

As implied by its name, the Generalised Linear Model is a generalisation of the ordinary linear regression. The linear regression is generalised by allowing the linear model to be linked to the response variable through a link function.

Anderson et al. (2007) summarizes the structure of a Generalised Linear Model as follow:-

$$\mu_i = \mathrm{E}[Y_i] = g^{-1}(\sum_j X_{ij} B_j + \varphi_i$$

where

Y_i is the vector of responses

g(x) is the link function: a specified (invertible) function which relates the expected response to the linear combination of observed factors

X_{ij} is a matrix produced from the factors

B_j is a vector of model parameters, which is to be estimated

φ_i is a vector of known effects

To determine the Motor insurance premium using Generalised Linear Models, the expected claims cost is computed based on a set of rating factors together with its vector of model parameters. Each individual risk profile is reflected through the rating factors, which in turn, translates into an estimated claims cost that the general insurance company is expected to incur.

Broadly speaking, the actuary determines the rating factors to be considered within the Generalised Linear Model. In determining the number of rating factors to be included, the actuary needs to consider the practicality of including such rating factors against the availability of data, customer expectations, and permitted regulations. Based on the selected rating factors,

the Generalized Linear Model computes the corresponding parameters based on the individual characteristics of these rating factors.

3. Results

Motor

A sample hypothetical output of the Generalized Linear Model is shown below:

Base Premium:	$ 850
Rating Factor 1 – Gender	
Male	1.000
Female	0.975
Rating Factor 2 – Age	
Below 25	1.250
26yrs – 30yrs	1.150
31yrs – 35yrs	1.125
36yrs – 40yrs	1.000
41yrs – 45yrs	1.005
46yrs – 50yrs	1.010
51yrs – 55yrs	1.005
56yrs – 60yrs	0.995
Above 60yrs	0.988
Rating Factor 3 – Vehicle Make/Model	
Toyota	1.000
Honda	1.015
Proton	1.045
Perodua	1.085
Others	0.995

Using the above sample hypothetical output, the computation of Motor insurance premium for four risk profile examples are shown below:

Individual 1's risk profile: Male, 40 years old, Honda
Motor insurance premium = $850 x 1.000 x 1.000 x 1.015 = $862.75

Individual 2's risk profile: Male, 21 years old, Perodua
Motor insurance premium = $850 x 1.000 x 1.250 x 1.085 = $1,152.81

Individual 3's risk profile: Female, 62 years old, Toyota
Motor insurance premium = $850 x 0.975 x 0.988 x 1.000 = $818.81

Individual 4's risk profile: Female, 53 years old, Ford
Motor insurance premium = $850 x 0.975 x 1.005 x 0.995 = $828.73

The Generalized Linear Model output is relatively easy to explain to stakeholders. The riskiness of a specific rating factor's characteristics is expressed through the fitted and computed parameters.

Based on the hypothetical examples above

- The lower relativity factor of 0.975 for females compared to males implies that, all else being equal, females are 2.5% less risky compared to males. Hence a 2.5% discount is justified for females relative to males.
- The higher relativity factor of 1.250 for young drivers below 25 years old compared to drivers between 36 to 40 years old implies that, all else being equal, young drivers below 25 years old are 25% more risky compared to drivers between 36 to 40 years old. Hence a 25% loading is justified for young drivers below 25 years old relative to drivers between 26 to 40 years old.

Fire

A sample hypothetical output of the Generalized Linear Model is shown below:

Base Premium: $ 2,500

Rating Factor 1 – Construction Class

Construction Class 1A	1.000
Construction Class 1B	1.050
Construction Class 2	1.250
Construction Class 3	2.500

Rating Factor 2 – Type of properties

Residential	0.450
Retail trading	2.000
Hotel/Offices	1.000
Food processing industries	0.950
Textiles	3.400
Timber	5.250

Using the above sample hypothetical output, the computation of Fire insurance premium for three risk profile examples are shown below:

Building 1's risk profile: Construction Class 1A, Residential
Fire insurance premium = $2,500 x 1.000 x 0.450 = $1,125.00

Building 2's risk profile: Construction Class 3, Timber
Fire insurance premium = $2,500 x 2.500 x 5.250 = $32,812.50

Building 3's risk profile: Construction Class 1B, Hotel
Fire insurance premium = $2,500 x 1.050 x 1.000 = $2,627.10

The Generalized Linear Model output is relatively easy to explain to stakeholders. The riskiness of a specific rating factor's characteristics is expressed through the fitted and computed parameters.

Based on the hypothetical examples above

- The higher relativity factor of 1.250 for Construction Class 2 compared to Construction Class 1A implies that, all else being equal, Construction Class 2 are 25% more risky compared to Construction Class 1A. Hence a 25% loading is justified for Construction Class 2 relative to Construction Class 1A.
- The lower relativity factor of 0.450 for Residential properties compared to Hotel/Offices properties implies that, all else being equal, Residential properties are 55% less risky compared to Hotel/Offices properties. Hence a 55% discount is justified for Residential properties relative to Hotel/Offices properties.

4. Discussion and Conclusion

Generalised Linear Models

One-way analyses are simple to understand and help summarize key explanatory metrics such as average premiums, loss ratios, claim frequencies, and average claims sizes. However, one-way analyses can be distorted by correlations between rating factors. For example, young drivers may tend to drive more affordable car make/models. Hence, a one-way analysis of car make/models may suggest that the affordable car make/models have higher claims experience. However, the underlying reason for the higher claims experience may not be car make/model specific. Rather, the underlying reason for the higher claims experience for the affordable car make/models may be the driver age/experience.

Using Generalised Linear Models would help adjust for correlations and allow for investigation into interaction effects.

Actuaries' role during the Phased Liberalisation of the Malaysian Motor and Fire Tariffs

As the Malaysian general insurance industry transitions into a phased liberalised environment, the board and senior management are expected to leverage on the advice of a professionally qualified actuary in fulfilling their governance requirements and the introduction of new products.

The actuaries have widely adopted Generalised Linear Models as their individual risk pricing model for Motor and Fire products. The outputs of the Generalised Linear Model are easy to explain and implement into information technology systems.

Practicality

The selection of rating factors is a balance between practicality and accuracy. Adopting more rating factors would help refine the accuracy of the insurance premium but at the risk of customers being discouraged by the number of required questions. Adopting less rating factors would be practical for business purposes but introduces the risk of adverse selection due to inadequate information.

In addition to the technical premium computed by the actuary, due regard also needs to be given to market competition and customer retention elasticity.

The application of the +/-10% cap and cup for Motor insurance products helps ensure a gradual transition to market-based pricing, however, it may introduce scenarios where different risk profiles are being charged the same premium (i.e. at the cap/ cup).

New products

Innovative general insurance companies have taken the lead to introduce various new products as allowed under the Phased Liberalisation of Motor and Fire Tariffs policy document.

The policy document specifies new products as

(a) a new type of Motor or Fire cover (inclusive of any add-on cover), that is not defined under the Tariffs; or

(b) any Motor or Fire product, which incorporates new or additional features or components not defined under the Tariffs; or

(c) any variation to, or extension of any cover (inclusive of any add-on cover) defined under the Tariffs.

The introduction of new products, such as smart key coverage, limited special perils coverage, and e-hailing coverage, gives customers a wider choice to tailor their insurance solutions to meet their specific individual needs.

A new product can only be introduced to the market after the policy wording for the new product has been approved by the relevant industry association.

Bank Negara Malaysia's approval is required for Motor new products with the following features:

a) new exclusions, limits, or other terms resulting in reduced coverage compared to products defined under the Motor Tariff;

b) term or duration of cover that is shorter or longer than a continuous period of 12 months;

c) advanced innovations including usage-based insurance and complex experience rating structures beyond existing Motor Tariff, No-Claims-Discount or loading structures;
d) any benefits relating to the policy owner's No-Claims-Discount status or amount; and
e) products where premium rates deviate from the current Motor Tariff rates for the Comprehensive, Third Party Fire and Theft, or Third Party products, by more than 10%, or any other limit subsequently specified by Bank Negara Malaysia.

Bank Negara Malaysia's approval is required for all Fire new products, with the exception of:

a) Large and Specialised Risks; and
b) Products where the premium rates for the Fire component of the deviate from the current Fire Tariff by less than 30%, or any other limit subsequently specified by Bank Negara Malaysia.

References

1. Bank Negara Malaysia. (2016). Phased Liberalisation of Motor and Fire Tariffs. Bank Negara Malaysia (BNM/RH/PD 029-8).
2. Anderson. D., Feldblum. S, Modlin. C, Schirmacher. D., Schirmacher. E., and Thandi. N.. (2007). A Practitioner's Guide to Generalized Linear Models. CAS Study Note, pp. 4-39 only.

Refugees in labor market: Do they act as a marginalisation tool? A decomposition of the wage gap in Palestine

Rabeh Morrar[1,2]
[1]An-Najah National University, Palestine
[2]Northumbria University, UK

Abstract

Given the importance of equal opportunities for poverty reduction, inclusive development and the integration of refugees, we will use the data from the more recent Palestinian Labor Force Survey (PLFS) conducted in the West Bank and Gaza Strip in 2016 to examine, the structure and the level of the income inequality including an empirical decomposition (Oaxaca–Blinder Decomposition) of the refugee and non-refugee disparities. For the decomposition of this inequality, we will use the quantile decomposition approach, based on the Recentered Influence Function (RIF) regression, recently proposed by Firpo, Fortin, and Lemieux (2009). This approach allows us to examine the main origins of the observed inequalities among the population subgroups (refugees and residents) taking into consideration a group of demographical variables (gender, education, place of living, age, etc.). It investigates more specifically how the disparities in the distribution of household features and in the returns to these features contribute to this inequality. This would undoubtedly shed more light on the role and effectiveness of current development and integration policies conducted by the authorities and others institutions in favor of the refugees. The study found non-refugees earn 17% more wages than their refugees counterparts. The composition effect explained by differences in productivity characteristics presents 8.01% of the mean wage gap, while the discrimination effect explains about 9% of the mean wage gap, so the highest percentage of wage differential is due to discrimination effects. The quantile decomposition reveals that the wage differentials are found to be much larger at higher and lower deciles than at the middle part of wage distribution. Also, the only composition effect is found to drive the overall wage gap at the first quantile; no significant effect of the discrimination part is observed at this part of distribution.

Keywords

Oaxaca–Blinder Decomposition; RIF-OLS regressions; Inequality; Refugee; Palestine

1. Introduction

A massive forced migration has been occurred in several periods during the recent Israeli-Arab wars. The largest group of Palestinian refugees originate from more than 500 cities, towns and villages located in Mandate Palestine during the first war in 1948, while a little number of Palestinians remain internally displaced and get the Israeli citizenship (Abu Sitta, 2000). The second largest group of Palestinian refugees were forced displaced during the second Israeli-Arab war in 1967 originated mainly from the West Bank and the Gaza Strip. In this period, huge number of the Palestinian refugees displaced in 1948 to these areas were displaced for a second time, while only few numbers of Palestinians were internally displaced because of this second war (IDMC, 2006; UNHCR, 2006). The most recent displacement were the forced migration of a third largest group of Palestinian refugees that comprise those displaced from the West Bank, East Jerusalem, and the Gaza Strip since 1967 due to Israel's protracted military occupation (IDMC, 2006; UNHCR, 2006).

The situation of the Palestine refugees was classified by the United Nations as one of the most protracted cases of forced displacement in the world today (UNHCR, 2006). Many of camps shelters suffer from unhealthy conditions, lake of safety, and poor construction of the barracks which creates very high temperatures in summer and freezing conditions in winter (Hanafi, 2009). In a survey conducted by Shaml Center (2003), two-thirds of who living in camps felt that their home was too small for their families, half felt that the camps do not meet their basic needs, and 57 percent stated that the camps lacked proper health conditions. Pierre Krähenbühl, Commissioner-General for the United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA)[1] addressed for the fourth Committee in the general assembly (21st meeting) of the UNRWA in 2015 that the Palestine refugees today feel further than ever 'left behind', and remained among the most marginalized due to Israel's ongoing occupation and the absence of sustainable, predictable funding. He added, "The vulnerability and isolation of the refugees had intensified, and broader gains in social and economic development across the region were very much at risk", and "they remained casualties of the unresolved conflict that had violated their fundamental human rights for more than six decades". In essence, due to the continuing occupation, absence of sustainable and predictable funding resources for refugee camps, and the unanticipated resolution for the refugees' problem in the near future, camps remain among the most marginalized localities in the Palestinian territories, among which increasing inequality and marginalisation exist. Therefore, we expect substantial difference in exist in Palestine between those who are camp dwellers and those who live in other urban and rural areas in terms of

[1]https://www.un.org/press/en/2015/gaspd599.doc.htm

socioeconomic status, living conditions, and quality of life. This marginalization often linked to exclusion and violence. In this article, we examine the structure and the level of the income inequality including an empirical decomposition of the refugee non-refugee disparities. More specifically, we will investigate how the disparities in the distribution of household features and in the returns to these features contribute to this inequality. This would undoubtedly shed more light on the role and effectiveness of current development and integration policies conducted by the authorities and others institutions in favor of the refugees.

2. Methodology

Using the basic Oaxaca and Blinder decomposition technique (Blinder, 1973; Oaxaca, 1973), wage difference equations will be estimated for refugee workers and non-refugee workers separately. To explain, suppose the mean log wage function for each group (2 groups) is described by the subsequent equation:

$$E(Y_G|X_G) = X_G\beta_G \qquad (1)$$

where Ydenotes the logarithmic real hourly wages, X is the vector of general (i.e. age, gender, education, marital status, experience, and residence) and labor market characteristics (i.e. occupation, sector of activity) (including the constant term), β is the vector of coefficients and G denotes the group of workers: refugee and non-refugees in labor market. Then the OLS estimate of β_G assesses the impact of X on the conditional or unconditional mean of Y for group G. It is noteworthy that the Oaxaca–Blinder decomposition has been widely used to decompose the mean wage gap between two opposite groups (initially between male and female groups) into a composition effect explained by differences in productivity features and an unexplained wage structure effect due to different returns to covariates. Accordingly, the mean log wage gap between non-refugee (G) and refugee ($\bar{G}$) workers can be written as follows:

$$\bar{Y}_G - \bar{Y}_{\bar{G}} = (\bar{X}_G - \bar{X}_{\bar{G}})\hat{\beta}_G + \hat{X}_{\bar{G}}\left(\hat{\beta}_G - \hat{\beta}_{\bar{G}}\right) \qquad (2)$$

Where $\hat{\beta}_G$ is the reference wage structure, and $(\bar{X}_G - \bar{X}_{\bar{G}})\hat{\beta}_G$ is the composition effect and $\hat{X}_{\bar{G}}\left(\hat{\beta}_G - \hat{\beta}_{\bar{G}}\right)$ represents the wage structure effect (discrimination effect).

Notwithstanding its usefulness in explaining whether differences in wages between different population sub-groups are due to variations in characteristics between them or alternatively due to the wage structure, the Oaxaca-Blinder decomposition method is recently criticized for considering only the decomposition of the mean wage differences, yielding an incomplete representation of the inequality sources. Accordingly, other conventional methods have extended the decomposition beyond the mean and allow the investigation of the entire distribution, yet they all share the same weaknesses

in that they entail a set of assumptions and computational issues (Fortin, Lemieux, & Firpo, 2010). In this regard, the Recentered Influence Function (RIF) regression approach recently suggested by Firpo, Fortin, and Lemieux (2009) addresses these weaknesses and provides a straightforward regression-based method for performing a detailed decomposition of some distributional statistics such as quantiles, variance, and other statistics. The RIF is the key concept of the unconditional quantile regression, the recently widely used method of decomposition in the recent literature.

For this analysis, RIF (Y, q_τ) is the function of explanatory variables:

$$E(\mathrm{RIF}(Y, q_\tau)|\mathrm{X}) = X\beta_\tau \quad (3)$$

Where q_τ is the τth quantile and β_τ is the vector of parameters associated to q_τ. Because RIF(Y, q_τ) is unobserved in practice, we use the estimated equation:

$$\widehat{RIF}(Y_G, \hat{q}_\tau) = \hat{q}_\tau + \frac{\tau - 1(Y_G \leq \hat{q}_\tau)}{\hat{f}_Y(\hat{q}_\tau)} \quad (4)$$

Where $\hat{f}_Y$ is the estimated marginal density function of Y and I is an indicator function.

After estimating the model in equation (3) for the 10th(lowest percentile) to 90th(highest percentile) quantiles of the population, we use the obtained unconditional quantile regression estimates to decompose the different gaps into a component attributable to differences in the distribution of characteristics (composition effect) and a component due to differences in the distribution of returns (wage structure) as follows:

$$\hat{q}_{G,\tau} - \hat{q}_{\bar{G},\tau} = \overline{RIF}(Y_G, \hat{q}_\tau) - \overline{RIF}\left(Y_{\bar{G}}, \hat{q}_{\bar{G},\tau}\right) = (\bar{X}_G - \bar{X}_{\bar{G}})\hat{\beta}_{\bar{G},\tau} + \bar{X}_{\bar{G}}\left(\hat{\beta}_{G,\tau} - \hat{\beta}_{\bar{G},\tau}\right) \quad (5)$$

It is noteworthy that this RIF-based decomposition permits, after computing both the composition effect and discrimination effect throughout the wage distribution, to divide up the two effects into the contribution of each explanatory variable. Moreover, the issue resulting from the use of categorical predictors can also be straightforwardly resolved using the Yun's method (2005) of normalization.

The empirical analysis is based on secondary data from the 2016 Palestinian Labor Force Survey (PLFS) that is prepared by the Palestinian Central Bureau of Statistics (PCBS). PLFS is available on an annual basis for each year from 1995 to 2016. The Palestinian Labour Force Survey Programme conducts surveys quarterly. The survey provides basic information on the relative size and structure of the Palestinian labour force, and the components of employment, unemployment and time related underemployment.

3. Results

The results of Oaxaca–Blinder decomposition in Table 1 reveals that on average, non-refugees earn 17% more wages than their refugees counterparts. The composition effect explained by differences in productivity characteristics

presents 8.01% of the mean wage gap, while the discrimination effect explains about 9% of the mean wage gap, so the highest percentage of wage differential is due to discrimination effects. The detailed decomposition shown in the same table reveal that general characteristics variables play the main role in both explained and unexplained parts with being respectively about -10% and 17.3% . Occupation dummies, being at the second place, are found to explain 2% of the explained part and 4.33% of the unexplained one.

Table 1: Oaxaca decomposition results

	(1)	(2)	(3)
VARIABLES	Overall	explained	unexplained
Refugees	2.144***		
	(0.00900)		
Non -refugees	2.314***		
	(0.00755)		
difference	-0.170***		
	(0.0118)		
explained	-0.0801***		
	(0.00982)		
unexplained	-0.0894***		
	(0.0107)		
General characreristics		-0.0959***	0.173*
		(0.00833)	(0.103)
Industry		-0.00314	0.0208**
		(0.00570)	(0.00909)
Occupation		0.0189***	-0.0433*
		(0.00418)	(0.0222)
Constant			-0.240**
			(0.105)
Observations	16,953	16,953	16,953
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table 2: Quantile decomposition results

	10th			50th			90th		
VARIABLES	**overall**	**explained**	**unexplained**	**overall**	**explained**	**unexplained**	**overall**	**explained**	**unexplained**
Refugees	1.128***			2.233***			3.080***		

	(0.0181)			(0.0116)			(0.0108)		
Non - refugees	1.363***			2.378***			3.331***		
	(0.0154)			(0.0107)			(0.0133)		
difference	-0.236***			-0.144***			-0.251***		
	(0.0238)			(0.0157)			(0.0171)		
explained	-0.197***			-0.0345***			-0.103***		
	(0.0176)			(0.0117)			(0.0100)		
unexplained	-0.0390			-0.110***			-0.148***		
	(0.0249)			(0.0156)			(0.0180)		
General characteristics		-0.274***	0.0343		-0.05***	0.0724		-0.036***	-0.255
		(0.0163)	(0.242)		(0.0107)	(0.153)		(0.00960)	(0.182)
Industry		0.0467***	-0.0411*		-0.00164	0.0371***		-0.059***	0.0102
		(0.0122)	(0.0215)		(0.00706)	(0.0137)		(0.00745)	(0.0163)
Occupation		0.0311***	-0.00222		0.022***	-0.0344		-0.00739	-0.0857**
		(0.00873)	(0.0526)		(0.00548)	(0.0331)		(0.00587)	(0.0383)
Constant			-0.0299			-0.185			0.182
			(0.247)			(0.157)			(0.185)
Observations	17,264	17,264	17,264	17,264	17,264	17,264	17,264	17,264	17,264
Standard errors in parentheses									
*** p<0.01, ** p<0.05, * p<0.1									

The quantile decomposition results in Table 2. The results of such decomposition reveal some important findings. First, the wage differentials are found to be much larger at higher and lower deciles than at the middle part of wage distribution. The overall wage gap at the 10th and 90th percentile is respectively 23.60% and 25.1%, compared with 14.4% at the median. Second, the results show that only the composition effect is found to drive the overall wage gap at the first quantile; no significant effect of the discrimination part is observed at this part of distribution. More specifically at the first quantile, the general characteristics variables are found to be the main drivers of the wage gap with being 27.4%. The figure is different when turning to the second quantile (median); at this part of the distribution the discrimination effect is found to be the main driver of the wage gap (11% against 3.45% for the composition effect). at this quantile, the Industry dummies contribute to this unexplained part by -3.71%. The difference in magnitude between the discrimination and composition effects is narrowing at the end of the wage

distribution with being 10.3% for the composition effect and 14.8% for the discrimination one. According to the results of decomposition shown in Table 2, the occupation dummies are found to have the highest effect (about 6%) on the explained part of the wage gap while the industry dummies are found to have the largest impact on the unexplained part with being 8.57%.

3. Discussion and Conclusion

Using the standard Oaxaca-Blinder decomposition and the RIF-quantile regression function techniques, we decompose the distributional non-refugees/refugees wage differentials among wage workers in Palestine into composition effects, explained by differences in productivity characteristics, and discrimination effects, attributable to unequal returns to covariates. The Oaxaca-Blinder decomposition shows a notable mean wage gap (about 17%) between non-refugee workers and refugee workers. Around 8% of the gap is attributed to the differences in productivity characteristics between non-refugee and refugee workers, while 9% of the mean wage gap is explained by the discrimination against refugees. This is consistent with what Pierre Krahenbuhl, the commissioner-general of the United Nations Relief and works agency for Palestine refugees in the near east (UNRWA) declared about the status of the Palestinian refugees in the general assembly of the UNRWA in 9 November 2015[2] . He confirmed that the Palestinian refugees in West Bank and Gaza feel left behind, they are still suffering from increasing vulnerability and isolation, and to severe inequalities and discrimination against them from the society they are living in.

However the unconditional quantile decomposition reveal that wage gap between the two groups (refugees and non-refugees) is not uniform throughout the wage distribution, and wage differentials are much higher at the bottom and top than at the middle of the wage distribution. No discrimination effect was traced at the 10th percentile, and the general characteristics of the workers are responsible for the wage gap (27.4%) between the two groups. This means that in low wage jobs both non-refugee and refugees have equal opportunities to access into the labor market, which based on the general characteristics of each group. The discrimination effect become very clear in the median part of the wage distribution. Here we might look to many of the jobs in the SMEs which mainly classified as family business which mostly located in the urban areas in Hebron and Nablus. In the top part of wage distribution, the discrimination gap was narrowed but still the non-refugees have more access to the highest wages' jobs in the Palestinian labor market. Many of the high skills jobs in the private sectors (ICT sector, knowledge intensive business services, etc), universities, international NGOs

[2] https://www.un.org/press/en/2015/gaspd599.doc.htm

requires high skills and competences, therefore refugee workers who achieve the cognitive skills are likely to obtain equal opportunities with non-refugees to access into these jobs.

The industry dummy shows that the lowest salaries were in the transportation, storage & communication sector for both refugees and non-refugees, but the negative wage gap in the first percentile changed to positive sign at the top end of the wage distribution with reference to the same basic occupational category. The same shift was experienced in the commerce, hotels and restaurants sector. For example, in the first percentile the wage for non-refugee and refugee workers in the transportation, storage & communication sector is respectively 89.8% and 101.1% less than that in the base category (elementary occupations), but it becomes significantly positive in the top end of the wage distribution for nonrefugees and refugees with respectively 24.1% and 11.7% more than the basic occupational category. This may reflect the very high growth in the salaries in the communication sector in Palestine which reflects the vast development and fast growth of productivity in the last decade (Morrar et al. 2019). The highest wages was found in the construction sector for both refugees and non-refugees. While we found no wage gap in the first quantile between the workers in the construction sectors and the base sector (elementary occupations), the gap expanded for the second and third percentile of the wage distribution. This is because the low wages in the construction sector is for whom working in West Bank and Gaza, while medium and high salaries basically for the Palestinian workers in the construction sector in Israel. The difference in magnitude between the discrimination and composition effects in the industry dummy is narrowing at the end of the wage distribution with being 10.3% for the composition effect and 14.8% for the discrimination one.

In terms of occupation, workers were penalized the lowest wages when employed in services and as vendors in markets. This is for both refugees and non-refugees. For refugees, the wage gap between workers in services and who work legislators and senior managers expand along the wage distribution. This means that the refugees are rewarded in an increasing rate when working in high-skilled jobs which based on competences and experience of the workers. This is not the case in low-skilled job which many not require high skills and competences. The jobs with highest wages for non-refugees are for workers in craft and related trades mainly in the median and upper part of the wage distribution. This supports the previous literature in Palestine (Hilala and McGrath, 2016 & Morrar et al. 2019) about the importance of vocational education and life-long learning in building human capital and high income generation for youth and vocational education students. For example, Hilala and McGrath (2016) found that graduates in vocational education and training in Palestine were half as likely to be unemployed than their peers, which

reflected in higher personal and household income. Those graduates also were able to provide the financial resources to get married and their independent houses. The very high ratio of graduates in all academic fields in Palestine coinciding with weak favoring for vocational education have increased steadily labor supply in most of high skilled jobs (Fallah, 2016), and with low labor demand (due to insufficient investment and high political risk), this were reflected on low wages and salaries.

References

1. Abu Sitta, S. (2000). The Palestinian Nakba 1948, The Register of Depopulated Localities in Palestine. London: The Palestinian Return Centre.
2. Blinder, S. A. (1973). Wage discrimination: Reduced form and structural estimates. Journal of Human Resources, 8, 436–455.
3. Belal Fallah , B. (2016). Evaluation of the Efficiency of the Palestinian Labor Market, Palestine Economic Policy Research Institute (MAS), Ramallah, Palestine.
4. Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional Quantile Regressions. Econometrica, 77 (3), 953-973.
5. Hanafi, S. 2009 'Governance, Governmentalities and the State of Exception in the Palestinian Refugee Camps in Lebanon' Journal of Refugee Studies 23(2) pp134-159.
6. Hilal, R. & McGrath, S. (2016). The Role of Vocational Education and Training in Palestine in Addressing Inequality and Promoting Human Development, Journal of International and Comparative Education, Volume 5, Issue 2.
7. Internal Displacement Monitoring Centre (IDMC) (2006). Occupied Palestinian Territories: West Bank Wall Main Cause of New Displacement Amid Worsening Humanitarian Situation.
8. Morrar, R., Abdeljawad, I., Jabr, S., Kisa, A., & Younis, M. (2019). The Role of Information and Communications Technology (ICT) in Enhancing Service Sector Productivity in Palestine: An International Perspective, Journal of Global Information Management (JGIM) 27(1).
9. Palestinian Diaspora and Refugee Center (Shaml) (2013). Survey about the Palestinian refugees in Lebanon, Leaded by Sari Hanafi.
10. United Nations High Commissioner for Refugees (UNCHR) (2006). Report of the Special Rapporteur of the Commission on Human Rights, John Dugard, on the situation of human rights in the Palestinian territories occupied since 1967, UN Doc. E/CN.4/2006/29.

Probit model of decent living house ownership determinants in DKI Jakarta Province 2017

Sahda Ratnasari[1], Ema Tusianti[2]
[1]BPS-Statistics Indonesia Kaimana Regency
[2]BPS-Statistics Indonesia

Abstract

Owning a decent living house is important for human wellbeing. Unfortunately, in a big city like DKI Jakarta, housing backlog is one of the social problems because of dense population. According to Indonesian Social Economic Survey 2017, only 48 percent household living in an owned private house. There are many factors contributing to housing ownership in DKI Jakarta Province. This study aims to analyse probability of households own a house by harnessing binary probit regression method. The result shows that household expenditure, gender, age, and employment status of the head of the household, household family members and the presence of children significantly determine access to home ownership. For the government, these factors can be used for a consideration in developing an inclusive housing ownership program. Currently, DKI Jakarta Government only considers income as a variable for the program.

Keywords

Housing ownership; Probit regression; DKI Jakarta

1. Introduction

Every human being needs to maintain their existence. According to Maslow's hierarchy of needs theory (1943), there are five levels of basic human needs, i.e. physiological, safety, love or belonging, self-esteem, and the need for self-actualization. According to Maslow, individuals will try to meet the needs of the lowest, then reach the needs at the next level. Physiological needs in the lowest level of Maslow's theory consist of food, drink, house, oxygen, sleep, and health. One of the physiological needs of humans is a place to live. Unfortunately, the increasing number of population makes the need for house even greater due to limited land area especially in a densely populated urban area like Jakarta.

The results of the National Congress on Housing and Settlements (KNPP) II in 2009 stated that the protection or recognition of the rights of all community members to habitable dwellings was carried out by providing opportunities and choices. On the other hand, Law of the Republic of Indonesia number 1 in 2011 article 5 paragraph 1 declares that the state is

responsible for residential areas where guidance is carried out by the government. It means that the problem of housing as a place of residence is a form of state responsibility to improve the human welfare. In article 6 paragraph 1 is also explained that the function of guidance includes planning, regulation, control and supervision. Furthermore, in 2015-2030 Sustainable Development Goals number 11, inclusive, safe, resilient and sustainable settlements are one of the main objectives of urban development.

The Housing and Urban Development Institute in 2017 explained that at least six main housing problems that must be resolved by the government. The issues cover the low number of home ownership, slum areas, low community self-reliance, lacking purchasing power, the occurrence of large-scale land tenure, and the absence of institutional systems and housing governance (Rinaldi, 2017).

Housing construction is intended to overcome the backlog. Backlog is the gap between houses built with the houses needed by residents (General of Budgeting Directorate, Ministry of Finance, 2015). Backlog is used as one of the indicators in the housing sector to measure the needs of homes in Indonesia as a part of National Medium Term Development Plan (RPJMN).

According to Statistics Indonesia (BPS), in 2015, there were 11.4 million home ownership backlogs, meaning that in 2015 there were 11.4 million households living in a rent house. DKI Jakarta is a province with the lowest level of home ownership in 2017 (48.33 percent). DKI Jakarta is a province with the highest population density in Indonesia. DKI Jakarta has 10.5 million people or population density reaches 15.3 thousand people / km2 (BPS-Statistics Indonesia DKI Jakarta Province, 2015). In 2017 only 48.33 percent of households in DKI Jakarta own their house. It means that more than half of DKI Jakarta's households do not have their own dwellings, but living by rent, free rent, or others. The large number of households that do not own a house if not regulated by government may create problems such as slums (including improper housing) or living in the area that are not supposed to build.

The DKI Jakarta Government has launched "the DP 0 Rupiah Houses Program". This program is intended for residents having a maximum of 7 million rupiah per month. They can buy a house with zero rupiah of down payment. Due to limited land, the housing offered is a vertical residence with the concept of a simple flat owned (rusunami). The program is targeted to be started in 2018 (Thalib, 2016).

The population of DKI Jakarta is a heterogeneous and complex society because it comes from all over Indonesia with different backgrounds, ethnicities, cultures, and mindsets. The target of the Housing Program is lacking if only looking at the income factor. Additionally, according to Governor Regulation number 182 of 2017, the Provincial Minimum Wage (UMP) of DKI Jakarta is only Rp. 3.35 million per month.

Actually, there are some variables affect home ownership. The first variable is accumulative household income. The research of Li (1977), Gandelman (2005), Lauridsen and Skak (2007), Guris, Caglayan and Un (2011), Drew (2014) and Aizawa and Helble (2016) state that income significantly influences the opportunity to own a house. Sykes (2005) on his study revealed that income have a significant effect on female household heads in both the white and black races in the United States. Tan (2008) with the factor analysis method found that income has a strong correlation with housing ownership in Malaysia.

The characteristics of the head of the household have an important role that influences housing ownership. The head of the household is a person who is responsible for household needs. The sex of the head of the household may determine housing ownership. Gendelman (2005), Guris, Caglayan and Un (2011) in their study concluded that men were more likely to become homeowners than that of women. Another characteristic of the head of the household is age Increasing age will increase the chances of owning a home (Li, 1977; Asiedu, 1999; Gandel man, 2005; Constant et al., 2007; Guris et al., 2011; and Drew, 2014).

The education level of the head of the household also may determine the ability of the household to own their own home. The level of education plays an important role in shaping mindset. Sykes (2005) found that education had a significant effect on unmarried female head of household, whereas in the white race it was not significant. Tan (2008) with factor analysis found that home ownership is highly correlated with head of household education. While Skak and Lauridsen (2007) state that education increases a person's chances of owning a home.

The marital status of the head of the household also influences the decision of owning a house. Married couples tend to consider owning a home. This marital status is proven to have an effect on home ownership status according to Gendelman (2005), Constant, Roberts, and Zimmermann (2008), and Guris, Caglayan, and Un (2011). The longer the marriage, the greater the chance to own a house (Skak and Lauridsen, 2007). On the contrary, according to Drew (2014), Lauridsen and Skak (2007) marital status does not affect one's intention to own a house.

Other related research found that the work of the head of the household also affected home ownership. People who already have permanent jobs and income who are also likely to think of owning a home. This is related to the effort that must be made to own the house. Study of Asiedu (1999) found that the type of work affected home ownership in Kumasi, Ghana. This is also in line with the research of Tan (2008) who conducted research in Malaysia. Sykes (2005) found that employment status was not significant for black-headed

households, but significant for the white race. Conversely, Drew (2014) states that work does not affect home ownership.

The larger household members who live together, the more space needed. Houses must be able to meet the needs of all household members. Aizawa and Helble (2016), states that the number of household members is the strongest variable affecting Japanese people to become homeowners than income variable. Likewise with the research of Guris, Caglayan, and Un (2011) which found that the type of extended family and couples without children significantly influenced the decision to own a house. But in the Skak and Lauridsen study (2007) the number of household members did not significantly increase the chances of a family owning a home. Fisher and Jeffe (2003) found that in macro terms, the number of household members did not affect the homeownership rate in a country.

Constant, Roberts, and Zimmermann (2007), Tan (2008), Drew (2014) also show that the existence of children influences the reason to own a home. According to Gendelman (2005), the significant influence is not the presence of children, but the number of children under the age of 18 in the household. On the contrary, Lauridsen and Skak (2007) found that children's existence did not significantly influence the decision to housing ownership in Denmark.

This study, however, tries to elaborate those factors as variables determining housing ownership in DKI Jakarta. By doing so, it can be a reference for the local government to evaluate variable of inclusive housingownership program.

2. Methodology

The data used in this study is sourced from the 2017 National Socio-Economic Survey (SUSENAS) by the BPS-Statistics Indonesia by using 5062 household samples. Probit model is used to analyse the determinants of housing ownership. The dependent variable in this study is home ownership status which is a dichotomous variable which are homeowner and not homeowner. Meanwhile, the independent variable to explain the dependent variable consists of eight variables, i.e. household expenditure, gender, age, education level, marital status, and employment status of the head of the household, number of household members, and the presence of children.

The dependent variable is a categorical dichotomous (binary) variable, so there are 3 models that can be used, namely the logit model, the probit model, and the gompit model (complementary log-log). Logit models and probit models can be used if the dependent variable data tends to be symmetrical, or the amount of data between categories is almost the same. Conversely, the gompit model is used for not symmetrical data (Agresti, 1990).

In this study, the model used is a probit model, because the comparison of data between categories on the dependent variable is quite

symmetrical.The output of the probit model is quite complete, including the regression coefficient, significance of variables, marginal effects, then the probability value which is the inverse of the value Z. The probit model has also been widely used in previous housingownership studies, including Gandelman's research (2005), Constant, Roberts, and Zimmermann (2007), Guris, Caglayan, and Un (2011), and Aizawa and Helble (2016).

3. Results

Overall, 48.33 percent of households in DKI Jakarta have their own homes and 51.67 percent of households occupy non-own homes. This makes DKI Jakarta the lowest region in terms of home ownership in Indonesia (BPS, 2018). From the results obtained, it can be concluded that more than half of the population of DKI Jakarta lives in houses that are not theirs.

The percentage of household owning a house based on many variables can be seen in Table 1. The interesting figures is that woman has higher number owning house than man. Informal worker also has higher probability to own house.

Table 1. Distribution of self-owned housing status based on independent variables

No.	Variables	Category	Percentage	Percentage of households with their own houses
(1)	(2)	(3)	(4)	(5)
1.	Household expenditure	>7.000.000	31	72,5
		≤7.000.000	69	37,7
2.	Sex of the head of the household	Man	83,4	46,1
		Woman	16,6	60,4
3.	Age of the head of the household	≥35	78	57,6
		<35	22	16,3
4.	Educational level of the head of the household	≥ Senior high school	60,2	49,7
		< Senior high school	39,8	46
5.	Marital status of the head of the household	Married or ever married	92,6	51,1
		Single	7,4	16,6
6.	Job status of the head of the household	Formal	55,4	39,5
		Informal or unemployment	44,6	57,2
7.	Number of household members	≥4	52,9	58,1
		<4	47,1	37,7
8.	Presence of children	Yes	60,6	47,7
		No	39,4	49,8

Source: Susenas Kor 2017 (processed)

However, to identify significant variable, a partial test is carried out to determine the effect of each independent variable on the housing ownership status. Partial test results can be observed in column 4 Table 2. An independent variable is said to significantly influence the dependent variable, if the p-value is less than the 0.05 significance level. Only educational level of the household head does not significantly affect housing ownership status in DKI Jakarta in 2017.

Table 2. Results of estimation model of home ownership status using all independent variables

Variables	Coefficient	Standard Error	p|z|	Marginal Effect
(1)	(2)	(3)	(4)	(5)
Household expenditure	0,7418	0,0445	0,0000	0,2441
Sex of the head of the household	-0,2733	0,0542	0,0000	-0,0899
Age of the head of the household	0,8174	0,0623	0,0000	0,2690
Educational level of the head of the household	0,0711	0,0420	0,0900	0,0234
Marital status of the head of the household	0,6372	0,0992	0,0000	0,2097
Job status of the head of the household	-0,3599	0,0407	0,0000	-0,1185
Number of household members	0,3392	0,0489	0,0000	0,1116
Presence of children	-0,2570	0,0496	0,0000	-0,0846
Constants	-1,1228	0,1050	0,0000	

To find out the opportunity for households to own a house, it is necessary to know the value of Z (Equation 1). Therefore, the non-significant variables are not included in the formation of the model, because from the value of Z then the opportunity value will be seen.

Table 3. Results of estimation model of housing ownership status using significant independent variables

Variable	Coefisien	Standard Eror	p|z|	Marginal Effect
(1)	(2)	(3)	(4)	(5)
Household expenditure	0,7616	0,0429	0,0000	0,2508
Sex of the head of the household	-0,2640	0,0540	0,0000	-0,0869
Age of the head of the household	0,8097	0,0621	0,0000	0,2666
Marital status of the head of the household	0,6215	0,0987	0,0000	0,2047

Job status of the head of the household	-0,3472	0,0400	0,0000	-0,1143
Number of household members	0,3394	0,0489	0,0000	0,1118
Presence of children	-0,2522	0,0495	0,0000	-0,0830
Constants	-1,0830	0,1022	0,0000	

The equation of the probability of households to own their own homes based on these independent variables, can be written as follows:

$$Zi = -1{,}0830 + 0{,}7616\ Expenditure - 0{,}2640\ Sex + 0{,}8097\ Age + 0{,}6215\ Marital\ status - 0{,}3472\ Job\ status + 0{,}3394\ Household\ members - 0{,}2522\ Children \quad (2)$$

From Equation 2 above, it can be seen that the expenditure variable, age and marital status of the head of the household, and the number of household members have a positive coefficient. Meanwhile, the variable of the sex of the head of the household, the employment status of the head of the household, and the presence of the child have a negative coefficient.

The sex variable of the head of the household has a negative influence on housing ownership status in DKI Jakarta. The marginal effect is -0.0869, which indicates the opportunity for households with male head of households to be 8.69 percent lower than owning a household headed by a female head. This is in line with Table 1, which states that the level of home ownership in households with female household head is higher than that of male. Kolomatsky (2017) explains that women who are not married are more likely to have a house than men with the same status. Mach (2018) also found similar results. Women have more than 50 percent of all types of property compared to men. Weintraub (2017) states that there are several reasons related to this, including women having a strong desire to own their own homes, women needing more space or wanting a smaller house, and the location of a house closer to work, school or family.

The employment status of the head of the household significantly influences the ownership status of the house. However, marginal effects and coefficients of this variable indicate a negative direction. Based on the marginal effect, formal workers have an opportunity of 11.43 percent lower than owning a house compared to households with a informal worker or unemployment. The possibility of this is due to the fact that even though it has own status, the inhabited house is a family heritage house. However, informal workers are a very potential resource in the economy of DKI Jakarta.

The presence of children also significantly affects home ownership status. Marginal effect generated -0.0830. Households with the presence of children in it have a 8.30 percent lower chance of owning a home than household without children. Previous research also found that the presence of children significantly affected home ownership, but in a positive direction, for example in the Drew study (2014). The opposite happens in the DKI Jakarta area. The

presence of children in the household can be one of the reasons why households delay to have house. This is because the cost of home ownership will compete with the costs of raising children due to the greater need compared to non-child households (Courgeau and Lelie`vre, 1992 in Mulder 2006).

Probability for housing ownership can be obtained by equation (2) which is the value of Z. Furthermore, the value of Z will see the opportunity. The opportunity for home ownership is the area under the Normal curve. With 7 independent variables, each of which has 2 categories, there will be 128 possible types of households that occur. Households that have the greatest chance of owning their own homes are households with expenditures above 7 million rupiah, and having female household heads, aged 35 years and over, married or ever married, working in informal sector, have 4 household members and above, and do not have children in the household.

Conversely, households with the lowest chance of home ownership are households with expenditures below 7 million rupiah having a male head of household, aged under 35 years who are not married, working in the formal sector, having household members less than 4 people, and have children in their households. Households with this category may be the main target in housing related policies. The policy can prioritize households with male household head working in the formal sector and having children, because it proves to be more difficult to own a house.

4. Discussion and Conclusion

Based on the results, several results are not in line with the hypothesis. Households with female household head are more likely to have their own decent house than that of male. Most households with female household head are divorced. In addition, household's heads who work in the informal sector or do not work have a higher probability to have an owned home.

The variables that significantly influence the ownership status of decent homes in DKI Jakarta are household expenditure, gender, age, marital status and employment status of the head of the household, family size and the presence of children.

Households that have the highest chance of owning a decent house are households with expenditure above 7 million rupiah, having a female head of household aged 35 years and over, married or ever married, working in the informal sector, having 4 household member or more and not having children. On the contrary, households with the lowest chance of housing ownership are households with expenditures below 7 million rupiah, having a male head of household, aged under 35 years who are not married, working in the formal sector, and having household members less than 4 people and have children. Based on that condition, the government should prioritize households with

male household head, working in the formal sector and having children, because it is evident that households with these conditions have the least chance of having a house.

Further researchers need to elaborate other variables, such as the area of residence, the source of housing ownership (buy by itself or from inheritance), social capital, accessibility, inflation, and changes in land price levels, because the household's ability to own a house may also be affected by these variables.

Some results in this study are not in accordance with the hypothesis, for example households with female household head working in the informal sector have a higher chance of ownership of their houses. Therefore, an in-depth (qualitative) study of these conditions is needed.

References

1. Agresti, A. (1990). *Categorical Data Analysis*. Toronto: John Wiley and Sons, Inc.
2. Aizawa, T., & Helble, M. (2016). Determinant of Tenure Choice in Japan : What Makes You A Homeowner. ADBI Working Paper Series, No. 625.
3. Asiedu, A. B. (1999). Determinants of home-ownership in Kumasi, Ghana. *Danish Journal of Geography* 99 , 81–88.
4. BPS-Statistics Indonesia. (2018, Januari 30*). Persentase Rumah Tangga menurut Provinsi dan Status Kepemilikan Rumah Milik Sendiri*, 1999-2017 . Available from bps.go.id/statictable: https://www.bps.go.id/statictable/2009/03/12/1539/persentase-rumah-tangga-menurut-provinsi-danstatus- kepemilikan-rumah-milik-sendiri-1999-2017.html
5. BPS-Statistics Indonesia DKI Jakarta Province. (2015). *Profil Kependudukan Hasil SUPAS 2015 Provinsi DKI Jakarta*. Jakarta.
6. Constant, A. F., Roberts, R., & Zimmermann, K. F. (2008). Ethnic Identity and Immigrant Homeownership. *CEPR Discussion Paper No. DP6490.*
7. Drew, R. B. (2014). Believing in Homeownership:Behavioral Drivers of Housing Tenure Decisions. *Joint Center for Housing Studies Harvard University.*
8. Fisher, L. M., & Jaffe, A. J. (2003). Determinants of International Home Ownership Rates. *Housing Finance International*.
9. Gandelman, N. (2005). Homeownership and Gender. *Inter American Development Bank .*
10. General of Budgeting Directorate, Ministry of Finance. (2015). *Peranan APBN dalam Mengatasi Backlog Perumahan Bagi Masyarakat Berpenghasilan Rendah (MBR)*. Jakarta.
11. Guris, S., Caglayan, E., & Un, T. (2011). Estimating of Probability of Home-ownership in Rural and Urban Areas: Logit, Probit and Gompit Model. *European Journal of Social Sciences – Volume 21, Number 3* .

12. Kolomatsky, M. (2017, September 28*). Most Unmarried Homeowners Are Women*. Available from https://www.nytimes.com:https://www.nytimes.com/2017/09/28/realestate/mostunmarriedhomeowners- are-women.html
13. Lauridsen, J., & Skak, M. (2007). Determinants of Homeownership in Denmark. *Discussion Papers on Business and Economics.*
14. Li, M. M. (1977). A Logit Model of Homeownership. Econometric Society , pages 1081-1097.
15. Mach, J. (2018, Juni 25). *Women own more properties than men and other findings from StatCan's housing report*. Available from https://www.lowestrates.ca: https://www.lowestrates.ca/news/women-own-moreproperties- men-statcan-housing-report-25184
16. Maslow, A. H. (1943). *A Theory of Human Motivation*. Calicut, India: Nalanda Digital Library, Regional Engineering College.
17. Mulder, C. H. (2006). Home-ownership and family formation. J *Housing Built Environ (2006) 21 ,* 281–298.
18. Rinaldi, M. (2017, January 19*). Ini Enam Masalah Pokok di Sektor Perumahan Rakyat.* Available from Liputan6.com:https://www.liputan6.com/bisnis/read/2831519/ini-enam-masalah-pokok-di-sektorperumahan- rakyat
19. Sykes, L. L. (2005). A home of her own: an analysis of asset ownership for non-married black and white women. *The Social Science Journal* 42 , 273–284.
20. Tan, T. H. (2008). Determinants of homeownership in Malaysia. Munich Personal RePEc Archive .
21. Thalib, R. (2016, November 12). *Program Hunian Terjangkau dan DP 0 Rupiah*. Available from Jakartamajubersama.com: http://jakartamajubersama.com/program-hunian-terjangkau dan-dp-nol-rupiah
22. Weintraub, E. (2017, Agustus 30). *Why More Single Women Are Becoming First-Time Home Buyers*. Available from https://www.thebalance.com: https://www.thebalance.com/single-women-buy-homes-too-17983

High frequency value-at-risk analysis: An empirical study for IPC Mexican Index

Chin Wen Cheong[1], Liu ChengZhi[2], Jing ShengZhe[2] & Ye ZhiQing[2]
[1]Department of Mathematics, Xiamen University Malaysia
[2]School of Economics and Management, Xiamen University Malaysia

Abstract
This study investigates the dynamic volatility movements and market risk of the high frequency Mexican IPC (Indice de Precios y Cotizaciones) index. Based on the heterogeneous market hypothesis framework, the high frequency 5-minute interval data have been utilized to examine the return and volatility of IPC index. Using high frequency realized volatility and bi-power volatility estimators in the heterogeneous autoregressive model, the IPC Mexican market is found to be in concordance with the investment structure suggested by the heterogeneous market hypothesis. Besides various volatility estimators, the heterogeneous autoregressive model is improved with the enhancement of autoregressive conditional heteroscedasticity effect in order to capture the volatility of the realized volatility. In order to obtain a better forecast, the combination forecasts have been applied using various averaging methods and the forecast evaluations are examined using various forecast loss functions. Finally, the forecasted results are utilized in determining the Mexican IPC stock market risk via the value-at-risk based on normal and heavy-tailed distributions.

Keywords
Heterogeneous market hypothesis; high frequency volatility; value at risk

1. Introduction

Over the last couple of decades, the efficient market hypothesis (Fama, 1998; Malkiel, 2003) in term of market information, has been rigorously investigated using the financial markets data which include closed daily and high-frequency data. Based on the traditional efficient market hypothesis (EMH), new proposed hypotheses such as the heterogeneous market hypothesis (HMH) has been introduced to complement the EMH. The relevant studies for HMH are commonly conducted using high-frequency data which normally included 1-minute or 5-minute interval data. This hypothesis suggested that the financial markets consist of market participants with various duration of investment strategies. The results of combining various investment time horizons have generated the 'seemingly' like long-range

dependence volatility (Cheong and Lee, 2018; Cheong et al, 2017) in some empirical financial market studies.

For this specific study, we have selected the Mexican IPC index which acts as an important indicator to reflect the general and comprehensive performance of the Mexican Stock Exchange (BMV). In addition, as the largest stock exchange in Mexico and the fifth stock exchange in America, BMV plays an irreplaceable role in the financial market. Recently, Horenstein and Snir (2017), Herrerra, et al. (2015) and Torre et al. (2016) have conducted the empirical studies regarding the portfolio planning in this area; besides, Choudhry (1996) and Aggarwal et al. (1999) completed relative researches focusing on the AR-GARCH models. To the authors' information, practical studies about this topic are limited, especially for the highfrequency data of the IPC index.

In our analysis, we use two high-frequency volatility estimators namely the realized volatility (RV) and bipower variation volatility (BV), to re-examine the HMH in the Mexican stock market. Using the Heterogeneous Autoregressive Model (Corsi, 2009) with enhancement of asymmetric ARCH feature, the Mexican Indice de Precios y Cotizaciones (IPC) index is modeled and estimated using the 5-minute data. After evaluating the best forecast model for volatility, we further examine the performances for the individual and average combined forecasts which will be further used in determining the market risk. Volatility usually connects with determining the market risk for investment decision. For the application in finance, the value-at-risk is determined based on the estimation results.

The remaining of this study is arranged as follows: Section 2 explains the formation of high-frequency RV and BV HAR models. Section 3 discusses the value-at-risk determination; Finally, Section 4 summarizes and concludes this research.

2. Research Methodology

The high-frequency Heterogeneous AutoRegressive (HAR) volatility models are based on the heterogeneous market hypothesis concepts. In this study, we use the HAR model with the improvement of asymmetric autoregressive conditional heteroskedastic (ARCH) impact. The specifications for HAR(RV)-TGARCH and HAR(BV)-TGARCH models are formulated as follows:

$$\ln\left(\sigma_{RV,t}^{2,day}\right) = \theta_{RV} + \theta_{RV,d}\ln\left(\sigma_{RV,t-1}^{2,day}\right) + \theta_{RV,w}\ln\left(\sigma_{RV,t-1}^{2,week}\right) + \theta_{RV,m}\ln\left(\sigma_{RV,t-1}^{2,month}\right) + \varepsilon_{RV,t}$$

$$\ln\left(\sigma_{BV,t}^{2,day}\right) = \theta_{BV} + \theta_{RV,d}\ln\left(\sigma_{BV,t-1}^{2,day}\right) + \theta_{BV,w}\ln\left(\sigma_{BV,t-1}^{2,week}\right) + \theta_{BV,m}\ln\left(\sigma_{BV,t-1}^{2,month}\right) + \varepsilon_{BV,t}$$

where $\varepsilon_{\cdot,t}$ follows a TGARCH model in the realized volatility (Corsi et al., 2008) and each of the HAR volatility components can be computed

using the equations $\sigma_t^{2,week} = \frac{1}{5}\sum_{t=1}^{5} \ln(\sigma_t^{2,day})$ and $\sigma_t^{2,month} = \frac{1}{22}\sum_{t=1}^{522} \ln(\sigma_t^{2,day})$.

In order to obtain better forecasted results, we combine several competitive forecasts into a single forecast by forecast averaging methods to improve our forecast results. We use a few commonly used averaging weighting (ω) schemes such as the simple-mean (SM), simple median (SMed) and least squares (LS) approaches.

3. Results and Discussion

This study has selected the Mexico Stock Exchange, which is ranked the second largest in Latin American stocks. The IPC index indicates the BMV overall performance. It is made up of a balanced weighted selection of shares that are representative of all the shares listed on the exchange from various sectors across the economy. In this study, the in-sample data are started from January 2010 and ended at 2015 December (1479 days). The forecast evaluation results are presented in Table 1.

Forecast evaluations

Table 1: Dynamic Forecast Evaluations

Actual: BV	Forecast evaluation		
Forecast method	MAE	RMSE	MAPE
Individual : HAR(RV)-TGARCH	0.00005110	0.00021915	95.14545878
HAR(BV)-TGARCH	0.00004161 *	0.00021773 *,**	56.18827691
GARCH-t	0.00004781	0.00022120	52.51741209 *
TARCH-t	0.00004969	0.00022159	57.40503898
Average : Simple mean	0.00003892 **	0.00021801	42.05743195
Simple median	0.00003900	0.00021835	39.40507510 **
Least-squares	0.00004161	0.00021773 **	56.18828106

Notes: * indicates the smallest (best) value for individual forecast only.

** indicates the smallest (best) value for individual and average forecasts.

Table 1 reports the dynamic forecast evaluations, which consist of 116 days from July 2015 until December 2015 for MAE, RMSE, and MAPE for the four models. Using the dynamic forecast approach, the estimated parameters will be used for the next one-day-ahead forecast. For the overall forecast evaluations among the individual forecasts and average forecasts, it is found that almost all the smallest forecast loss functions are dominant by the average forecasts such as least squares and simple median methods. These

outcomes suggested that the average forecasts gathered all the advantages of each individual forecast and provided the best forecasts by averaging them. In other words, it is worthy to implement the average forecasts in order to obtain a more accurate forecast result.

The value-at-risk determination

For the application in finance, we computed the market risk for the Mexican IPC using the value-at-risk approach. Three student-t models namely the HAR(RV)-TGARCH, HAR(BV)-TGARCH and TGARCH are used for the purpose of comparisons. In this specific evaluation, we only calculated the one-day-ahead forecast and the student-t distributed return is obtained by the AR-TGARCH model. According to the dynamic forecast evaluations in Table 1, we can find that, in most of the times, combination forecasts present better forecast outcomes compared to the individual models. Therefore, the forecasts come from these methods may be more reliable for investors. The overall market risks results are presented in Table 2.

Table 2: Value-at-Risk Determination based on actual BV (Dynamic forecast)

	HAR(RV) - TGARCH	HAR(BV) - TGARCH	TGARCH	Simple mean	Simple median	Least-squares
$\widehat{\sigma_t^2}(1)$	0.00003490	0.00002320	0.00001250	0.00002035	0.00001789	0.00002324
VaR						
5% quantile	-0.00956950	-0.00778527	-0.00569012	-0.00728579	-0.00682484	-0.00779196
5% VaR	-95969.4991	-7785.2711	-5690.1161	-7285.7887	-6824.8422	-7791.9619
VaR						
1% quantile	-0.01477851	-0.01203231	-0.00880755	-0.01126353	-0.01055407	-0.012014261
1% VaR	-14778.5077	-12032.3112	-8807.5509	-11263.5322	-10554.0659	-12042.6092

Note: Value-at-risk calculates with $1 million of capital

4. Conclusion

This study uses a modified heterogeneous autoregressive model with various high frequency realized volatility to re-examine the heterogeneous market hypothesis for the Mexican stock market. The empirical discoveries show that the jump-robust volatility outperformed the standard realized volatility and ARCH-type volatility in the forecast evaluations. For better forecast outcomes, the combination forecasts from three models are used and the forecasts are utilized in determining value-at-risk. In conclusion, the study enhances the literature on market information efficiency analysis especially in the empirical case study of high frequency heterogeneous market hypothesis. The empirical results offer an alternative way to forecast and determining market risk particularly in the analysis of investment portfolio strategy and risk management.

Acknowledgment

This work was supported by Xiamen University Malaysia Campus Research Fund (XMUMRF/2019-C3/IMAT/0006).

References

1. Aggarwal, C. I. & Leal, R., 1999. Volatility in emerging stock markets. The Journal of Financial and Quantitative Analysis, 34(1), pp. 33-35.
2. Cheong, W. C. & Lee, M. C., 2018. S&P500 volatility analysis using high-frequency multipower variation volatility proxies. Empirical Economics, 54(3), pp. 1297-1318.
3. Cheong, W. C., Lee, M. C. & Tan, P. P., 2017. Heterogenous market hypothesis evaluation using multipower variation volatility. Communication in Statistics- Simulation and Computation, 46(8), pp. 6574-6587.
4. Choudhry, T., 1996. Stock market volatility and the crash of 1987: Evidence from six emerging markets. Journal of International Money and Finance, 15(6), pp. 969-981.
5. Corsi, F., 2009. A simple approximate long memory model of realized volatility. Journal of Financial Economics, 7(2), pp. 174-196.
6. Fama, E., 1998. Market efficiency, Long-term returns, and behavioral finance. Journal of Financial Economics, 49(3), pp. 283 - 306.
7. Herrera, F. L., Salgado, R. J. S. & Akec, S. C., 2015. Volatility dependence structure between the Mexican Stock Exchange and the World Capital Market. Investigación Económica, 74(293), pp. 69-97.
8. Horenstein, A. R. & Snir, A., 2017. Portfolio choice in Mexico. Journal of Behavioral and Experimental Finance, Volume 16, pp. 1-13.
9. Malkiel, B. G., 2003. The Efficient Market Hypothesis and Its Critics. Journal of Economic Perspectives, 17(1), pp. 59-82.
10. Torre, O. D. l., Galeana, E. & Aguilasocho, D., 2016. The use of the sustainable investment against the broad market one. A first test in the Mexican stock market. European Research on Management and Business Economics, 22(3), pp. 117-123.

Modeling seasonal epidemic data using integer autoregressive model

Manik Awale
Savitribai Phule Pune University, Pune, India

Abstract

In this paper we attempt to model the epidemic data using a seasonal integer valued autoregressive time series model. A seasonal stationary model is proposed for modeling such data. Various probabilistic and inferential properties of the model are studied. Simulation studies are carried out to see the performance of the parameter estimators and to study the forecasting performance of the model. The model is illustrated with a real data set.

Keywords

Autoregression;Binomial thinning; Coherent forecasting; Surveillance data; Public health.

1. Introduction

Most of the epidemic surveillance data are counts data and hence researchers uses the integer-valued autoregressive time series models for the modeling such type of data. Public health officials collect daily, weekly or monthly data on number of cases of a various diseases. Here, we consider a stationary seasonal model based on binomial thinning operator for epidemic time series data, which is similar to the one introduced by Bourguignon et al. (2016), but with geometric marginal distributions. In seasonal stationary models, current value X_t is regressed on the last sth observation X_{t-s}, where 's' is the seasonal period. All the calculations have been performed using R language for statistical computing (*URL: http://www.R-project.org/*).

2. Seasonal geometric INAR(1) model based on binomial thinning

The integer-valued auto-regressive process of order one with geometric marginal distribution and seasonal period 's', (GINAR(1)s) is defined as,

$$X_t = \phi \circ X_{t-s} + Z_t, \quad t \geq s, \tag{1}$$

where, ' $\circ$ ' is a binomial thinning operator, $\phi \circ X = \sum_{i=0}^{X} W_i$, W_i are i.i.d. as

$$P(W_i = 0) = 1 - \phi = 1 - P(W_i = 1), \quad \phi \in (0,1).$$

Here, $Z_t = U_t M_t, \forall\ t$, with U_t independent of M_t,

$$P[U_t = 0] = \phi = 1 - P[U_t = 1],$$

and $\{M_t\}$ an i.i.d. geometric sequence with

$$P[M_t = j] = (1 - \theta)\theta^j, j \geq 0.$$

Then, the marginal distribution of $\{X_t\}$ is given by

$$P[X_t = x] = (1 - \theta)\theta^j, x \geq 0, 0 < \theta < 1.$$

The probability distribution of Z_t is

$$P(Z_t = z) = \begin{cases} (1 - \theta + \phi\theta), & if\ z = 0 \\ (1 - \theta)(1 - \phi)\theta^z & if\ z \geq 1. \end{cases}$$

The marginal mean, variance and pgf of $\{X_t\}$ are

$$E(X_t) = \theta/(1 - \theta),$$
$$V(X_t) = \theta/(1 - \theta)^2$$

and

$$\Phi_{X_t}(v) = (1 - \theta)/(1 - v\theta), \ |v| < \frac{1}{\theta}.$$

The conditional mean and the variance are

$$E(X_t|X_{t-S}) = \phi X_{t-S} + (1 - \phi)\frac{\theta}{1 - \theta}$$

and

$$V(X_t|X_{t-S}) = \phi(1 - \phi)X_{t-S} + (1 - \phi)\theta(1 + \phi\theta)\frac{1}{(1 - \theta)^2}.$$

The conditional pmf (analogous to conditional pmf in GINAR(1)) is given by

$$P(X_{t=y}|X_{t-S} = x) = \begin{cases} (1 - \phi)\theta^{y-x} \sum_{m=0}^{y} \binom{x}{m} \theta^{x-m} \phi^m (1 - \phi)^{x-m+1} \\ + \binom{x}{y} \phi^{y+1}(1 - \phi)^{x-y}, & if\ y \leq x, \\ (1 - \phi)\theta^{y-x}(1 - \phi)\{\phi + (1 - \phi)\theta\}^x, & if\ y > x. \end{cases} \quad (2)$$

The conditional pgf is given by,

$$G_{X_{t+k}}|X_{t-r}(v) = (1 - \phi^q + v\phi^q)^{X_{t-r}} \left(\frac{1 - \theta + \theta\phi^q(1 - v)}{1 - \theta v}\right), q = \begin{bmatrix} k \\ s \end{bmatrix}. \quad (3)$$

From (3) it can be shown that,

$$G_{X_{t+k}}|X_{t-r}(v) \to G_{X_t}(v) = \frac{1-\theta}{1-\theta v} \text{ as } k \to \infty.$$

If $\{X_t\}$ is a process defined as in (1), then the autocorrelation function of the process $\{X_t\}$ is given by

$$\rho(h) = \begin{cases} \rho^{h/s}, if \text{ h is a multiple of s,} \\ 0, \text{otherwise.} \end{cases}$$

The *k*-step ahead conditional pmf (analogous to GINAR(1)) is

$$P(X_{t+k} = y|X_{t-r} = x) = \begin{cases} (1-\theta-\phi^q+\theta\phi^q)^{\theta^{y-x}} \sum_{m=0}^{y} \binom{x}{m}^{\theta^{x-m}} \phi^{qm}(1-\phi^q)^{x-m} \\ +\binom{x}{y}\phi^{q(y+1)}(1-\phi^q)^{x-y}, \quad if\ y \le x, \\ (1-\theta-\phi^q+\theta\phi^q)\theta^{y-x}(\phi^q+(1-\phi^q)\theta)^x, \quad if\ y > x. \end{cases} \quad (4)$$

where, $q = \left[\frac{k}{s}\right]$. The *k*-step ahead conditional mean and variance are respectively,

$$E(X_{t+k}|X_{t-r}) = \phi^q X_{t-r} + (1-\phi^q)\frac{\theta}{1-\theta} \quad (5)$$

and

$$V(X_{t+k}|X_{t-r}) = \phi^q(1-\phi^q)X_{t-r} + \left(\frac{\theta}{1-\phi^2}(1-\phi^{q-1}-\phi^q+\phi^{2q-1})\right)\mu_z + \frac{1-\phi^{2q}}{1-\phi^2}\sigma_z^2, \quad (6)$$

where,

$$\mu_z = \frac{\theta(1-\phi)}{1-\theta} \quad \text{and } \sigma_z^2 = \frac{(1-\phi)\theta(1+\phi\theta)}{(1-\phi)^2},$$

are the mean and variance of Z_t From the equations (5) and (6), we observe that, as $k \to \infty$ the conditional mean and variance converge to the marginal mean and variance respectively.

3. Estimation of the parameters of GINAR(1)$_s$model

In this section we consider the maximum likelihood and conditional least squares estimation of the model parameters. Conditional maximum likelihood estimators can be obtained by maximizing the conditional log-likelihood function

$$\log L(x_1, \ldots, x_n; \phi, \theta) = \sum_{t=s+1}^{n} \log P(\mathrm{X}_t|\mathrm{X}_{t-s}),$$

where, $P(\mathrm{X}_t|\mathrm{X}_{t-s})$ is given in (2). The conditional least squares estimates of the parameters can be obtained by minimizing the function

$$S_n(\phi, \theta) = \sum_{t=s+1}^{n} \left(\mathrm{X}_t - E(\mathrm{X}_t|\mathrm{X}_{t-s})\right)^2$$

with respect to $\phi$and θ. The differentiation results in the following estimating equations,

$$\sum_{t=s+1}^{n}\left((\mathrm{X}_t-\phi\mathrm{X}_{t-s}-(1-\phi)\frac{\theta}{1-\theta}\right)\mathrm{X}_{t-s}=0 \qquad (7)$$

and

$$\sum_{t=s+1}^{n}\left((\mathrm{X}_t-\phi\mathrm{X}_{t-s}-(1-\phi)\frac{\theta}{1-\theta}\right)=0 \qquad (8)$$

Solving (7) and (8) for ϕ and θ, we get the following estimators,

$$\hat{\phi}_{CLS}=\frac{(n-s)\sum_{t=s+1}^{n}\mathrm{X}_t\mathrm{X}_{t-s}-(\sum_{t=s+1}^{n}\mathrm{X}_t)(\sum_{t=s+1}^{n}\mathrm{X}_{t-s})}{(n-s)\sum_{t=s+1}^{n}X_{T-S}^2-(\sum_{t=s+1}^{n}\mathrm{X}_{t-s})^2}$$

and

$$\hat{\theta}_{CLS}=\frac{\sum_{t=s+1}^{n}\mathrm{X}_t-\hat{\phi}_{CLS}\sum_{t=s+1}^{n}\mathrm{X}_{t-s}}{(n-s)(1-\hat{\phi}_{CLS})+(\sum_{t=s+1}^{n}\mathrm{X}_t-\hat{\phi}_{CLS}\sum_{t=s+1}^{n}\mathrm{X}_{t-s})}.$$

The The conditional least square (CLS) estimator $\widehat{\Theta}_{CLS}=(\hat{\phi},\hat{\theta})'$ of the parameter $\Theta=(\phi,\theta)'$ of INAR(1) model with geometric marginal distribution as defined in (1) has the following asymptotic distribution,

$$\sqrt{n}\begin{pmatrix}\hat{\phi}_{CLS}\\ \hat{\theta}_{CLS}\end{pmatrix}\xrightarrow{d}N\left[\binom{0}{0},\begin{pmatrix}\frac{1-\theta-\theta^2-\phi\theta(2\theta-3+\theta^2)+\phi^2(\theta^3+3\theta^2-2\theta-1)}{\theta} & (1-\theta)(1-\phi\theta)\\ (1-\theta)(1-\phi\theta) & \frac{(1+\phi)(1-\theta)^2\theta}{(1-\phi)}\end{pmatrix}\right].$$

4. Forecast accuracy measures

Suppose the observed data set $\{X_1,\ldots,X_n,X_{n+1},\ldots,X_{n+m}\}$ is partitioned into two sets. The first n observations are used for the estimation of the parameters of the model and rest m observations called the test set is used for the computation of these measures. These measures are defined as,

$$\mathrm{PRMSE}\,(k)=\sqrt{\frac{1}{m-k+1}\sum_{t=n}^{n+m-k}\left(X_{t+k}-\hat{X}_{t+k}^{Mean}\right)^2},$$

$$PMAE\,(k)=\frac{1}{m-k+1}\sum_{t=n}^{n+m-k}\left|X_{t+k}-\hat{X}_{t+k}^{Med}\right|$$

and

$$\mathrm{PTP}=\frac{1}{m-k+1}\sum_{t=n}^{n+m-k}I(X_{t+k}=\hat{X}_{t+k})\times 100\%$$

where, I(·) is the indicator function. In PTP, we have used $\hat{X}_{t+k}=\hat{X}_{t+k}^{Med},\hat{X}_{t+k}^{Mode}$ and the k-step ahead conditional mean, after rounding it to the nearest integer. $\hat{X}_{t+k}^{Mode},\hat{X}_{t+k}^{Median}$ and $\hat{X}_{t+k}^{Mean}$ are obtained from the k-step ahead conditional pmf.

The standard prediction intervals assume the predictive probability distribution to be symmetric. However, for INAR models, it is uni-modal and positively skewed. Hence, we find the $100(1-\gamma)\%$ highest predictive probability interval (HPP) for X_{t+k} as, $C_k = (X_L, X_U)$, with $C_k = \{y : p_k(y|x) \geq k_y\}$ and k_y is the largest number such that,

$$P(X_L \leq X_{t+k} \leq X_U | X_t = x) = \sum_{y=X_L}^{X_U} p_k(y|x) \geq (1-y)$$

where,

$$p_k(y|x) = P(X_{t+k} = y | X_t = x).$$

4.1 Simulation study

The simulation results in Table 1 present the parameter estimates and their mean squared errors associated with a seasonal GINAR(1) (GINAR(1)$_s$) model. From this table, it can be seen that both CML and CLS estimators perform well and get stabilized for larger sample sizes. CML estimates have smaller mean squared error(MSE) than the CLS estimators. Therefore, we have used CML estimates for data analysis. For this study, we have simulated 1000 series each of size 300, 500 and 1000 with parameter values as given in Table 1, with seasonal period s = 52. In this table, the average values of the estimates over 1000 simulations and their corresponding MSE's are given. Table 2 presents the comparison of seasonal GINAR(1) model and non seasonal GINAR(1) model. A simulation study have been carried out to examine the performance of the seasonal as well as non seasonal models for seasonal data. We have carried out 1000 simulations of the series with size 600. Out of these 600 observations first 400 were used for the parameter estimation and remaining for computation of forecast accuracy measures reported in Table 2. It can be observed that the PRMSE and PMAE for the seasonal model are smaller than the non-seasonal model and it is observed for all the parameter combinations. Here, we conclude that the GINAR(1)$_s$ model performs better than the GINAR(1) model in terms of PRMSE and PMAE. Here, we have obtained the coherent forecasts such as median and mode, using the *k*-step ahead conditional distribution given in (4). The traditional mean forecast is obtained using the expression (5).

We have computed the PTP values for all these parameter combinations and for both the models, (see Table 3). PTP for mean is computed using the rounded part of the *k*-step ahead conditional mean. Here also, it can be observed that the PTP for all mean, median and mode are higher in the seasonal model than the non seasonal model. This implies that the proportion of forecasting exact values is higher in the seasonal model than the non-seasonal model. It can also be observed that the model performs much better in terms of forecast accuracy measures (see Table 2 and 3) for relatively higher values of the thinning parameter.

5. Application to Campylobacter infection data

In this section, we consider the application of the seasonal geometric INAR(1) model to the seasonal epidemic data. The Campylobacter infection data has seasonal structure. The data are from Rottweil County of BadenWu¨rttemberg state in Germany. It consist of 312 weekly observations pertaining to 2001 to 2006. The data are available from Robert Koch Institute, Germany (https://survstat.rki.de). From the the ACF plot in Figure 1, it can be seen that the series has seasonality and from PACF plot the AR(1) structure is evident. The mean and variance are 1.2756 and 2.0781 respectively, implying overdispersion. From mean, variance it can be confirmed that the INAR(1) model with geometric marginal distribution would be a suitable model for the data. The suitability of geometric model can be seen from the bar-plot and AIC , BIC computed for the similar models in Table 4.

Table 1: Parameter estimates and their mean square errors.

n	$\hat{\phi}_{cml}$	$\hat{\theta}_{cml}$	$\hat{\phi}_{cls}$	$\hat{\theta}_{cls}$
		$\phi = 0.2, \theta = 0.3$		
300	0.1976	0.2989	0.1915	0.2996
	0.0031	0.0007	0.0048	0.0007
500	0.1959	0.2988	0.1968	0.2990
	0.0019	0.0004	0.0031	0.0004
1000	0.1992	0.2998	0.1987	0.2993
	0.0009	0.0002	0.0015	0.0002
		$\phi = 0.5, \theta = 0.6$		
300	0.4978	0.5978	0.4916	0.5969
	0.0010	0.0009	0.0033	0.0010
500	0.5001	0.5983	0.4953	0.5987
	0.0005	0.0005	0.0021	0.0006
1000	0.4991	0.5989	0.4961	0.5991
	0.0002	0.0002	0.0009	0.0003
		$\phi = 0.9, \theta = 0.5$		

Table 2: PRMSE and PMAE comparison of GINAR(1) and GINAR(1)$_s$ models.

k	GINAR (1)		GINAR $(1)_S$		GINAR(1)		GINAR $(1)_S$	
	PRMSE	PMAE	PRMSE	PMAE	PRMSE	PMAE	PRMSE	PMAE
	$\phi = 0.2, \theta = 0.3$				$\phi = 0.4, \theta = 0.5$			
1	0.7776	0.4282	0.7623	0.4199	1.4047	1.001	1.2882	0.7667
2	0.7776	0.4281	0.7626	0.4200	1.4036	0.9989	1.2876	0.7666
3	0.7774	0.4281	0.7624	0.4199	1.4033	0.9985	1.2873	0.7663
4	0.7776	0.4282	0.7624	0.4201	1.4033	0.9985	1.2873	0.7662
5	0.7774	0.4280	0.7624	0.4200	1.4026	0.9982	1.2865	0.7657
	$\phi = 0.6, \theta = 0.7$				$\phi = 0.8, \theta = 0.9$			
1	2.7882	1.9505	2.2244	1.2330	9.3811	6.6127	5.6197	2.4265
2	2.7871	1.9488	2.2244	1.2328	9.3782	6.6074	5.6204	2.4273
3	2.7867	1.9489	2.2235	1.2323	9.3785	6.6081	5.6213	2.4282
4	2.7864	1.9491	2.2235	1.2321	9.3769	6.6066	5.6184	2.4269
5	2.7861	1.9492	2.2225	1.2321	9.3789	6.6082	5.6213	2.4277

Table 3: PTP comparison of GINAR and GINAR(1)$_s$ models.

	GINAR (1)			GINAR (1)$_S$			GINAR(1)			GINAR (1)$_S$		
	Med	Mode	Mean	Med	Mode	Mean	Med	Mode	Mean	Med	Mode	Mean
k	$\phi = 0.2, \theta = 0.3$						$\phi = 0.4, \theta = 0.5$					
1	70.02	70.04	65.87	69.52	70.08	63.65	38.37	49.77	25.12	52.56	53.93	26.66
2	70.04	70.04	66.31	69.51	70.07	63.66	38.34	49.79	25.13	52.56	53.92	26.67
3	70.03	70.03	66.29	69.51	70.07	63.66	38.39	49.80	25.14	52.57	53.93	26.67
4	70.03	70.03	66.29	69.50	70.06	63.65	38.37	49.79	25.15	52.57	53.93	26.68
5	70.03	70.03	66.30	69.51	70.07	63.66	38.38	49.80	25.15	52.59	53.94	26.68
	$\phi = 0.6, \theta = 0.7$						$\phi = 0.8, \theta = 0.9$					
1	18.52	29.74	13.47	45.33	46.06	19.45	5.42	9.99	3.98	38.36	38.36	7.75
2	18.51	29.89	13.48	45.34	46.07	19.46	5.42	10.02	3.98	38.36	38.35	7.74
3	18.52	29.89	13.48	45.35	46.07	19.46	5.42	10.04	3.98	38.36	38.35	7.75
4	18.51	29.89	13.47	45.35	46.07	19.46	5.42	10.04	3.99	38.36	38.36	7.75
5	18.50	29.89	13.47	45.34	46.06	19.46	5.41	10.04	3.98	38.37	38.36	7.76

Table 4: Model selection using AIC and BIC.

Model	Parameter Estimates	AIC	BIC
GINAR(1)	$\hat{\phi}$=0.1197, $\hat{\theta}$=0.5508	969.1888	970.1771
GINAR(1)$_S$	$\hat{\phi}$=0.0050, $\hat{\theta}$=0.5595	814.7428	815.7311
NGINAR(1)	$\hat{\phi}$=-=0.2432, $\hat{\theta}$=1.2211	964.5258	965.5141
NGINAR(1)$_S$	$\hat{\phi}$=0.0058, $\hat{\theta}$=-=1.2703	814.7420	815.7303

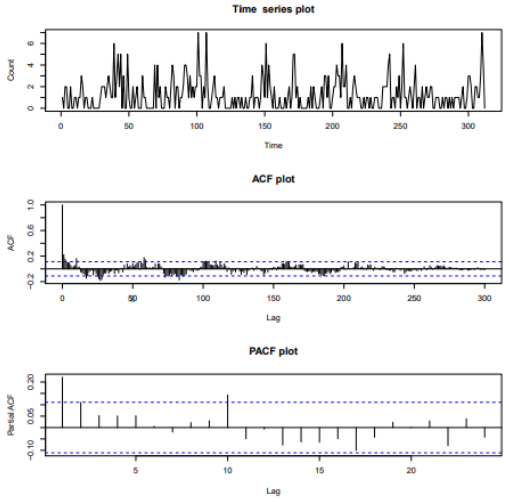


Figure 1: Time series, ACF and PACF plots for Compylobacteriosis data.

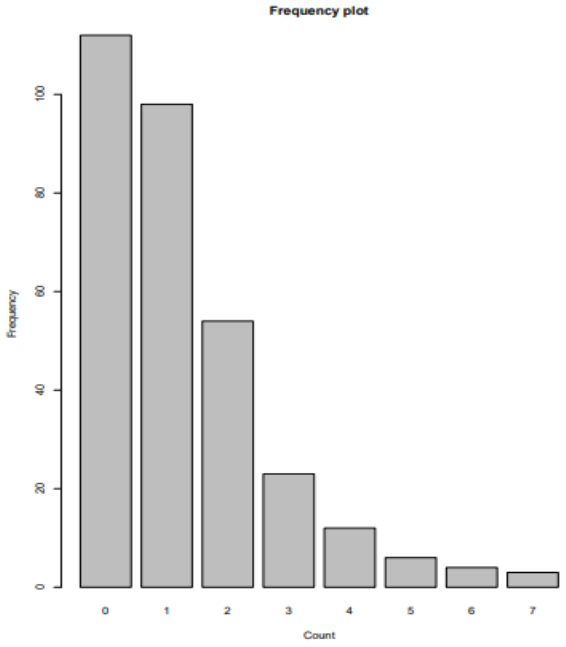


Figure 2: Frequency plot for Compylobacteriosis data.

6. Conclusions

INAR models without seasonality are been studied by the researchers. Here we studied an INAR(1) model with seasonal structure, which is a more general model, for $s = 1$ it rdces to usual INAR(1) model. Recently Tian et al. (2018) proposed similar model (NGINR(1)$_s$) based on negative binomial thinning, this model is bit complicated as compared to the model with binomial thinning. Here we have shown that the model with binomial thinning works as good as negative binomial thinning and some times better. Hence, the simple one can be used in practice. This model can be used for forecasting the future cases of a seasonal disease.

References

1. Bourguignon, M., Vasconcellos, K. L. P., Reisen, V. A and Isp´any, M. (2016). A Poisson INAR(1) process with a seasonal structure, *Journal of Statistical Computation and Simulation*, **86(2)**, 373-387.
2. Tian S., Wang D. and Cui S. A Seasonal geometric INAR(1) process based on negative binomial thinning. *Statistical Papers*, https://doi.org.org/10.1007/s00362-018-1060-7.

Latin square designs: Causal inference in a potential outcomes framework

Rahul Mukherjee[1], Peng Ding[2]
[1]Indian Institute of Management Calcutta, Kolkata 700104, India
[2]Department of Statistics, University of California, Berkeley, CA 94720, USA

Abstract
This article develops a randomization-based theory of causal inference from Latin square designs. For any treatment contrast, we propose an unbiased estimator and obtain its sampling variance. In contrast to many other situations in causal inference, it is found that Latin square designs do not admit a quadratic variance estimator which is conservative in the sense of having a nonnegative bias and which becomes unbiased under Neymannian strict additivity. Use of replicated Latin squares is seen to be helpful in overcoming this difficulty. The issue of minimaxity is also discussed in this context.

Keywords
Minimaxity; Strict additivity; Treatment contrast; Unbiased estimator; Variance estimation

1. Introduction

Causal inference in a potential outcome framework has been of significant current interest; see, for example, Imbens & Rubin (2015), Dasgupta et al. (2015), Mukerjee et al. (2018) and Zhao et al. (2018), where further references are available. Inspired by the findings in Sabbaghi & Rubin (2014), the present article aims at developing a randomization-based theory of causal inference from Latin square designs. Unlike the traditional method of analysis of these designs, the approach here does not require rigid linear model assumptions.

For any treatment contrast in a Latin square design, we propose an unbiased estimator and work out its sampling variance. We next examine if, as in many other situations of causal inference, one can obtain a quadratic variance estimator which is conservative in the sense of having a nonnegative bias and which becomes unbiased under Neymannian strict additivity (Neyman, 1923/1990). A matrix analysis is employed for this purpose and the answer turns out to be in the negative, apparently in disagreement with a result in Hinkelmann & Kempthorne (2007), and we show how this conflict can be resolved. Finally, use of replicated Latin squares is seen to be helpful in variance estimation, leading to a variance estimator which enjoys a minimaxity property.

2. Methodology

Consider a Latin square design that randomizes $n(\geq 2)$ treatments $1, \ldots, n$ over n^2 experimental units, arranged in an $n \times n$ array, so that each treatment occurs once in every row and every column, all such configurations being equiprobable. Let $Y(krc)$ be the potential outcome from the unit at row r and column c, when exposed to treatment k, $1 \leq k, r, c \leq n$. Then $\tau(rc) = \sum_{k=1}^{n} l(k)Y(krc)$ represents a typical unit-level treatment contrast, the $l(k), 1 \leq k \leq n$, being known constants, not all zeros, that sum to zero. Hence

$$\bar{\tau} = \sum_{r=1}^{n}\sum_{c=1}^{n} \tau(rc)/n^2 = \sum_{k=1}^{n} l(k)\bar{Y}(k) \tag{1}$$

defines a treatment contrast for the finite population of n^2 units, where

$$\bar{Y}(k) = \sum_{r=1}^{n}\sum_{c=1}^{n} Y(krc)/n^2 \tag{2}$$

is the mean of all the n^2 potential outcomes for treatment k. Interest lies in inference on $\bar{\tau}$ via the potential outcomes observed from the Latin square design.

Under a Latin square design, let $W(krc)$ be an indicator which equals 1 if the unit at row r and column c is assigned to treatment k, and 0 otherwise. Following the supplement to Sabbaghi & Rubin (2014), then for $1 \leq k, k^*, r, r^*, c, c^* \leq n, r \neq r^*, c \neq c^*$,

$$E\{W(krc)\} = 1/n, \tag{3}$$

while

$$E\{W(krc)W(k^*rc)\} = \delta(k, k^*)/n,$$
$$E\{W(krc)W(k^*rc^*)\} = E\{W(krc)W(k^*r^*c)\} = \{1 - \delta(k, k^*)\}/\{n(n-1)\},$$
$$E\{W(krc)W(k^*r^*c)\} = \{n - 2 + \delta(k, k^*)\}/\{n(n-1)^2\},$$

where $\delta(.,.)$ is Kronecker delta which equals 1 if the two arguments are equal, and 0 otherwise. As a result, for $1 \leq k, k^*, r, r^*, c, c^* \leq n$, if we write

$$\rho(k, k^*) = \delta(k, k^*) - 1/n, \tag{4}$$

and similarly define $\rho(r, r^*)$ and $\rho(c, c^*)$, then

$$\text{cov}\{W(krc), W(k^*r^*c^*)\} = \{n(n-1)^2\}\rho(k, k^*)\rho(r, r^*)\rho(c, c^*). \tag{5}$$

3. Results

We now find an unbiased estimator of the treatment contrast $\bar{\tau}$ in (1) and obtain its variance. Some matrix notation will help. For any k, let $Y(k)$ be the $n^2 \times 1$ vector with elements $Y(krc), 1 \leq r, c \leq n$, arranged in the lexicographic order of r and c, e.g., if $n = 2$, then $Y(k) = \big(Y(k11), Y(k12), Y(k21), Y(k22)\big)'$, with the prime denoting transpose. Thus

$$Y = (Y(1)', \ldots, Y(n)')' \tag{6}$$

is the $n^3 \times 1$ vector of all potential outcomes. Let

$$\hat{\bar{Y}}(k) = \sum_{r=1}^{n} \sum_{c=1}^{n} W(krc)Y(krc)/n \tag{7}$$

denote the mean of the observed potential outcomes for treatment $k, 1 \leq k \leq n$. Also, define

$$L = \text{diag}\{l(1), \ldots, (l(n)\}, P = I - n^{-1}J, \Phi = P \otimes P, V = [1/\{n(n-1)^2\}]LPL \otimes \Phi \tag{8}$$

where $\otimes$ stands for Kronecker product, while I and J denote, respectively, the identity matrix and the matrix of all ones, both of order n. Then the following result holds.

Theorem 1. (a) *An unbiased estimator of* $\bar{\tau}$ *is given by* $\hat{\bar{\tau}} = \sum_{k=1}^{n} l(k)\hat{\bar{Y}}(k)$, (b) $\text{var}(\hat{\bar{\tau}}) = Y'VY$. Proof. The truth of (a) is evident from (1), as $E\{\hat{\bar{Y}}(k)\} = \bar{Y}(k)$, by (2),(3) and (7). Now by (a),

$$\text{var}(\hat{\bar{\tau}}) = \sum_{k=1}^{n} \sum_{k^*=1}^{n} l(k)l(k^*)\text{cov}\{\hat{\bar{Y}}(k), \hat{\bar{Y}}(k^*)\}. \tag{9}$$

while by (5),(7),(8),

$$\text{cov}\{\hat{\bar{Y}}(k), \hat{\bar{Y}}(k^*)\} = (1/n^2) \sum_{r=1}^{n} \sum_{c=1}^{n} \sum_{r^*=1}^{n} \sum_{c^*=1}^{n} \text{cov}\{W(krc), W(k^*r^*c)\}Y(krc)Y(k^*r^*c^*)$$

$$= [1/\{n\left(n-1\right)^2\}]\rho\left(k, k^*\right) \sum_{r=1}^{n} \sum_{c=1}^{n} \sum_{r^*=1}^{n} \sum_{c^*=1}^{n} \rho(r, r^*)\rho(c, c^*)Y(krc)Y(k^*r^*c^*)$$

$$= [1/\{n\left(n-1\right)^2\}]\rho\left(k, k^*\right) Y(k)'\Phi Y(k^*). \tag{10}$$

The last step follows because by (4) and (8), $\rho(k, k^*)$ is the (k, k^*)th element of P and $\rho(\bar{r}, r^*)$ and $\rho(c, c^*)$can also be similarly interpreted. If one substitutes (10) in (9) and recalls the definitions of Y and V from (6) and (8), then one obtains (b).

In many situations of causal inference in a potential outcome framework, given a treatment contrast like $\bar{\tau}$, one can find an estimator of $\text{var}(\hat{\bar{\tau}})$ which is (a) quadratic in the observed potential outcomes, (b) conservative in the sense of having a nonnegative bias and (c) unbiased under Neymannian strict additivity; see e.g., Mukerjee et al. (2018). We shall show in Theorem 2 below that this is not the case for the Latin square design.

In our context, strict additivity (Neyman, 1923/1990) means constancy of $Y(krc) - Y(k^*rc)$, over $1 \leq r, c \leq n$ for every k, k^*. Consider now an estimator of $\text{var}(\hat{\bar{\tau}})$ in Theorem 1(b) which is quadratic in the observed potential outcomes. In the most general form, such an estimator is given by

$$\widehat{var}\,(\hat{\bar{\tau}}) = Z'AZ, \tag{11}$$

where A is any symmetric matrix of order n^3 and Z is the $n^3 \times 1$ vector with elements

$$Z(krc) = W(krc)Y(krc), \;\; 1 \leq k, r, c \leq n, \tag{12}$$

arranged in the lexicographic order of k, r and c. Note that by the definition of $W(krc)$, the only elements of Z that are possibly nonzero are the observed potential outcomes. Suppose the range of each potential outcome, say T, includes the points 0 and 1. Thus T can, in particular, be the real line, or the non-negative part thereof, or only the set {0, 1} which corresponds to binary potential outcomes. Then the following result holds.

Theorem 2. *For no choice of A in* (11), *one can have* $E\{\widehat{var}\,(\hat{\bar{\tau}})\} \geq \mathrm{var}(\hat{\bar{\tau}})$, *irrespective of the values of the potential outcomes in T, with equality attained whenever strict additivity holds.*

Proof. We first present a useful lemma and then complete the proof in three steps. As seen in Step 1, $E\{\widehat{var}\,(\hat{\bar{\tau}})\} = Y'BY$, for a symmetric matrix B, analogously to $\mathrm{var}(\hat{\bar{\tau}}) = Y'VY$ in Theorem 1(b). Still, however, the inequality in the statement of the theorem does not necessarily imply that $B - V$ is nonnegative definite (nnd), because the range T of any potential outcome may not be the entire real line. As a result, a more subtle argument is required in the proof; see e.g., Step 3.

Lemma 1. *If G* is a *symmetric matrix and $s'Gs = 0$ for every binary vector s, then $G = 0$.*

Proof. Taking s as the unit vectors, diagonal elements of G are zeros. Hence, taking s as the pairwise sums of distinct unit vectors, off-diagonal elements of G are also zeros, as G is symmetric.

Step 1 (expectation of variance estimator): Consider $\widehat{var}\,(\hat{\bar{\tau}})$ as given by (11) and (12). We label the rows and columns any square matrix of order n^3 by triplets $krc, 1 \leq k, r, c \leq n$, in the lexicographic order of k, r and c. Thus, if $a(krc, k^*r^*c^*)$ denotes the $(krc, k^*r^*c^*)$th element of A in (11), then

$$\widehat{var}\,(\hat{\bar{\tau}}) = \sum a(krc, k^*r^*c^*)W(krc), W(k^*r^*c^*)Y(krc)Y(k^*r^*c^*),$$

the sum being over $1 \leq k, r, c, k^*, r^*, c^* \leq n$. As a result, by (3) and (5),

$$E\{\widehat{var}\,(\hat{\bar{\tau}})\} = Y'BY \tag{13}$$

where the symmetric matrix B of order n^3 has $(krc, k^*r^*c^*$th element $b(krc, k^*r^*c^*) = a(krc, k^*r^*c^*)[(1/n^2) + \{n/(n-1)^2\}\rho(k, k^*)\rho(r, r^*)\rho(c, c^*)]$. In particular, then for every k, r, c, and $k^* \neq k, r^* \neq r, c^* \neq c$, by (4),

$$b(krc, k^*rc) = b(krc, kr^*c) = b(krc, krc^*) = 0 \tag{14}$$

Step 2 (invoking unbiasedness under strict additivity): Let ε be the $n \times 1$ vector of ones $\tilde{\varepsilon} = \varepsilon \otimes \varepsilon$, and I denote the identity matrix of order n. Recall that the range T of each potential outcome includes 0 and 1. Let $Y = (I \otimes \tilde{\varepsilon})g$, where $g = (g(1), \ldots, g(n))'$ is any binary vector. Then $Y(krc) = g(k), 1 \leq k, r, c \leq n,$,

so that each $Y(krc)$ is in T and strict additivity holds. Now, suppose $\widehat{var}\,(\hat{\tau})$ is unbiased for $\mathrm{var}(\hat{\tau})$ under strict additivity. Then by (13) and Theorem 1(b), $Y'BY = Y'VY$, for $Y = (I \otimes \tilde{\varepsilon})g$, irrespective of the binary vector g. As B and V are symmetric, from (8) and Lemma 1, we get

$$(I \otimes \tilde{\varepsilon}')B(I \otimes \tilde{\varepsilon}) = (I \otimes \tilde{\varepsilon}')V(I \otimes \tilde{\varepsilon}) = 0 \qquad (15)$$

Next, let $Y = (\varepsilon \otimes \tilde{I})h$, , where $\tilde{I} = I \otimes I$ and h is any binary vector of order n^2. Then again each $Y(krc)$ is in T and strict additivity holds. So, if $\widehat{var}\,(\hat{\tau})$ is unbiased for $\mathrm{var}(\hat{\tau})$ under strict additivity, then analogously to (15), by (8) and Lemma 1,

$$(\varepsilon' \otimes \tilde{I})B(\varepsilon \otimes \tilde{I}) = (\varepsilon' \otimes \tilde{I})V(\varepsilon \otimes \tilde{I}) - \xi\Phi, \qquad (16)$$

where $\xi = l'l/\{n(n-1)^2\}$, with $(l(1), \ldots, l(n))'$ and the fact that $l'J = 0$ is used.

Partition B as $B = [B(k, k^*)], 1 \le k, k^* \le n$, where each $B(k, k^*)$ is square of order n^2 and has elements $b(krc, k^*r^*c^*), 1 \le r, c, r^*, c^* \le n$. Then (15) and (16) are equivalent, respectively, to

$$\tilde{\varepsilon}'B(k, k^*)\tilde{\varepsilon} = 0, 1 \le k, k^* \le n, \qquad and \ \sum_{k=1}^{n}\sum_{k^*=1}^{n} B(k, k^*) = \xi\Phi. \quad (17)$$

If we take $k = k^*$ in the first set of identities in (17) and invoke the last two identities in (14), then

$$\sum_{r=1}^{n}\sum_{c=1}^{n} b(krc, krc) + \sum_{r=1}^{n}\sum_{c=1}^{n}\sum_{r^*(\neq r)=1}^{n}\sum_{c^*(\neq c)=1}^{n} b(krc, kr^*c^*) = 0, \qquad 1 \le k \le n \qquad (18)$$

Similarly, if we equate the diagonal elements from the last identity in (17), invoke the first identity in (14), and recall (8), then

$$\textstyle\sum_{k=1}^{n} b(krc, krc) = \xi\{(n-1)/n\}^2 = l'l/n^3, \quad 1 \le r, c \le n \qquad (19)$$

By (18) and (19), as $l \neq 0$,

$$\sum_{k=1}^{n}\sum_{r=1}^{n}\sum_{c=1}^{n}\sum_{r^*(\neq r)=1}^{n}\sum_{c^*(\neq c)=1}^{n} b(krc, kr^*c^*) = -\sum_{k=1}^{n}\sum_{r=1}^{n}\sum_{c=1}^{n} b(krc, krc) = -l'l/n < 0. \quad (20)$$

Step 3 (invoking conservativeness): Let $e(1), \ldots, e(n)$ be the unit vectors of order n. Take $Y = E(K) \otimes h$, where $1 \le k \le n$, and h is any binary vector of order n^2. Then each element of Y is in T, and in view the partitioning of B indicated above, together with (8),

$$Y'BY = h'B(k, k)h, \quad Y'VY = [\{l(k)\}^2/\{n^2(n-1)\}]h'\Phi h. \qquad (21)$$

Now, suppose $\widehat{var}\,(\hat{\tau})$ is conservative with a nonnegative bias, i.e., $Y'BY \ge Y'VY$ for every Y, by (13) and Theorem 1(b). If we add the consequent inequalities from (21) over $1 \le k \le n$, then

$$\textstyle\sum_{k=1}^{n} h'B(k, k)h \ \ge [l'l/\{n^2(n-1)\}h'\Phi h. \qquad (22)$$

Label the n^2 elements of h by pairs rc, $1 \le r, c \le n$, in the lexicographic order of r and c, and take h such that it has ones in the positions $1c_1, \ldots, nc_n$, and zeros elsewhere, where $\{c_1, \ldots, c_n\}$ is any permutation of $\{1, \ldots, n\}$. Then by (19) and (8), the left- and right-hand sides of (22) equal

$$\sum_{k=1}^{n}\left\{\sum_{r=1}^{n} b(krc_r, krc_r)\right) + \sum_{r=1}^{n}\sum_{r^*(\neq r)=1}^{n} b(krc_r, kr^*c_{r^*})\} = (l'l/n^2) + \sum_{k=1}^{n}\sum_{r=1}^{n}\sum_{r^*(\neq r)=1}^{n} b(krc_r, kr^*c_{r^*})\},$$

and $[l'l/\{n^2(n-1)\}][n\{(n-1)/n\}^2 + n(n-1)(-1/n)^2] = l'l/n^2$ respectively. Hence (22) yields

$$\sum_{k=1}^{n}\sum_{r=1}^{n}\sum_{r^*(\neq r)=1}^{n} b(krc_r, kr^*c_{r^*})\} \ge 0,$$

for each $\{c_1, \ldots, c_n\}$. Summing the above over all such $n!$ permutations,

$$\{(n-2)!\}\Sigma_{k=1}^{n}\Sigma_{r=1}^{n}\Sigma_{c=1}^{n}\Sigma_{r^*(\neq r)=1}^{n}\Sigma_{c^*(\neq c)=1}^{n} b(krc, kr^*c^*) \ge 0,$$

which contradicts (20) and completes the proof of the theorem.

We now reconcile Theorem 2 with a result in Hinkelmann & Kempthorne (2007) on unbiased estimation of $\mathrm{var}(\hat{\tau})$ under strict additivity, by showing that their estimator is not conservative in general, i.e., its expectation can be less than $\mathrm{var}(\hat{\tau})$ when such additivity does not hold. It is easily seen that strict additivity, as considered above, is equivalent to

$$Y(krc) = g(k) + h(rc), \qquad 1 \le k, r, c \le n. \tag{23}$$

for some constants $g(k)$ and $h(rc)$. For $n \ge 3$, under strict additivity as formulated in (23), Hinkelmann & Kempthorne (2007, § 10.2) reported an unbiased estimator of $\mathrm{var}(\hat{\tau})$ as $\widehat{var}_{HK}(\hat{\tau}) = \left(\frac{l'l}{n}\right) MSE$ where $l = \left(l(1), \ldots, l(n)\right)'$ as before, and MSE is the mean square due to error in the traditional analysis of variance of a Latin square design. With a view to examining this estimator in some detail, we now obtain its expectation in general, without assuming strict additivity.

Theorem 3. $E\{\widehat{var}_{HK}(\hat{\tau})\} = Y'V_1Y$, where

$$V_1 = [l'l/\{n(n-1)\}^3]\{n(n-3)P \otimes P \otimes P + (n-1)(J \otimes P \otimes P + P \otimes J \otimes P + P \otimes P \otimes J)\}.$$

Proof. Note that $MSE = SSE/\{(n-1)(n-2)\}$, where

$$\mathrm{SSE} = \Sigma_{k=1}^{n}\Sigma_{r=1}^{n}\Sigma_{c=1}^{n} W(krc)\{Y(krc) - \hat{\bar{Y}}(k\bullet\bullet) - \hat{\bar{Y}}(\bullet r\bullet) - \hat{\bar{Y}}(\bullet\bullet c) + 2\hat{\bar{Y}}\}^2, \tag{24}$$

in our notation, with

$$\begin{aligned} &\hat{\bar{Y}} = \Sigma_{k=1}^{n}\Sigma_{r=1}^{n}\Sigma_{c=1}^{n} W(krc)Y(krc)/n^2, && \hat{\bar{Y}}(k\bullet\bullet) = \Sigma_{r=1}^{n}\Sigma_{c=1}^{n} W(krc)Y(krc)/n, \\ &\hat{\bar{Y}}(\bullet r\bullet) = \Sigma_{k=1}^{n}\Sigma_{c=1}^{n} W(krc)Y(krc)/n, && \hat{\bar{Y}}(\bullet\bullet c) = \Sigma_{k=1}^{n}\Sigma_{r=1}^{n} W(krc)Y(krc)/n. \end{aligned} \tag{25}$$

Because the sum of the $W(krc)$, over any one argument is 1, from (24) and (25), one can check that

$$\text{SSE} = \Sigma_{k=1}^{n}\Sigma_{r=1}^{n}\Sigma_{c=1}^{n}W(krc)\{Y(krc)\}^2 - n\Sigma_{k=1}^{n}\{\hat{\bar{Y}}(k\bullet\bullet)\}^2 - n\Sigma_{r=1}^{n}\{\hat{\bar{Y}}(\bullet r\bullet)\}^2 - n\Sigma_{c=1}^{n}\{\hat{\bar{Y}}(\bullet\bullet c)\}^2 + 2n^2\hat{\bar{Y}}^2 \quad (26)$$

By (3)-(5) and (25), writing $\sum$ and $\Sigma 1$ for sums over $1 \le k, r, c, k^*, r^*, c^* \le n$ and $1 \le k, r, c, r^*, c^* \le n$ respectively,

$$E[\Sigma_{k=1}^{n}\Sigma_{r=1}^{n}\Sigma_{c=1}^{n}W(krc)\{Y(krc)\}^2] = (1/n)Y'(I\otimes I\otimes I)Y,$$

$$E(n^2\hat{\bar{Y}}^2) = (1/n^2)\Sigma[(1/n^2) + \{n/(n-1)^2\}\rho(k,k^*)\rho(r,r^*)\rho(c,c^*)]Y(krc)Y(k^*r^*c^*)$$
$$= (1/n^4)Y'(J\otimes J\otimes J)Y + [1/\{n(n-1)^2\}]Y'(P\otimes P\otimes P)Y,$$

$$E[n\Sigma_{k=1}^{n}\{\hat{\bar{Y}}(k\bullet\bullet)\}^2] = (1/n)\Sigma_1[(1/n^2) + \{n/(n-1)^2\}\rho(k,k)\rho(r,r^*)\rho(c,c^*)]Y(krc)Y(kr^*c^*)$$
$$= (1/n^3)Y'(I\otimes J\otimes J)Y + [1/\{n(n-1)\}]Y'(I\otimes P\otimes P)Y.$$

Similarly,

$$E[n\Sigma_{r=1}^{n}\{\hat{\bar{Y}}(\bullet r\bullet)\}^2] = (1/n^3)Y'(J\otimes I\otimes J)Y + [1/\{n(n-1)\}]Y'(P\otimes I\otimes P)Y,$$

$$E[n\Sigma_{c=1}^{n}\{\hat{\bar{Y}}(\bullet\bullet c)\}^2] = (1/n^3)Y'(J\otimes J\otimes I)Y + [1/\{n(n-1)\}]Y'(P\otimes P\otimes I)Y.$$

If we now calculate E(SSE) term-by-term from (26), use the fact that $I = P + n^{-1}J$ and recall the definition of $\widehat{var}_{HK}(\hat{\tau})$, then Theorem 3 follows after some simplification.

As before, let ε be the $n \times 1$ vector of ones, and $\tilde{\varepsilon} = \varepsilon\otimes\varepsilon$. If strict additivity holds then by (23), $Y = g\otimes\tilde{\varepsilon} + \varepsilon\otimes h$, for some vectors g and h of orders $n \times 1$ and $n^2 \times 1$, respectively. As a result, by (8), Theorem 3 and Theorem 1(b), both $E\{\widehat{var}_{HK}(\hat{\tau})\}$ and $\text{var}(\hat{\tau})$ reduce to $[1/\{n(n-1)^2\}](l'l)(h'\Phi h)$. So, perfectly in agreement with Hinkelmann & Kempthorne (2007), $\widehat{var}_{HK}(\hat{\tau})$ is an unbiased estimator of $\text{var}(\hat{\tau})$ under strict additivity. It remains to satisfy ourselves that this estimator is not conservative in general. To that end, let h be the binary vector of order n^2, with elements labelled by pairs $rc, 1 \le r, c \le n$, in the lexicographic order of r and c, such that h has ones in positions 11, 22,..., nn, and zeros elsewhere. Then

$$h'(J\otimes P)h = h'(P\otimes J)h = n\{(n-1)/n\} + n(n-1)(-1/n) = 0, \quad (27)$$

$$h'\Phi h = h'(P\otimes P)h = n\{(n-1)/n\}^2 + n(n-1)(-1/n)^2 = n-1. \quad (28)$$

As in the context of Theorem 2, suppose the range of each potential outcome, say T, includes the points 0 and 1. Considering again the unit vectors $e(1), \ldots, e(n)$ of order n, take $Y = e(k)\otimes h$, where $1 \le k \le n$. Then each $Y(krc)$ is in T, and from (8), (27), (28), Theorem 3 and Theorem 1(b),

$$E\{\hat{\text{var}}_{\text{HK}}(\hat{\bar{\tau}})\} = [l'l/\{n(n-1)\}^3]\{(n-3)(n-1)^2 + (n-1)^2\} = (n-2)l'l/\{n^3(n-1)\},$$
$$\text{var}(\hat{\bar{\tau}}) = [\{l(k)\}^2/\{n^2(n-1)\}]h'\Phi h = \{l(k)\}^2/n^2.$$

If $\widehat{var}_{HK}(\hat{\bar{\tau}})$ is a conservative estimator of $\text{var}(\hat{\bar{\tau}})$, i.e., $E\{\widehat{var}_{HK}(\hat{\bar{\tau}})\} \geq \text{var}(\hat{\bar{\tau}})$ in general, then from the above, $(n-2)l'l \geq n(n-1)\{l(k)\}^2, 1 \leq k \leq n$. Because $l'l > 0$, summing these inequalities over k, we get $n(n-2) \geq n(n-1)$, and a contradiction is reached.

4. Discussion and Conclusion

Use of replicated Latin squares appears to be the simplest and most flexible way of overcoming the difficulty in variance estimation as noted in Theorem 2. Consider Un^2 experimental units arranged in $U(\geq 2)$ groups, where each group has n^2 units, arranged in an $n \times n$ array. Within each group, the n treatments are randomized as a Latin square, while randomization is done independently across groups. Let $Y_u(krc)$ be the potential outcome from the unit at row r and column c of group u, when exposed to treatment $k, 1 \leq u \leq U, 1 \leq k, r, c \leq n$. Then $\bar{\tau}_u = \sum_{k=1}^{n} l(k)\bar{Y}_u(k)$ is the counterpart of (1) for group u, where $\bar{Y}_u(k) = \sum_{r=1}^{n}\sum_{c=1}^{n} Y_u(krc)/n^2$. Hence, $\bar{\tau} = \sum_{u=1}^{U} \bar{\tau}_u/U$ now defines a treatment contrast for the finite population of Un^2 units. Write $\bar{\tau}_{vec} = (\bar{\tau}_1, \ldots, \bar{\tau}_u)'$.

Analogously to (7), one can define $\bar{Y}_u(k)$ as the mean of the observed potential outcomes for treatment k in group u. Then $E\{\hat{\bar{Y}}_u(k)\} = \bar{Y}_u(k)$, so that $\hat{\bar{\tau}}_u = \sum_{k=1}^{n} l(k)\hat{\bar{Y}}_u(k)$ is unbiased for $\bar{\tau}_u, 1 \leq u \leq U$, and as a result, $\hat{\bar{\tau}} = \sum_{u=1}^{U} \hat{\bar{\tau}}_u/U$ is an unbiased estimator of $\bar{\tau}$, with $\text{var}(\hat{\bar{\tau}}) = \sum_{u=1}^{U} var(\hat{\bar{\tau}}_u)/U^2$, as $\bar{\tau}_1, \ldots, \hat{\bar{\tau}}_U$ are independent due to independent randomization across groups. Moreover, $\text{var}(\hat{\bar{\tau}}_u)$ is given by Theorem 1(b) for any u, with Y there replaced by the corresponding vector of the n^3 potential outcomes from group u. Let $\hat{\bar{\tau}}_{vec} = (\hat{\bar{\tau}}_1, \ldots, \hat{\bar{\tau}}_U)'$ Then the following result is not hard to obtain; cf. Alquallaf et al. (2018) in the context of strip-plot designs.

Theorem 4. *Consider a variance estimator* $\widehat{var}(\hat{\bar{\tau}}) = \hat{\bar{\tau}}'_{vec} Q \hat{\bar{\tau}}_{vec}$, *where* Q *is any nnd matrix of order* U, *with* (i) *each row sum zero and* (ii) *each diagonal element* $1/U^2$. *Then*

(a) $E\{\widehat{var}(\hat{\bar{\tau}})\} = var(\hat{\bar{\tau}}) + \bar{\tau}'_{vec} Q \bar{\tau}_{vec}$.

(b) *The bias* $\bar{\tau}'_{vec} Q \bar{\tau}_{vec}$ *vanishes for every treatment contrast if* $\bar{Y}_u(k) - \bar{Y}_u(k^*)$ *is constant over* $1 \leq u \leq U$, *for every* k, k^*.

By Theorem 4(a), $\widehat{var}(\hat{\bar{\tau}})$ is a conservative estimator of $\text{var}(\bar{\tau})$ because of having a nonnegative bias. Moreover, the condition in Theorem 4(b) for this bias to vanish for every contrast amounts to between-group additivity. This is milder than strict additivity which now demands the constancy of $Y_u(krc) - Y_u(k^*rc)$, over $1 \leq u \leq U$ and $1 \leq r, c \leq n$ for every k, k^*. Along the lines of Mukerjee et al. (2018), it can also be seen that the

minimax choice of the nnd matrix Q subject to (i) and (ii) in Theorem 4, in the sense of minimizing the maximum of the bias $\bar{\tau}'_{vec} Q \bar{\tau}_{vec}$ over every spherical contour of the form $\bar{\tau}'_{vec} \bar{\tau}_{vec} = \lambda (> 0)$, is $Q_0 = [1/\{U(U-1)\}](I_U - U^{-1} J_U)$, where I_U is the identity matrix and J_U is the matrix of ones, both of order U. With $Q = Q_0$, the expression for $\widehat{var}(\hat{\bar{\tau}})$ in Theorem 4 becomes $\widehat{var}(\hat{\bar{\tau}}) = \frac{\sum_{u=1}^{U} (\hat{\bar{\tau}}_u - \hat{\bar{\tau}})^2}{\{U(U-1)\}}$.

In the spirit of randomized block designs replicating each treatment more than once in every block (Dasgupta et al., 2015), one may as well wish to consider a frequency-square or *F*-square design as an alternative to replicated Latin squares. Following Hedayat & Seiden (1970), given fixed positive integers $f_1, \ldots, f_n$ which sum to N an *F*-square design randomizes treatments $1, \ldots, n$ over N^2 experimental units, arranged in an $N \times N$ array, so that treatment k occurs f_k times in every row and every column, $1 \leq k \leq n$, all such configurations being equiprobable. The case $f_1 = \ldots = f_n = 1$ corresponds to the Latin square design. One can show that an *F*-square design with $f_1, \ldots, f_n \geq 2$ admits an estimator of the variance of any treatment contrast estimator meeting the requirements spelt out in the previous section, that is, the variance estimator is conservative in the sense of having a nonnegative bias, and the bias vanishes under strict additivity. Unlike $\widehat{var}(\hat{\bar{\tau}})$ in Theorem 4, however, it turns out that such a variance estimator is not guaranteed to be nonnegative even in the balanced case $f_1 = \ldots = f_n$. A more compelling concern from a practical viewpoint is that the *F*-square approach is less flexible than the use of replicated Latin squares. For instance, the smallest *F*-square design with $f_1 = \ldots = f_n \geq 2$ corresponds to $f_1 = \ldots = f_n = 2$ and involves $4n^2$ experimental units, as against $2n^2$ experimental units that two replications of a Latin square will require. Because of these reasons, we do not pursue the *F*-square design any further.

References

1. Alquallaf, F. A., Huda, S. & Mukerjee, R. (2018). Causal inference from strip-plot designs in a potential outcomes framework. Preprint, arXiv:1805.06663
2. Dasgupta, T., Pillai, N. S. & Rubin, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. Roy. Statist. Soc. Ser. B*, **77**, 727-753.
3. Hedayat, A. & Seiden, E. (1970). *F*-square and orthogonal *F*-squares design: A generalization of Latin square and orthogonal Latin squares design. *Ann. Math. Statist.*, **41**, 2035-2044.

4. Hinkelmann, C. & Kempthorne, O. (2007). *Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design*, *2nd ed*. Hoboken, NJ: Wiley.
5. Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction.* New York: Cambridge University Press.
6. Mukerjee, R., Dasgupta, T. & Rubin, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *J. Amer. Statist. Assoc.*, **113**, 868-881. 7.
7. Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles (with discussion). Section 9 (translated). *Statist. Science*, **5**, 465-480.
8. Sabbaghi, A. & Rubin, D. B. (2014). Comments on the Neyman-Fisher controversy and its consequences. *Statist. Science*, **29**, 267-284.
9. Zhao, A., Ding, P., Mukerjee, R. & Dasgupta, T. (2018). Randomization-based causal inference from split-plot designs. *Ann. Statist.*, **46**, 1876-1903.

Undercoverage bias of web survey on smoking and heavy drinking

Jason Hsia, Guixiang Zhao,Machell Town
US Centers for Disease Control and Prevention

Abstract

The Behavioral Risk Factor Surveillance System (BRFSS) is the world's largest annual telephone survey that collects data on adult US residents' health, use of preventive services, health care access and related behavioral risk factors, such as tobacco and alcohol use and sedentary lifestyle/getting regular physical activity. In the last decade, BRFSS response rates have been going down. The cost of making more calls to improve response rates is expensive. New modes of cost-effective data collection could be helpful. To explore the possibilities of using a web survey to complement or replace the existing BRFSS, we examine the web survey undercoverage bias of target population. We used 2017 BRFSS questions on Internet use, current smoking, and alcohol use to examine undercoverage bias by partitioning it into a product of two components: proportion of non-Internet use and difference in prevalence of interests between Internet users and non-Internet users. For current smoking, the highest absolute bias was found among those 45 years and older (-1.37 to -3.13%) and those with less than a high school education (3.70%). For heavy drinking, the highest bias was found with respondents 75 years and older (1.22%) and with less than a high school education (1.11%)

Keywords

Web survey; Undercoverage bias; Behavioral risk factors

1. Introduction

Web surveys save field operational costs, which is a major advantage over traditional telephone or face-to-face interviews. Its common disadvantage, however, is target-population undercoverage. Each year, the Behavioral Risk Factor Surveillance System (BRFSS)--the world's largest telephone survey--collects data on participants' health, use of preventive services, health care access, and their related behavioral risk factors, such as tobacco and alcohol use, sedentary lifestyle/regular physical activity and other behaviors in 50 states, the District of Columbia, and participating US territories. The survey started in 1984 and, since that time, has had a target population of noninstitutionalized US residents aged 18 years and older. BRFSS is subject to high survey costs, especially in addressing the decreasing response rates over the last decade. To date, little is known about web surveys' associated undercoverage bias regarding surveillance on health and related behavioral

risk factors. To explore possibilities of complementing or replacing the current BRFSS with a web survey, it is necessary to examine this method's undercoverage bias of its target population. The 2017 BRFSS included a question about participants' Internet use and we used that for this study of web surveys and undercoverage bias.

2. Methodology

Data used in this study are from the 2017 BRFSS and were aggregated from all 50 states and the District of Columbia data. A total of 424,262 eligible persons had been interviewed by telephone. We used binary answers (yes and no) to the question: "Have you used the Internet in the past 30 days?" to differentiate Internet users. We studied outcome variables that were the two most-commonly surveyed behaviors: current smoking and heavy alcohol drinking. In this study, we used 403,078 observations that had complete information on three variables: Internet use, current smoking, and heavy drinking (men having more than 14 drinks per week and women having more than 7 drinks per week). In order to assess the undercoverage bias, we assume in this study that all Internet users in the past 30 days will participate in a web survey of similar content and that the response rate would be 100%.

Let P,P_I, and P_{NI} be prevalence for both Internet users (IU) and non-Internet users (NIU), IU only and NIU only, respectively, and N and N_{NI} be total and NIU size, respectively. We use

$$P_I - P = \frac{N_{NI}}{N}(P_I - P_{NI})$$

to be the bias of undercoverage when only IU are included, that is, the difference between prevalence for Internet use and for both Internet use and non-Internet use. The estimation was based on a correspondent formula of sample:

$$p_I - p = \frac{n_{NI}}{n}(p_I - p_{NI})$$

where lower-case symbols represent estimated correspondent parameters using the BRFSS sample and correspondent sample sizes. Variances of those estimated parameters were estimated using linearization methods while taking complex survey design into account.

3. Results

Among 403,078 respondents to the 2017 BRFSS, 15.06% did not use the Internet in the past 30 days prior to the interview. The proportion of the NIU attenuated with decrease of age (Table 1). The youngest age group had only about 3.84% of NIU, while the highest proportion of the NIU was among respondents aged 75 years and older (49.63%). For other demographic variables, the highest proportion of the NIU was found among those with lowest attained education (45.55%), lowest federal poverty level of <100%

(29.59%) and 100-199 % (25.35%) as well as among Hispanic participants (23.79%) and Non-Hispanic Black participants (22.18%).

Overall prevalence differences between IU and NIU was -4.41% for current smoking and 2.23% for heavy drinking. The differences by demographic variables are shown in Tables 2 and 3. For current smoking, the highest absolute differences of prevalence of current smoking were in the age groups of 45-54, 55-64, and 65-74 years (-12.73%, -16.10%, and -9.34%, respectively), men (-8.03%), Non-Hispanic Black (-7.70%), Non-Hispanic Other (-8.88%), and those who received less than a high school education (8.13%). For heavy drinking, the differences of prevalence between IU and NIU were in general smaller than the differences for current smoking. The highest differences were in those aged 18-34 and 75 years and older (3.48% and 2.46%, respectively), women (4.23%), Non-Hispanic Whites (2.77%), and those with lower educational attainment (2.44% for less than high school and 2.39 for high school).

For current smoking, the highest absolute bias was found among those 45 years and older (-1.37 to -3.13%), men (-1.18%), Non-Hispanic Black (-1.71%), and those with less than high school education (3.70%). For heavy drinking, the highest bias was found with those 75 years and older (1.22%), women (0.65%), and participants with less than a high-school education (1.11%).

4. Discussion and Conclusion

The proportions of NIU are shown as in Table 1. Findings that the proportion of NIU were higher among older people and those with lower education, and were lower among Non-Hispanic Whites are consistent with a previous report (Cohen and Adams 2011). Undercoverage in a web survey would not be a problem if IU did not differ systematically from NIU. We noticed differences, however, between IU and NIU (Tables 2 and 3). IU therefore could not be considered a random sample from the target population, i.e., valid conclusions could not be drawn from the web survey.

In terms of decomposition, there were three types of significant biases. First, both proportion of NIU and the absolute difference of prevalence between IU and NIU was relatively large, e.g., among the 55-64 age group or with less high school education for current smoking. Second, the proportion of NIU was large but the difference of prevalence between IU and NIU was small, e.g., among 75 years and older for current smoking. Third, the proportion of NIU is small but the difference of prevalence between IU and NIU was relatively large, e.g., among 18-34 age group for heavy drinking. The findings within these subgroups suggest we need to examine both the proportion of NIU and the difference between IU and NIU at the same time when we evaluate the undercoverage bias, rather than focus on the proportion of NIU only (Bethlehem and Biffignandi 2012).

A few points should be highlighted in this study. First, both biases and the differences between IU and NIU for the current smoking variable were more substantial than those for the heavy drinking variable. This might be because the prevalence of heavy drinking is lower than that of current smoking. Further, those participants who were older, Non-Hispanic Black, Hispanic, less than high-school educated, and in a lower federal poverty level (FPL) had larger proportion of NIU. This is generally consistent with size of biases in our study with exceptions for Hispanic and low FPL—a finding that suggests when one examines biases of a web survey, one should examine both the proportion of NIU and the difference of prevalence of interests between IU and NIU as well as low prevalence overall. Reporting a Hispanic background (in general) is associated with a lower prevalence of current smoking (American Lung Association, 2018). Last, for those with less than a high school education, the prevalence of current smoking among IU was smaller than what was found overall, and the prevalence of heavy drinking was larger among this group than it was overall—findings that deserve further investigation.

There are some limitations to this study. First, this study was based on a sample survey, BRFSS, instead of a population level data, such as compete population registry. Second, we assume that all who have used the Internet during the last month will respond to current smoking and heavy drinking questions. This is not true in practice. Our study, therefore, examined a part of undercoverage biases.

This is a preliminary study of biases of a web survey on gauging current smoking and heavy drinking. It primarily used data from the BRFSS. Different levels of undercoverage biases were detected for overall and subgroup prevalences. The proportion of NIU and the difference of prevalence between IU and NIU play roles in assessing undercoverage biases. This research can be expanded to continue to explore the possibility of using a web survey to complement or replace the existing BRFSS.

References

1. Cohen RA, Adams PF. Use of the Internet for Health Information: United States, 2009. NCHS Data Brief No. 66, July 2011.
2. Bethlehem J, Biffignandi S. Handbook of Web Surveys, John Wiley & Sons, 2012.
3. American Lung Association. Tobacco Use in Racial and Ethnic Populations. https://www.lung.org/stop-smoking/smoking-facts/tobacco-use-racial-and-ethnic.html

Table 1 Proportion of Individuals Who Did Not Use the Internet in the Past 30 Days, Prevalence of Current Smoking and Heavy Drinking in the US (BRFSS, 2017)

		Non-Internet User		Current Smoking		Heavy Drinking	
	Total	%	SE	%	SE	%	SE
Overall	403078	15.06	0.12	16.06	0.13	6.25	0.08
Age (yrs)							
18-34	64864	3.84	0.15	16.95	0.26	7.10	0.18
35-44	46358	7.39	0.25	20.00	0.38	6.92	0.26
45-54	62139	12.17	0.29	18.08	0.31	6.60	0.20
55-64	87371	19.47	0.32	17.65	0.29	6.13	0.17
65-74	84042	26.54	0.37	11.23	0.24	5.21	0.17
75+	58304	49.63	0.51	5.14	0.21	3.03	0.17
Sex							
Men	176870	14.65	0.17	18.16	0.20	6.79	0.13
Women	226208	15.44	0.17	14.10	0.16	5.75	0.11
Race/Ethnicity							
NH White	315220	12.44	0.11	16.81	0.14	7.10	0.10
NH Black	31851	22.18	0.43	17.98	0.41	4.51	0.23
Hispanic	28782	23.79	0.49	12.93	0.40	5.11	0.28
NH Other	27225	9.37	0.34	13.37	0.46	4.15	0.28
Education							
<High School	27567	45.55	0.58	26.75	0.51	5.33	0.28
High School	108747	20.55	0.23	20.70	0.26	6.34	0.18
Some College	112709	8.28	0.15	16.30	0.23	6.54	0.15
≥College	154055	3.20	0.08	6.28	0.12	6.25	0.12
FPL[¶] (%)							
<100	32715	29.59	0.54	25.21	0.48	4.96	0.27
100-199	73221	25.35	0.35	21.24	0.34	5.45	0.20
≥200	196640	6.41	0.11	11.70	0.17	7.33	0.13
Unknown	100502	16.57	0.23	16.12	0.24	5.57	0.14

[¶] Federal Poverty Level

Table 2 Prevalence of Current Smoking among Internet Users and Non-Internet Users, Their Difference, and Bias in the United States (BRFSS, 2017)

	Internet Users		Non-Internet Users		Difference[§]		Bias		
	%	SE	%	SE	%	SE	%	SE	
Overall	15.39	0.14	19.80	0.34	-4.41	0.37	-0.66	0.18	*
Age (yrs)									
18-34	16.87	0.27	18.88	1.57	-2.01	1.60	-0.08	0.30	
35-44	19.54	0.39	25.80	1.64	-6.26	1.69	-0.46	0.46	
45-54	16.53	0.31	29.26	1.11	-12.73	1.15	-1.55	0.43	*
55-64	14.52	0.28	30.61	0.86	-16.10	0.90	-3.13	0.43	*
65-74	8.75	0.24	18.09	0.58	-9.34	0.63	-2.48	0.44	*
75+	3.77	0.24	6.53	0.34	-2.76	0.41	-1.37	0.55	*
Sex									
Men	16.98	0.21	25.01	0.56	-8.03	0.60	-1.18	0.26	*
Women	13.90	0.18	15.19	0.41	-1.29	0.44	-0.20	0.23	
Race/Ethnicity									
NH White	16.15	0.15	21.46	0.40	-5.31	0.43	-0.66	0.18	*
NH Black	16.27	0.46	23.97	0.92	-7.70	1.02	-1.71	0.59	*
Hispanic	12.94	0.45	12.90	0.83	0.03	0.95	0.01	0.63	
NH Other	12.53	0.49	21.42	1.33	-8.88	1.42	-0.83	0.57	
Education									
<High School	30.46	0.75	22.33	0.67	8.13	1.00	3.70	0.77	*
High School	21.05	0.30	19.36	0.46	1.69	0.55	0.35	0.35	
Some College	16.10	0.25	18.50	0.77	-2.40	0.81	-0.20	0.28	
≥College	6.16	0.12	9.89	0.78	-3.73	0.78	-0.12	0.15	
FPL[¶] (%)									
<100	25.86	0.57	23.64	0.90	2.22	1.07	0.66	0.72	
100-199	21.61	0.41	20.16	0.63	1.45	0.75	0.37	0.49	
≥200	11.43	0.18	15.64	0.63	-4.21	0.66	-0.27	0.20	
Unknown	15.49	0.26	19.27	0.59	-3.78	0.64	-0.63	0.33	

[¶] Federal Poverty Level
[§] Difference between Internet users and non-Internet users.
* 95% confidence interval does not include zero.

Table 3 Prevalence of Heavy Drinking among Internet Users and Non-Internet Users, Their Difference, and Bias in the United States (BRFSS, 2017)

	Internet Users		Non-Internet Users		Difference[§]		Bias		
	%	SE	%	SE	%	SE	%	SE	
Overall	6.59	0.09	4.35	0.19	2.23	0.21	0.34	0.13	*
Age (yrs)									
18-34	7.23	0.18	3.76	0.70	3.48	0.72	0.13	0.26	
35-44	7.00	0.27	5.90	1.06	1.10	1.09	0.08	0.37	
45-54	6.49	0.20	7.35	0.69	-0.86	0.72	-0.10	0.28	
55-64	6.22	0.19	5.74	0.40	0.48	0.44	0.09	0.25	
65-74	5.65	0.19	4.01	0.38	1.64	0.43	0.44	0.26	
75+	4.25	0.26	1.79	0.22	2.46	0.34	1.22	0.31	*
Sex								0.00	
Men	6.78	0.14	6.81	0.36	-0.03	0.38	0.00	0.19	
Women	6.40	0.12	2.17	0.18	4.23	0.21	0.65	0.16	*
Race/Ethnicity								0.00	
NH White	7.45	0.11	4.68	0.22	2.77	0.24	0.35	0.15	*
NH Black	4.59	0.26	4.24	0.52	0.35	0.58	0.08	0.34	
Hispanic	5.59	0.33	3.57	0.49	2.03	0.59	0.48	0.43	
NH Other	4.05	0.30	5.08	0.86	-1.03	0.91	-0.10	0.41	
Education								0.00	
<High School	6.44	0.42	4.00	0.34	2.44	0.54	1.11	0.50	*
High School	6.83	0.21	4.45	0.30	2.39	0.36	0.49	0.27	
Some College	6.72	0.17	4.56	0.34	2.16	0.38	0.18	0.23	
≥College	6.28	0.12	5.46	0.99	0.82	1.00	0.03	0.17	
FPL[¶] (%)								0.00	
<100	5.15	0.31	4.53	0.51	0.62	0.60	0.18	0.41	
100-199	5.89	0.25	4.16	0.32	1.73	0.41	0.44	0.32	
≥200	7.40	0.14	6.27	0.54	1.13	0.56	0.07	0.19	
Unknown	6.03	0.16	3.27	0.25	2.76	0.30	0.46	0.22	*

[¶] Federal Poverty Level

[§] Difference between Internet users and non-Internet users

* 95% confidence interval does not include zero

Spatial fay-herriot model for estimating expenditure in Bangka Belitung Province, Indonesia

Ahmad Risal[1], Siti Muchlisoh[2]
[1]BPS - Statistics Indonesia, Jakarta
[2]Polytechnic of Statistics STIS, Jakarta

Abstract

In Indonesia, the local government has the right to self-regulate the government affairs known as regional autonomy system. One indicator that assist government in carrying out development of the region is expenditure per capita. STIS, a campus for future employments in Statistics Indonesia has done survey named Praktik Kerja Lapangan (PKL), to estimated expenditure up to village level in Bangka Belitung Province with Small Area Estimation (SAE), Fay-Herriot Model. Fay-Herriot hasn't included spatial influences into model. Meanwhile, there's possibility of spatial influence on expenditure. Hence, this research will estimate SAE method by including spatial influence or Spatial Fay-Herriot (SFH) Model. Data used were raw data PKL, Potential Village and shapefile of Bangka Belitung. Based on Mean Square Error (MSE), MSE value of estimated SFH is smaller than direct estimation. It means, estimation with the SFH provide more precise estimation results than direct estimation. So, SFH is better for further investigation.

Keywords

Small Area Estimation; Fay-Herriot Model; Spatial Fay-Herriot Model; Expenditure

1. Introduction

Since January 1st, 2001, the Indonesian government system went through a transformation from centralized to decentralized known as regional autonomy. In this format, local governments have the right, authority, and obligation to regulate and manage their own government affairs and the interests of the local community. Therefore, program development by local governments began to be directed at the level of a smaller area, ranging county, district, or even village. In order for the development program to be carried out more effectively and efficiently, it is necessary to have the availability of data up to the micro scale.

During a present, providing data through surveys conducted by Statistics Indonesia – Badan Pusat Statistik (BPS) as the official provider of statistics is not entirely capable of directly estimating small areas to districts and villages yet. If an estimation is made directly to the level of a small area, it will produce a low precision value because the number of samples is insufficient. A solution

to this problem is to provide a budget to increase the number of samples, so that the survey design carried out is able to provide a statistical output of small areas.

The next problem arises because of budget constraints, so information about an indicator is not available evenly in a small area. By paying attention to the information needs of small areas and seeing the conditions of limited resources, a statistical method that is able to meet the availability of information needs to be applied even though the resources owned are limited. This method is known as the Small Area Estimation.

Small Area Estimation (SAE) is a technique that is often used and is expected to produce better precision to estimate smaller areas. Estimates in SAE are based on the model, so additional information is needed from variables that have a relationship with the variables being observed which are called covariates. The additional information is in the form of previous census data or administrative data from the area concerned. All this additional information must be related to the parameters observed (Rao, 2003).

Poverty is a problem faced by every region in Indonesia. In addressing these problems made an effort to make the estimation of the poorer regions. This estimation is expected to help the government to reduce poverty more accurately. The main indicator used by the BPS in determining someone is said to be poor or not is the average per capita expenditure.

So far, BPS has obtained per capita expenditure data per region from a survey named National Socio-Economic Survey (SUSENAS). However, published results have limitations in the number of samples for estimation in small areas. The number of samples used can only estimate up to the district level or even the province. This situation does not meet the needs of the regional government, which requires local statistical data to the lower area level.

Politeknik Statistika STIS as a campus for future employments in BPS, already done a survey named Field Work Practice – Praktik Kerja Lapangan (PKL) in 2017. The aim of the field work was conducted on the study of poverty and income distribution in the Bangka Belitung Province using Fay-Herriot model. The use of Fay-Herriot model to estimate parameters has not incorporated the spatial influence into the model. On the other hand, it does not rule out the possibility of spatial influence on expenditure per capita in a region. Therefore, Fay-Herriot model was developed by incorporating spatial influences into the model. The Fay-Herriot model estimator takes into account the random influence of the spatial correlated area known as the Spatial Fay-Herriot model. Spatial Fay-Herriot model can improve the variety structure of the estimation model for small areas that have spatial correlations between areas. The model used in the area-based on Spatial Fay-Herriot model is because the spatial modelling included in the SAE model is modelling the

spatial data type area. The estimator of Spatial Fay-Herriot model has been used by Pratesi & Salvati (2007) by entering a weighting spatial matrix, spatially nearest neighbours into the Spatial Fay-Herriot model.

Based on that, this study examines the estimated expenditure per capita in Bangka Belitung Province at 2017 by using direct estimation and Spatial Fay-Herriot model. After that based on the results of the estimation, carried out a comparison from that two method of estimation method to determine the most appropriate estimation method. so that we hope to produce estimates of expenditure per capita to the level of small areas with good precision can be achieved.

Several studies have been conducted relating to the Spatial Fay-Herriot Model. Matualage (2012) estimated the per capita expenditure in Jember Regency, East Java Province using the Spatial Fay-Herriot model. The results showed that the per capita expenditure of each village in Jember Regency obtained with estimators of Fay-Herriot model was more diverse when compared to village per capita expenditures produced with the Spatial Fay-Herriot model with an average per capita expenditure for each village with the Fay-Herriot model higher than the average expenditure per capita village with the Spatial Fay-Herriot model. In addition, the Spatial Fay-Herriot model is better used to estimate the average village/district expenditure per capita in Districts Jember with the participation variables compared to the direct estimator method or the the Fay-Herriot model. Whereas when compared to the direct estimator, the Fay-Herriot model estimator RRMSE value is not much different

2. Methodology

The data used in this study are secondary data from various sources. The following are details of the data used:

- Data used as direct estimation, namely the average per capita expenditure of each sub-district and village in the Bangka Belitung Islands Province with 7053 samples. This data is estimated from per capita expenditure for each region selected as a sample in Field Work.
- Village Potential (PODES) 2014 as covariates in Small Area Estimation. The variables chosen are as follows:
    - The main source of income for the majority of the population
        - District: agriculture (X_1); mining and quarrying (X_2); processing industry (X_3); wholesale/retail trade and restaurants (X_4); transportation, warehousing and communication (X_5); and services (X_6).
        - Village: mining and quarrying (d_1); wholesale/retail trade and restaurants (d_2); transportation, warehousing and communication (d_3); and services (d_4).

- Number of hospital facilities (X_7) .
- Number of village unit cooperatives – Koperasi Unit Desa (KUD) that are still active/operating (X_8) .

➢ Queen type of spatial weighting matrix originating from the shapefile of Bangka Belitung Islands Province.

3. Small Area Estimation

Small Area Estimation (SAE) is an indirect estimation method that combines survey data with other supporting data such as from the previous census. The supporting data must contain variables with the same characteristics as the survey data. This is done so that smaller areas can be estimated and provide a better level of precision.

Based on data availability, SAE is grouped into two types, namely area-level and unit-level. The study in this paper used area-level of SAE. This is because the supporting data (covariates) available only reaches the area level. The area level model connects the small area direct estimator with supporting data from other domains for each area, that is $\mathbf{x_i}^{\mathbf{T}} = (x_{1i}, \dots, x_{Pi})$. The parameter of small area that will be estimated is θ_i. The linear model that explains these relationships is:

$$\theta_i = \mathbf{x_i}^{\mathbf{T}}\boldsymbol{\beta} + z_i v_i \qquad i = 1,2,\dots, m \tag{1}$$

Where:

$\beta = (\beta_{0,\dots}\beta_P)^T$ is SAE regression coefficient estimates

z_i = positive constant

v_i = random effect area, assumed $v_i \sim$ iid $N(0, \sigma^2_v)$

m = number of observations (area)

In making conclusions about the population under equation 1, it is assumed that the direct estimation value is known $\hat{\theta}_i$ and can be written as follows:

$$\theta i = \theta i + ei \qquad i = 1, 2 \dots, m \tag{2}$$

where e_i is a sampling error assumed to be $e_i \sim N(0, \Psi_i)$.

The SAE model for area level consists of two levels of the model component, namely the indirect estimation model component showed by equation (1) and the direct component showed by equation (2). If the models in equations (1) and (2) are combined they will form the following equation:

$$\hat{\theta}_i = \mathbf{x_i^T}\boldsymbol{\beta} + z_i v_i + e_i \qquad i = 1,2 \dots, m \tag{3}$$

The above model has not included the effect of spatial correlation in it. The SAE model which included the spatial correlation between areas was first introduced by Cressie (Cressie 1991 referred to in Rao 2015), assuming spatial dependence following the model of Conditional Autoregressive (CAR) process. The SAE model was later developed by other researchers, including Pratesi and Salvati (2007) by assuming that the spatial dependence included in the error

component of random effect area follows the model of Simultaneous Autoregressive (SAR) process. The SAR model (Spatial Fay-Herriot Model) itself was first introduced by Anselin (Anselin 1992) where area v random influence vectors is as follows:

$$\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u} \quad (4)$$

Where:

ρ = spatial Autoregressive coefficient
W = spatial weighted matrix
v = random effect area
u = error vector of random effect area

4. Results

4.1 Direct Estimation of Per Capita Expenditure

Direct estimates of the per capita expenditure can only be done in regions that have at least a sample of the Field Work Practices (PKL STIS). Bangka Belitung Islands Province has a total of 43 sub-districts and 387 villages. PKL STIS just have sample on 42 sub-districts and 135 villages with number of samples are 7053 samples. So, the direct estimation is done only on that sample.

Table 1. Descriptive Statistics on Per Capita Expenditure Based on Direct Estimation

Statistic	Subdistrict	Village
Number of observations	42	135
Mean	IDR.1,238,894	IDR.1,208,024
Std. Deviation	IDR.232,380	IDR.285,505
Minimum	IDR.880,850	IDR.512,963
Median	IDR.1,168,565	IDR.1,184,435
Maximum	IDR.1,788,768	IDR.1,877,401

Based on the results of direct estimation, per capita expenditures for sub-districts and villages were obtained as shown in table 1. At the sub-district level, sub-districts that have the lowest average per capita expenditure is the districts in South Bangka District, Tukak Sadai and sub-districts that have the highest average per capita expenditure is Rangkui in Pangkalpinang City. At the village level, village which has the lowest per capita expenditure average is Peradong village, West Bangka District and the village which has the highest average per capita expenditure is Masjid Jamik Village, Pangkalpinang City.

4.2 Spatial Fay-Herriot of Per Capita Expenditure

In estimating the Spatial Fay-Herriot model, a spatial weighted matrix is needed. In this study, the spatial weighted matrix used is the Queen-type which has been standardized in the row. Spatial weighted type queen

contiguity takes into account the proximity of a region to another region. If the area is right around the observation area, then the area is given code one (1) whereas if it is not right around the observation area then the area is given code zero (0). The Illustration is on Figure 1.

Figure 1. Illustration of Queen Contiguity Matrix

Queen		
1	1	1
1	i	1
1	1	1

After obtaining a spatial weighting matrix, a spatial autocorrelation test will be conducted using the Moran's I. Test. A summary of the results of the Moran's I test is presented in Table 2. The hypotheses used are as follows:

H0: I = 0 (no spatial autocorrelation)

H1: I ≠ 0 (there is spatial autocorrelation)

Based on the test results of the Autocorrelation in Table 2, it is shown that using the queen type spatial weighting matrix shows there is spatial autocorrelation in the random effect area at the village level.

Table 2. Spatial Autocorrelation Test Results on Average per Capita Expenditure

Statistics	Sub-District	Village
Moran's I	-0.1404	-0.0074
Z count	-1.0151	1.5197
P-value	0.8450	0.0643*

*sign at level significant of 10%

The next step is applying the Spatial Fay-Herriot model with the REML procedure. The covaites selection method that will be used to estimate the per capita expenditure is done by backward elimination. The result of using backward elimination at the sub-district level shows that from the 8 covariates used, only the remaining 3 variables are left, namely variables X_1, X_5 and X_6. Whereas for the village level, from all of the covariates used, only remaining five variables, they are variables d_1, d_2, d_3, d_4 and x_7. A summary of the estimation from covariates is presented in Table 3.

Table 3. Estimated coefficient of covariates in the Spatial Fay-Herriot Model

Sub-district			Village		
Variable	β	p-*value*	Variable	β	p-*value*
(Intercept)	1295961	0.0000	(Intercept)	1061677	0.0000
x_1	-23180	0.0000	d_1	186360	0.0001
x_5	85008.2	0.0000	d_2	610388	0.0000
x_6	92033.5	0.0000	d_3	494445	0.0000
			d_4	378201	0.0023
			x_7	235344	0.0359

The estimation of random effect variance (σ_v^2) is carried out by the maximum likelihood (MLE) method. The random effect variance calculation using the R program produces a random effect variance value of 3918068286 at the sub-district level. At the village level, the results of the random effect variance are 34761096201. Besides estimating σ_v^2, the Spatial Fay-Herriot model also estimates the coefficient of spatial autoregression (ρ). At the sub-district level, ρ value is -0.981 and at the village level the value is 0.1961.

Statistical values from the estimation of sub-district and village level per capita expenditure is presented in Table 4.

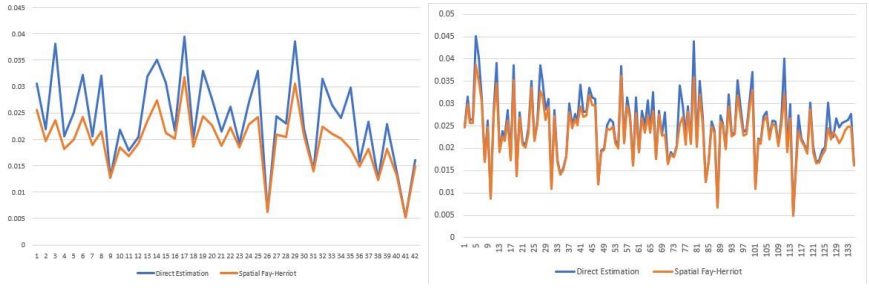
Table 4. Descriptive Statistics on Per Capita Expenditure Based on Spatial Fay-Herriot Model

Statistics	Sub-District	Village
Number of observations	42	135
Mean	IDR.1,240,177	IDR.1,192,000
Std. Deviation	IDR.212,025	IDR.266,385
Minimum	IDR.854,922	IDR.537,006
Median	IDR.1,175,266	IDR.1,158,716
Maximum	IDR.1,723,181	IDR.1,834,688

4.3 Comparison of MSE and RRMSE

To see the estimation method that gives the best estimation results, a comparison of the estimation results between the direct estimation method and the Spatial Fay Herriot Model will be compared. Comparisons will be made through Mean Square Error (MSE). After that, MSE value used to count Relative Root Mean Square Error (RRMSE) value. In this section a comparison is made for each sub-district and village.

Figure 2. Comparison of RRMSE of expenditure per capita sub-district (a) and village (b)



Based on Figure 2, it can be seen at the sub-district and village levels, the RRMSE value of the Spatial Fay-Herriot model estimation method is better than the direct estimate. Therefore, it can be said that the Spatial Fay-Herriot model is the best method of estimating the value of expenditure per capita at the sub-district and village level.

5. Conclusion

Based on the results discussed earlier, it can be seen that at the sub-district level, the largest value of per capita expenditure is in Rangkui sub-district that is IDR.1,788,768 and the lowest is in Tukak Sadai sub-district. While at the village level, the largest value of per capita expenditure is in Masjid Jamik village that is IDR.1,877,401.and the lowest value is in peradong village. After that, based on the RRMSE value we can see that the RRMSE value of the estimated Spatial Fay-Herriot model is smaller than the direct estimation. This shows that estimation with the Spatial Fay-Herriot model provide more precise estimation results than direct estimation. So we can say that the Spatial Fay-Herriot model (Small Area Estimation) is better for further investigation.

References

1. Anselin L. (1992). *Spatial econometrics: method and models*. Boston: Kluwer.
2. Matualage D. (2012). *Metode prediksi tak bias linier terbaik empiris spasial pada area kecil untuk pendugaan pengeluaran per kapita: studi kasus Kabupaten Jember Provinsi Jawa Timur [tesis].* Bogor: Institut Pertanian Bogor.
3. Rao JNK. (2003). *Small Area Estimation*. New York: John Wiley and Sons.
4. Rao JNK, Molina I. (2015). *Small Area Estimation, second edition*. New York: John Wiley and Sons.
5. Pratesi M, Salvati N. (2008). *Small area estimation: the EBLUP estimator based on spatially correlated random area effects. Statistical methods and applications*, Stat. Meth. & Appl.

On modelling the frequency of antenatal care visits in Bangladesh

Kakoli Rani Bhowmik[1], Sumonkanti Das[2,3], Md. Atiqul Islam[2]
[1]Center for Statistics, Hasselt University, Hasselt, Belgium
[2]Department of Statistics, Shahjalal University of Science & Technology, Sylhet, Bangladesh
[3]Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands

Abstract

The standard Poisson (PR) and negative binomial (NBR) regression models are used for modelling number of antenatal care (ANC) visits based on the overdispersion assumption. Overdispersion can be due to excessive zeros (called zero-inflation) or having intra-cluster correlation (ICC) in the ANC data. Zero-inflated count data are modelled using two-part (zero and count) models - zero-inflated (ZIPR and ZINBR) and Hurdle (HPR and HNBR) regression models. Mixed effects Poisson (MPR) and negative binomial (MNBR) regression models may reduce overdispersion but not zero-inflation. In such case, mixed effects zero-inflated (MZIPR and MZINBR) and Hurdle (MHPR and MHNBR) models can be used. The main aim of this study is to identify a proper count regression model for the number of ANC visits of pregnant women in Bangladesh covering the issues of overdispersion, zero-inflation and ICC. Women ANC information have been extracted from 2014 Bangladesh Demographic and Health Survey data where 4493 women provide ANC information for their last birth. Since the existing zeros (about 22% women did not take ANC) cannot be defined as genuine (either structural or sampling), both zeroinflated and Hurdle models are investigated. Demographic, socio-economic, environmental, regional and women empowerment related covariates have been utilized to develop the models. Standard models (PR and NBR) with their mixed effects versions (MPR and MNBR) are fitted at first to examine whether overdispersion and zero-inflation have been accounted by the considered explanatory variables. Zeroinflated and Hurdle models without and with random effects are developed if zero-inflation still exists. Several model diagnostics including Vuong, likelihood ratio (LR) and uniformity (simulation-based approach) tests have been employed for model selection. Among the standard (PR, NBR) and mixed effects (MPR, MNBR) regression models, only MNBR overcomes overdispersion problem partially (ratio: 0.913, p-value=0.002), however zero-inflation was remained (ratio: 1.223, p-value: 0.000). Comparison of ZINBR and HNBR models with MZINBR and MHNBR indicates that cluster-specific random effects are significant at both zero and count components. The uniformity test (H0: Fitted model suits well the data) and the significance of extra random effects at zero-part indicate that MHNBR model with extra random effects performs better for the ANC data. The

findings suggest random community effects should be considered along with overdispersion and zero-inflation in modelling the ANC data of Bangladeshi women. Also, selection of either Hurdle or Zero-inflated type model should be taken carefully since assumption of all structural zeros is tough to meet in real world data.

Keywords
Hurdle model; negative binomial model; random effects; uniformity test; zero-inflated model

1. Introduction

Poisson regression (PR) and negative binomial regression (NBR) models have been widely used for modelling count response variable such as the frequency of antenatal care (ANC) visits (Rose *et al.* 2006, Beeckman *et al.*, 2010, Amrin, 2016, Islam and Masud, 2018). However, these models provide inconsistent estimates when overdispersion (mean greater than variance) and excess zeros (also called zero-inflation) exist in the data (Hinde & Demetrio, 1998). In such situation, hurdle and zero-inflated regression models (either Poisson or negative binomial distribution) are applied for accounting excess zeros (Staub & Winkelmann, 2013). In zero-inflated regression models (zero-inflated PR (ZIPR) and zero-inflated NBR (ZINBR)), it is assumed that the response values are generated from two generating process: only zero counts (structural zeros) from first process, and non-negative counts from a standard model (either a Poisson or a negative binomial model) which could produce zero count as well called sampling zeros (Lambert, 1992). On the other hand, in hurdle models (Mullahy, 1986) – Hurdle PR (HPR) and Hurdle NBR (HNBR), it is assumed there are only structural or genuine excess zeroes (i.e., a pregnant woman had zero ANC visit during a given period because she never visits). The details of these models can be found in Zeileis *et al.* (2008) and Yousuf *et al.* (2018). These regression models have also been widely applied for modelling ANC data, a proper example of data with excess zeros (Yusuf *et al.*, 2018; Fan & Habibov, 2009). In addition to the problem of overdispersion and zeroinflation, the correlation between measurements (a common phenomenon in longitudinal, repeated survey and clustered data) needs to be considered (Min and Agresti, 2005). Considering cluster/subject specific random effects model, overdispersion and correlation problem can be partially solved, but the issue of zero-inflation remains. Thus zero-inflated and Hurdle models have been extended for accounting this correlation by assuming cluster specific random effects at either count or both count and zero components (Hall, 2000; Yau and Lee, 2001).

In Bangladesh, a significant proportion of women do not take any ANC visits during their pregnancy period. During the last two decades, this

proportion reduce from 50% in 1994 to 20% in 2014 (NIPORT *et al.*, 2016). Consequently, there are great chance of excess zeros in the distribution of the number of ANC visits in Bangladesh. Also, the data are usually collected through standard cluster-sampling design and hence intra-cluster-correlation (ICC) should be considered for modelling the number of ANC visits. Previously, the zero-inflation and overdispersions are covered by fitting either standard models (PR, NBR) or zero-inflated and Hurdle models on modelling the frequency of ANC visits (Amrin, 2016; Islam and Masud, 2018). In this study, a proper count regression model has been explored covering all the three issues of overdispersion, zero-inflation and correlation in the data by employing the standard models (PR, NBR), zero-inflated models (ZIPR, ZINBR), Hurdle models (HPR, HNBR), mixed effects PR and NBR (MPR, MNBR), mixed effects zero-inflated (MZIPR, MZINBR) and Hurdle (MHPR, MHNBR) models considering cluster-specific random effects for count component only and both count and zero components (sometimes referred as extra random effects). The mixed models with extra random effects are denoted hereafter as MZINBR.ERE (for example) for the MZINBR model. Both zero-inflated and Hurdle type models are examined, since types of excess zeros are not possible to determine and hence selection of an appropriate model will be model driven instead of data driven.

2. Methodology

In this study, data are extracted from the nationally representative 2014 Bangladesh Demographic and Health Survey (BDHS), where the country was stratified into 20 sampling strata according to urban and rural enumeration areas of 7 divisions (NIPORT *et al.*, 2016). Two-stage stratified sampling design was implemented for collecting data: 600 clusters (393 from rural and 207 from urban areas) were drawn with probability proportional to the enumeration area size at first stage and 30 Households per cluster were selected with an equal probability systematic procedure at second stage. The information on ANC visits were collected from 4493 ever married women who gave birth in the three years preceding the survey. Among women with two or more live births within the given period, information only for the last birth were recorded. Mothers were asked a number of questions about ANC visits and the received health care during the antenatal visits. Mother's age, education, and mass media exposure, husband's education, wealth status, place of residence, and regional settings have been considered as explanatory variables in the study for developing regression models. The considered explanatory variables are found significant predictors of the number of ANC visits in several previous researches (Rahman *et al.*, 2008; Ali *et al.*, 2018; Islam and Masud, 2018).

In this study the number of ANC visits (non-negative integer) is the target response variable for which a proper count regression model is aimed to identify. Before model development, the bivariate relationship of the considered explanatory variables with the response variable has been checked by developing Poisson regression model for each of the explanatory variables. Let y_{ij} denotes number of ANC visits of i^{th} women living in j^{th} cluster, and the vector $\boldsymbol{X}_{ij}$ denotes the corresponding values of the considered explanatory variables. Assuming independence of ANC visits of i^{th} women, the PR and NBR model can be defined as:

$$log(\mu_{ij}) = \beta_0 + \beta \boldsymbol{X}^T{}_{ij}$$

where μ_{ij} is the expected number of ANC visits, β_0 is the overall intercept, and $\boldsymbol{\beta}$ is the vector of regression coefficients. The difference is the assumed distribution of y_{ij} in respective models. The cluster-specific random effects (the simplest form of mixed effects model) can be easily add in the previous models as

$$log(\mu_{ij}) = (\beta_0 + b_{oj}) + \beta X_{ij}{}^T$$

where, b_{oj} stands for random intercept at cluster level and assumed to follow a normal distribution with constant variance.

There are two components (count component and zero component) for both zero-inflated and Hurdle models and separate explanatory variables can be used. Again let $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ are vectors of known explanatory variables used for developing zero- and count-component models respectively. Then the simplified zero- and count-component models can be expressed respectively as

$$logit(\varphi_{ij}) = \gamma_0 + \gamma Z_{ij}{}^T \text{ and } log(\mu_{ij}) = \beta_0 + \beta X_{ij}{}^T$$

where μ_{ij} is the mean of the underlying parent distribution and φ_{ij} is the probability of zero counts from the binary process (binary logistic model). The difference in these two types of model can be explained by the distribution of y_{ij}. The distribution of the number of ANC visits is modeled using zero-inflated regression model with y_{ij} as:

$$P[y_{ij} = 0] = \varphi_{ij} + (1 - \varphi_{ij}) f(y_{ij}) \text{ and } P[y_{ij} \geq 1] = (1 - \varphi_{ij}) f(y_{ij})$$

where $f(.)$ is density of either Poisson or negative binomial distribution. While in Hurdle model,

$$P[y_{ij} = 0] = \varphi_{ij} \text{ and } P[y_{ij} \geq 1] = (1 - \varphi_{ij}) f(y_{ij}) / \{1 - f(y_{ij} = 0)\}$$

where $f(.)$ is the density of standard truncated-at-zero distribution and φ_{ij} is the probability of an observation being zero. The straightforward random intercept zero-inflated and Hurdle models can be expressed by adding a random component (b_{0j}) with β_0 and another extra random component c_{0j}

with γ_0 in zero-component which reflect cluster-specific random effects. R packages "pscl", "lme4", and "GLMMadaptive" have been employed for model development.

Standard models (PR and NBR) with their mixed effects versions (MPR and MNBR) are fitted at first to examine whether overdispersion and zero-inflation have been accounted by the considered explanatory variables. If zero-inflation still exists, zero-inflated and Hurdle models without and with random effects are then developed. As model diagnostics, Vuong test for non-nested models (Vuong, 1989), likelihood ratio (LR) test for nested models, Akaike's information criteria (AIC), overdispersion test, zero-inflation test and uniformity test (H_0: fitted model suits well for the data) are respectively implemented for model selection. The later three tests are based on residual diagnostics for hierarchical (multi-level / mixed) regression models available in DHARMa package of Hartig (2018). The final model is selected based on the uniformity test and the significance of the cluster-specific random effects in the model.

3. Results

The mean and median number of ANC visits were observed about 2.75 and 2 respectively. Though ANC from any source is considered, about 22% pregnant women did not have ANC visit. Bivariate analysis indicates that mean and median frequency of ANC visits significantly vary with administrative region, place of residence, women's education, partners' education, women's mass media exposure, status of women's pregnancy wanted, household wealth status and household decision maker on healthcare (results are not shown here).

The comparison of PR and NBR models without and with cluster-specific random intercepts shown in **Table 1** indicates that PR and MPR models fails to capture the overdispersion, while both NBR and MNBR accounts overdispersion but all models are infected by the issue of Zero-inflation. Among these models NBR model seems better for the considered data according to DHARMa uniformity test with very lower p-value (0.066), however AIC, log-likelihood, and LR test indicate inclusion of random intercept was still needed for the data. Thus zero-inflated and Hurdle models without and with random intercepts were developed. The results of Vuong tests for non-nested models shown in Table 2 indicates that either ZINBR or HNBR can be considered as better model for accounting excess zeroes. Table 2 also reflects that the over-dispersion is captured better by the NBR based models than the PR based models. The results of LR tests for nested models shown in **Table 3** indicate that cluster-specific random intercepts should be considered in the NBR based models. Even random effects are found important for both count-

and zero-part models in both cases of ZINBR (MZINBR and MZINBR.ERE) and HNBR (MHNBR and MHNBR.ERE) models.

Since there are four possible candidates as the best model for the ANC data, the DHARMa's uniformity test has been done for finding the best suitable model. **Table 4** shows that ZINBR with random effects at count-part (MZINBR) confirms uniformity with the observed count data but LR test in **Table 4** shows that this model still require random intercepts at the zero part (MZINBR.ERE). However, the MZINBR.ERE failed the uniformity test. On the other hand, HNBR with random intercepts at countpart (MZINBR) and also MZINBR with extra random effects at zero-part (MZINBR.ERE) passed the uniformity test. Thus MZINBR.ERE can be considered the best model among the possible best candidate models for the ANC data. Also, MHNBR and MHNBR.ERE models show lower clusterspecific variance components than MZINBR and MZINBR.ERE models. It is noted that different sets of explanatory variables are found significant in the zero and count parts of the fitted models, and same sets are maintained in all the two-part models for comparison purpose.

According to the selected Hurdle model with random effects at both count and zero components (MHNBR.ERE), division, place of residence, economic status, media exposure, mother's and partner's education status, mother's health care decision, and desire for pregnancy have highly significant effects on ANC visits. The odds ratio (OR) of having no ANC visits and incidence rate ratio (IRR) of the number of ANC visits are obtained respectively from zero-part and count-part of the fitted Hurdle-based model. Since both parts have cluster-specific random effects, hence the estimated parameters are interpreted differently from ordinary Hurdle model. It is worthy to mention that other models are not presented in this study but only stated them for comparison purpose.

From the count-part model, it is observed that women residing in Khulna (IRR:1.182) and Rangpur division (IRR:1.239), women from richer (IRR: 1.155) and richest (IRR: 1.306) households, women having access of mass media at least once a week (IRR:1.156), women attending secondary (IRR: 1.213) and higher education (IRR:1.430), and women with desire of pregnancy (IRR:1.232) had significantly higher IRR of attending ANC visits while women residing in rural area (IRR: 0.840) and women without power on health care decision (IRR:0.870) had significantly lower IRR after accounting the variation at the cluster level. On the other hand, the regression coefficients of the zero-part model indicate that mothers residing in Khulna (OR: 0.496) and Rangpur (OR:0.629) divisions, women from middle (OR:0.633), richer (OR:0.401) and richest (OR:0.197) households, women having access of mass media at least once a week (OR:0.567), women having primary (OR: 0.725), secondary (OR: 0.450) and higher education (OR: 0.243), and women's partners having secondary (OR:0.628) and higher education (OR:0.474) had significantly lower

odds of not attending ANC visit. The higher value of cluster-specific variance component (σ_z^2 =0.582) at zero-part indicates a significant amount of variation in the number of ANC visits is due to between-cluster heterogeneity.

4. Discussion and Conclusion

This study has attempted to find a proper way of identifying an appropriate count regression model for the number of ANC visits among the pregnant women in Bangladesh. A number of studies have been done to determine the risk factors of the number of ANC visits considering the PR and NBR models and their extensions for the excess-zero (zero-inflated and Hurdle model). Since the household survey data are conducted through any cluster-sampling design, there are obvious possibility of correlation among the women behaviour for taking ANC visits. Consequently, ignorance of this correlation in the model might provide bias estimates with unfortunate under-coverage rate due to lower standard errors. So, the mixed effects version of PR, NBR, ZIPR, ZINBR, HPR and HNBR are explored in this study. Only the random intercept model has been covered in this study, though further investigation can be done for checking significance of any random slope in the model. The findings of this study indicate that cluster-specific variation is significant for both excess zero and positive counts. These indicates that community (cluster) has a significant effect in the variation of the number of ANC visits and consequently community level variations should be considered in the model development for identifying risk factors of not attending in any ANC visit as well as the number of ANC visits. Finally, application of multilevel modelling technique has allowed to account community-level variations in the number of ANC visits, although the maximum variations originated mainly from women, household, and regional level factors that are explained by the considered variables. The findings recommend to consider community effects (ICC) along with overdispersion and zero-inflation in modelling the ANC data of Bangladeshi women. Also, selection of either Hurdle or Zero-inflated type model should be taken carefully since assumption of all structural zeros is tough to meet in real world data. It is better to take decision based on model (whether the fitted model can explain all the zeros) rather than based on types of zeros.

References

1. Ali, N., Sultana, M., Sheikh, N., Akram, R., Mahumud, R. A., Asaduzzaman, M., & Sarker, A. R. (2018). Predictors of Optimal Antenatal Care Service Utilization Among Adolescents and Adult Women in Bangladesh. *Health services research & managerial epidemiology*, 5, 1-8.
2. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1-14.

3. Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, *33*(3), 341-365.
4. Rahman, M., Islam, R., & Islam, A. Z. (2008). Rural-urban differentials of utilization of ante-natal health-care services in Bangladesh. *Health policy and development*, *6*(3), 117-125.
5. Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307-333.
6. Hartig, F. (2018). DHARMa: Residual Diagnostics for Hierarchical (multi-level/mixed) regression models. R package version 0.1. 5.
7. Staub, K. E., & Winkelmann, R. (2013). Consistent estimation of zero-inflated count models. *Health economics*, 22(6), 673-686.
8. NIPORT, Mitra and Associates, & ICF International. (2016). Bangladesh Demographic and Health Survey 2014. Dhaka, Bangladesh and Rockville, Maryland, U.S.A.: NIPORT, Mitra and Associates, and ICF International.
9. Hall, D. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56, 1030–39.
10. Yau, K. K. and Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, 20, 2907–20.
11. Beeckman, K., Louckx, F., & Putman, K. (2010). Determinants of the number of antenatal visits in a metropolitan region. *BMC Public Health*, 10(1), 527.
12. Islam, M. M., & Masud, M. S. (2018). Determinants of frequency and contents of antenatal care visits in Bangladesh: Assessing the extent of compliance with the WHO recommendations. *PloS one*, 13(9), e0204752.
13. Yusuf, O. B., Afolabi, R. F., & Ayoola, A. S. (2018). Modelling Excess Zeros in Count Data with Application to Antenatal Care Utilisation. *International Journal of Statistics and Probability*, 7(3), 22.
14. Min, Y., & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical modelling*, *5*(1), 1-19.
15. Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of statistical software*, *27*(8), 1-25.
16. Amrin, A. (2016). An Analysis of the Status of Antenatal Care in Bangladesh. *International Journal of Science and Research Methodology*, *5*(2), 49-57.
17. Fan, L., & Habibov, N. N. (2009). Determinants of maternity health care utilization in Tajikistan: learning from a national living standards survey. *Health & place*, *15*(4), 952-960.
18. Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zeroinflated and hurdle models for modeling vaccine

adverse event count data. *Journal of biopharmaceutical statistics*, *16*(4), 463-481.

Table 1: Akaike's information criteria (AIC), log-likelihood, likelihood-ratio (LR), dispersion, zero-inflation and uniformity tests of PR, NBR, MPR and MNBR models

Model	AIC	log-likelihood (df)	LR test (p-value)	Dispersion Test (Ratio & p-value)	Zero-Inflation Test (Ratio Statistic & p-value)	Uniformity Test (D-Statistic & p-value)
PR	19274.66	-9613.33 (24)	712.52	1.358 & 0.00	2.001 & 0.00	0.114 & 0.00
MPR	18565.31	-9257.66 (25)	(0.000)	1.096 & 0.004	1.685 & 0.00	0071 & 0.00
NBR	18314.42	-9132.21 (25)	216.47	0.937 & 0.000	1.208 & 0.00	0.019 & 0.066
MNBR	18099.96	-9023.98 (26)	(0.000)	0.913 & 0.002	1.223 & 0.00	0.022 & 0.026

Table 2: Vuong Tests for the non-nested models PR, ZIPR, HPR, NBR, ZINBR and HNBR models

Model 1	Model 2	Test Statistic (AIC Corrected)	p-value	Better Model
PR	NBR	-11.852	< 2.22e-16	NBR
PR	ZIPR	-12.316	< 2.22e-16	ZIPR
PR	HPR	-12.222	< 2.22e-16	HPR
ZIPR	HPR	0.532	0.297	ZIPR/HPR
NBR	ZINBR	-7.820	< 2.639e-15	ZINBR
NBR	HNBR	-7.719	< 5.882e-15	HNBR
ZINBR	HNBR	0.024	0.490	ZINBR/HNBR
ZIPR	ZINBR	-7.829	2.461e-15	ZINBR
HPR	HNBR	-7.766	4.053e-15	HNBR

Table 3: Likelihood Ratio (LR) tests for the nested models ZINBR, MZINBR, MNZINBR.ERE, HNBR, MHNBR, and MHNBR.ERE models

Model 1	Model 2	LR Test Statistic	DF	p-value
ZINBR	MZINBR	128.40	1	< 2.2e-16
MZINBR	MZINBR.ERE	66.71	2	< 0.0001
HNBR	MHNBR	86.59	1	< 2.2e-16
MHNBR	MHNBR.ERE	109.81	2	< 0.0001

Table 4: DHARMa Uniformity Test of MZINBR, MNZINBR.ERE, MHNBR, and MHNBR.ERE models and the corresponding count-part (σ_c^2) and zero-part (σ_z^2) variance components

Model	MZINBR	MZINBR.ERE	MHNBR	MHNBR.ERE
D-Statistic	0.016	0.022	0.011	0.018
p-value	0.199	0.025	0.653	0.121
σ_c^2	0.074	0.060	0.064	0.066
σ_z^2	-	1.264	-	0.582
Log-likelihood	-8911.048	-8877.694	-8932.08	-8877.176

P<0.05 indicates the model doesn't fit well for the count data

Table 5: Estimated incidence rate ratio (IRR) of having ANC visits and odds ratio (OR) of not attending any ANC visit (with 95% CI and p-values) from the hurdle negative binomial regression with random intercept at both count- and zero-part (HNBR.ERE) models, BDHS 2014

Factors	Category	Count-part				Zero-part			
		IRR	95% CI		p-value	*OR*	95% CI		p-value
Region	Barisal [R]								
	Chittagong	0.907	0.80	1.03	0.133	1.226	0.83	1.82	0.309
	Dhaka	0.998	0.88	1.13	0.981	0.796	0.53	1.20	0.272
	Khulna	1.182	1.04	1.35	0.012	0.496	0.31	0.78	0.002
	Rajshahi	1.008	0.88	1.15	0.907	0.900	0.59	1.37	0.627
	Rangpur	1.239	1.09	1.41	0.001	0.629	0.41	0.96	0.032
	Sylhet	0.931	0.81	1.07	0.308	1.468	0.98	2.20	0.061
Place of Residence	Urban [R]								
	Rural	0.840	0.78	0.90	0.000				
Wealth Status	Poorest [R]								
	Poorer	1.048	0.95	1.15	0.348	0.804	0.64	1.01	0.065
	Middle	1.026	0.93	1.14	0.615	0.633	0.48	0.83	0.001
	Richer	1.155	1.04	1.28	0.007	0.401	0.29	0.55	0.000
	Richest	1.306	1.16	1.47	0.000	0.197	0.13	0.31	0.000
Mass Media Exposure	Not at all [R]								
	Less than once a week	1.100	0.99	1.22	0.064	0.766	0.57	1.03	0.082
	At least once a week	1.156	1.08	1.24	0.000	0.567	0.45	0.71	0.000
Mother's Education	Illiterate [R]								
	Primary	1.088	0.98	1.21	0.130	0.725	0.57	0.92	0.009
	Secondary	1.213	1.09	1.35	0.001	0.450	0.35	0.59	0.000
	Higher	1.430	1.25	1.63	0.000	0.243	0.14	0.43	0.000
Pregnancy wanted	No [R]								
	Yes	1.232	1.12	1.36	0.000				
Decision on health care	Woman alone [R]								
	Woman & husband	0.969	0.90	1.05	0.418				
	Husband alone	0.920	0.85	1.00	0.051				
	Other	0.870	0.78	0.97	0.012				
Partner's Education	Illiterate [R]								
	Primary	0.985	0.91	1.07	0.722	0.969	0.78	1.20	0.772
	Secondary	1.005	0.92	1.10	0.911	0.628	0.48	0.82	0.000
	Higher	1.092	0.98	1.21	0.103	0.474	0.30	0.75	0.001
Intercept		1.906	1.59	2.29	0.000	1.090	0.75	1.59	0.654

[R] refers to reference group

Identification of disaggregated hotspots of child morbidity in Bangladesh: An application of small area estimation method

Bappi Kumar[1], Sumonkanti Das[1,2], Luthful Alahi Kawsar[1]
[1]Department of Statistics, Shahjalal University of Science & Technology, Sylhet, Bangladesh
[2]Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands

Abstract

Acute respiratory infection (ARI) and diarrhoea are two major causes of child morbidity and mortality in Bangladesh. National and regional level prevalence of ARI and diarrhoea can be calculated from a nationwide survey, however prevalence at micro-level administrative units (say, district and sub-district) is not possible due to lack of data. In such case, small area estimation (SAE) methods are applied by combining a survey data with a contemporaneous census data. Using a SAE method for dichotomous response variable (parallel to the World Bank SAE method), this study aims to estimate the proportions of under-5 children experienced with ARI and diarrhoea separately as well as either ARI or diarrhoea within a period of two-week preceding the survey. The ARI and diarrhoea information extracted from Bangladesh Demographic and Health Survey 2011 are used to develop a random effect logistic model for each of the indicators, and then the prevalence is estimated following the World Bank SAE approach using 5% data of the Census 2011. The estimated prevalence of each indicator significantly varied by district and sub-district (as for example 1.4-11.3% for diarrhoea, 2.2-11.8% for ARI and 4.3-16.5% for ARI/diarrhoea at sub-district level). In a number of districts and sub-district, the proportions are found double than the national level. district and sub-district levels spatial distributions of the indicators might help the policy makers to identify the vulnerable disaggregated and remote hotspots. Particularly, aid industries can provide effective interventions at the highly vulnerable spots to overcome the gaps between micro and macro level administrative units.

Keywords

ARI; diarrhoea; dichotomous response variable; mixed effects logistic model; spatial distribution

1. Introduction

Diarrhoea and acute respiratory infection (ARI) are recognized as important causes of global child morbidity and mortality. Pneumonia (a form of ARI) and diarrhoea are still remained major causes of child death. The impact of these two diseases is about 29% of all under-5 child deaths which causes the loss of 2 million young lives each year (WHO, 2018). Only

pneumonia accounts about 16% of these young child deaths, though the number of deaths due to pneumonia can be reduced by the early diagnosis and treatment of ARI. The integrated Global Action Plan for Pneumonia and Diarrhoea (GAPPD) also aims to reduce mortality from pneumonia and diarrhoea in under-5 children to fewer than 3 per 1000 and 1 per 1000 live births respectively by 2025 (WHO & UNICEF, 2013). As per GAPPD, the solutions to tackling pneumonia and diarrhoea do not require major advances in technology since proven interventions are already existed. Children are dying because services are not provided at the required demand and the children at higher risk are not being reached. Use of effective interventions such as exclusive breastfed of first 6 months, proper life-saving treatment for the children with suspected pneumonia and providing oral rehydration therapy to the children with diarrhoea are still enough to achieve these goals. Identifying the children at greatest risk, hardest to reach and most neglected, and targeting them with interventions of proven efficacy will enable us to close the gap, ultimately ending the heavy toll of preventable child deaths.

The prevalence of ARI and diarrhoea at national and divisional levels are estimated usually from nationwide household survey. In Bangladesh, nationwide data on ARI and diarrhoea information are collected through household surveys conducted by Bangladesh Bureau of Statistics (BBS), and Demographic and Health Survey (DHS). Diarrhoea and ARI data are collected by asking mothers whether their children are experienced with diarrhoea episode or ARI symptom during a two-week period preceding the survey. In the recent 2014 Bangladesh DHS (BDHS) survey, the episodes of ARI and diarrhoea were found around 5% for both cases [NIPORT et al., 2016], while the rates were 6% and 5% respectively in 2011 BDHS [NIPORT et al., 2013]. Though national and divisional estimates of diarrhoea and ARI prevalence are estimable from the survey data, those at disaggregated level (such as district and sub-district) are not estimable solely from the survey data due to lack of observation at the desired micro level. Consequently, it is impossible to find out the disaggregated hotspots highly vulnerable to ARI and diarrhoea prevalence for the government. Identifying these disaggregated hotspots might help the concerned aid industries for targeting them with efficient interventions.

Small area estimation (SAE) is a statistical technique to obtain estimates of a target parameter with better precision at any desired disaggregated administrative units of a country. The basic idea of SAE method is to combine a survey data with a recent census or administrative data via a statistical model (Rao, 2003). Survey data consist of the target variable and a regression model is specified with some explanatory variables which are common in both survey and census data. The World Bank has been utilising an SAE method known as ELL (Elbers, Lanjouw, & Lanjouw, 2003) for poverty and nutrition mapping in

many developing countries including Bangladesh using a continuous response variable. The basic idea of the methodology is to develop a regression model using a continuous response variable (such as weight-for-age Z-score). Since the variable of interest for diarrhoea and ARI prevalence is dichotomous (whether a child has experienced with diarrhoea or not during a fixed time period) instead of continuous, the ELL methodology cannot be implemented. However, the basic idea can be implemented after developing a generalized linear mixed model (GLMM) more specifically a random effect logistic model for the dichotomous response variable (Haslett et al., 2014). The main difficulty is to develop a proper GLMM model incorporating the available survey data with a recent census or administrative data for a country. The main aim of this study is to estimate the prevalence of ARI and diarrhoea for under-5 children at district and sub-district levels using a SAE technique for dichotomous response variable. In addition, the proportion of children suffering either from diarrhoea or ARI during the 2-week period (hereafter refereed as ARI/diarrhoea) is also aimed to estimate at disaggregated levels.

Finally, disaggregated level spatial distributions of all the three indicators are mapped to highlight the most vulnerable hotspots.

2. Methodology

To explain the small area statistical methodology for the proposed research, let the child-level measurement Y_{ijk} is either 1 (occurrence of diarrhoea during a fixed time period) or 0 (non-occurrence of diarrhoea) and $P_{ijk} = P(Y_{ijk} = 1)$ represents the probability of having diarrhoea for k^{th} child belonging to j^{th} household of i^{th} cluster. According to ELL methodology, the first target is to develop a nested error logistic regression model of P_{ijk} as below:

$$\text{logit}\left(P_{ijk}\middle|X_{ijk}, \eta_i, \mu_{ij}\right) = \log\left[P_{ijk}/\left(1 - P_{ijk}\right)\right] = \mathbf{X}_{ijk}^T\boldsymbol{\beta} + \mu_{ij}$$

where $\boldsymbol{X_{ijk}}$ is vector of explanatory information, $\boldsymbol{\beta}$ is vector of regression parameters, and uij corresponds to cluster specific random errors respectively. The random errors are usually assumed to be independent and identically distributed with mean zero and constant variance. The considered GLMM can be fitted by estimating regression parameters and variance components (Rao, 2003). The regression model can be extended to higher level, however the ELL methodology assumes heterogeneity at cluster level rather than higher levels (ELL, 2003). Following the ELL approach, the small area estimates and their root mean squared error (RMSE) can be calculated via either a parametric or semi-parametric bootstrap procedure. In each bootstrap, response values P_{ijk}^* can be predicted for all census children by generating regression parameters and level-specific errors from their parametric (or empirical) distributions. The estimates at target level can be obtained by aggregating P_{ijk}^* belong to the level. The bootstrap procedure can be done for say B=500 times and then the

ultimate parameters with their RMSEs can be calculated by taking average and standard deviation of B estimates respectively.

The BDHS 2011 children data are aimed to combine with the Bangladesh Population and Housing Census 2011 data to conduct the study. The main reason for using BDHS 2011 instead of recent BDHS 2014 is that the sampling design of BDHS 2011 is based on the Census 2011 and sub-districts are unidentifiable in BDHS 2014. The survey data is collected following a two-stage stratified sampling design by covering all 7 divisions, 64 districts, and 396 (out of 544) sub-districts. In the BDHS 2011, a total of 8341 children were found whose ARI and diarrhoea information were available (NIPORT et al., 2013). The full census data of Bangladesh is unavailable for academic purposes, however, 5% of the full census data is available from BBS. A number of important socio-demographic characteristics such as age, sex, education, schooling, employment, disability and housing characteristics are available in the census data, and consequently some district and sub-district level contextual variables can be created and included in the model specification. In model development, two-way interactions of residence, division, sex, and age are utilized to develop the best models.

3. Results

Fixed effect logistic models (GLM) and random intercept logistic models (GLMM) are developed and compared to find an appropriate model for each of the health indicators. The final models with the corresponding inputs are utilized in the SAE approach to estimate the proportion of each child health indictor with their RMSEs and CVs. Table 1 shows that the two-level GLMM are performing better than the fixed effect logistic models (GLM) in terms of AIC, Likelihood Ratio Test (LRT), and area under the ROC curve (AUC) for each of the three child health indicators. In all cases, GLMM can classify children health status more correctly than the GLM, particularly for ARI about 73% children are correctly classified. Though the cluster-specific random errors are assumed to follow the normal distribution, the normality assumption is dissatisfied for diarrhoea and ARI since more than 300 clusters (out of 600) had zero prevalence. However, the assumption is satisfied for ARI/diarrhoea in which case only 100 clusters had zero prevalence. To avoid the impact of this non-normality, non-parametric bootstrap procedure was employed in the prediction of health indicators.

To examine the performance of SAE method at the higher administrative levels, division level estimates are also estimated and compared with the design-based direct estimates (hereafter referred as DIR). The plots under the first two columns of Figure 1 show that the ELL provides very similar estimates of the prevalence with higher accuracy measure (CV×100) than the DIR estimates at division level. Both DIR and ELL estimators indicate that the

children of Khulna division had less experience with the occurrence of diarrhoea however they were more experienced with ARI. While in Sylhet and Barisal, prevalence of ARI/diarrhoea was more frequent compared to the other divisions. The estimated district level prevalence with the estimated CVs are plotted against the district (ordered by the number of children) in the plots under the third and fourth columns of Figure 1. The plots under third column show that the ELL estimates go through the direct estimates, which indicates the ELL estimator provides approximately unbiased estimates. The ELL estimator provides considerably lower CVs compared to the direct estimator as expected.

Since the survey data covers only 396 out of 544 sub-districts and sub-district specific sample sizes are too small, only the ELL estimator is applied to calculate sub-district specific prevalence. Summary statistics of the estimated sub-district level prevalence of diarrhoea, ARI and ARI/diarrhoea shown in Table 2 indicate that the mean prevalence are respectively 4.91% (SD: 1.70%), 6.05% (SD: 1.34%) and

10.37% (SD: 2.08%). Sub-district level prevalence picked up to 11.3% for diarrhoea, 11.8% for ARI and 16.5% for ARI/diarrhoea. It is observed that about 25% sub-districts have more than 6.0% and 7.0% prevalence of diarrhoea and ARI respectively, while about 75% sub-districts have more than 9.0% prevalence of ARI/diarrhoea. Sub-district level estimates can be considered efficient based on the CV estimates, of which 75% are below 17% for diarrhoea, 14% for ARI and 11% for ARI/diarrhoea. Figure 2 shows district and sub-district level maps of Bangladesh for the considered three indicators. The first map (a.1) shows that the districts of western-south (Khulna region) had lower diarrhoea prevalence, while the districts of northern (Mymensingh region), north-eastern (Sylhet region) and south-eastern (Chittagong region) had comparatively higher diarrhoea prevalence. More specifically, the highest diarrhoea prevalence is found in Cox's Bazar (8%) and lowest in Joypurhat district (3%). Map (a.2) shows that the districts of Southern region particularly coastal areas are more vulnerable to ARI compared to north-eastern parts of Bangladesh. The distribution of ARI/Diarrhoea prevalence shown in the map (a.3) reveals very similar distribution as for the diarrhoea prevalence. The sub-district level map for diarrhoea (b.1) indicates that a significant number of sub-districts particularly from Sunamganj, Sylhet, Chandpur, and Cox's Bazar districts have about 7% prevalence of diarrhoea, which is more than about 1.5 times of the national level. For ARI, sub-districts with prevalence greater than 6.8% were found more or less scattered all over the districts. Sub-district level map of ARI/diarrhoea shows that the sub-districts vulnerable from diarrhoea were also vulnerable from ARI/diarrhoea (such as Halishahar, Teknaf, Companiganj, Ramu, and Ukhia sub-districts are vulnerable from both diarrhoea as well as ARI/diarrhoea).

4. Discussion and Conclusion

The study shows disaggregated level (district and sub-district) prevalence of ARI and diarrheal episodes with the accuracy measures using a SAE method for dichotomous response variable. The findings suggest that the SAE method for the dichotomous variables are providing unbiased and accurate estimates compared to the direct estimator. Based on the study, it can be recommended for the policy makers that though the national level even division level estimates seem very low, there are significant inequalities in health indicators among the districts and sub-district. A comparison between district and sub-district level maps shows that which sub-districts contributes more for higher prevalence of any indicator (say, diarrhoea) for a specific district. The variation becomes higher when the level of administrative units goes down. Since the study find a suitable multilevel model for each of the health indicators using the 5% Census data, which can be used more accurately for the lower administrative units (say union, sum of clusters) if the full census data sets were available. Both ARI and diarrhoea are highly correlated with acute child malnutrition known as wasting. A national target is to reduce the wasting level at 5% by the year of 2025. The target of reducing mortality from pneumonia and diarrhoea along with wasting among under-5 children, the incidence of ARI and diarrhoea should be monitored at different administrative tiers so that the targets can be reached by the time.

References

1. WHO (2018). WHO Global Health Observatory. World Health Organization. Retrieved from httt://www.who.int/gho/child_health/en/index.html.
2. WHO & UNICEF (2013). Ending preventable child deaths from pneumonia and diarrhoea by 2025: The integrated Global Action Plan for Pneumonia and Diarrhoea (GAPPD). France: World Health Organization and UNICEF.
3. NIPORT, Mitra and Associates, & ICF International. (2013). *Bangladesh Demographic and Health Survey 2011.* Dhaka, Bangladesh and Calverton, Maryland, USA: NIPORT, Mitra and Associates, and ICF International.
4. NIPORT, Mitra and Associates, & ICF International. (2016). *Bangladesh Demographic and Health Survey 2014.* Dhaka, Bangladesh and Rockville, Maryland, U.S.A. : NIPORT, Mitra and Associates, and ICF International.
5. Rao, J. (2003). *Small Area Estimation.* New Jersey: John Wiley & Sons, Inc.
6. Elbers, C., Lanjouw, J. O. & Lanjouw, P., 2003. 'Micro-Level Estimation of Poverty and Inequality'. *Econometrica,* 71(1), p. 355–364.
7. Haslett, S., Jones, G., Isidro, M. & Sefton, A., 2014. *Small Area Estimation of Food Insecurity and Undernutrition in Nepal,* Kathmandu: Central

Bureau of Statistics, National Planning Commissions Secretariat, World Food Programme, UNICEF & World Bank.

Table 1. Summary statistics and diagnostics of the fitted logistic (GLM) and random intercept logistic (GLMM) models for Diarrhoea, ARI, and ARI/diarrhoea prevalence, BDHS 2011♀

Health Indicators	n	P	Model	AIC	logLik	$\sigma_u 2$	LRT of H_0:$\sigma_u^2 = 0$	AUC (%)
Diarrhoea	8341	21	GLM	3111.8	-1534.9	-	χ^2: 3.7120	54.86
			GLMM	3110.1	-1533.0	0.1687	p-value: 0.0215	71.55
ARI	8341	22	GLM	3680.2	-1818.1	-	χ^2: 6.2343	61.11
			GLMM	3676.0	-1815.0	0.1921	p-value: 0.0063	72.26
ARI/Diarrhoea	8341	21	GLM	5371.6	-2664.8	-	χ^2: 2.7519	58.41
			GLMM	5370.8	-2663.4	0.0791	p-value: 0.0448	65.78

♀ n=sample size, p= # of covariates, σ_u^2=cluster-level variance component, LRT= likelihood ratio test, AUC= Area under the receiving operating characteristics curve

Table 2. Summary statistics of diarrhoea, ARI, and ARI/diarrhoea prevalence among under-5 children at sub-district level with their root mean squared errors (RMSE) and CV using ELL estimator

Health Indicators		Minimum	Q1	Mean	Median	Q3	SD	Maximum
Diarrhoea	Prevalence (%)	1.38	3.57	4.91	4.66	5.96	1.70	11.31
	RMSE x 1000	3.93	4.80	7.36	6.01	8.36	3.77	27.80
	CV (%)	8.08	11.07	15.61	13.32	16.85	6.91	45.35
ARI	Prevalence (%)	2.21	5.21	6.05	6.04	6.79	1.34	11.83
	RMSE x 1000	4.57	5.91	7.86	6.82	9.04	2.91	21.30
	CV (%)	7.27	9.85	13.64	11.36	14.22	6.30	43.33
ARI/ Diarrhoea	Prevalence (%)	4.27	8.96	10.37	10.37	11.85	2.08	16.45
	RMSE x 1000	6.05	7.94	10.11	9.14	11.27	3.28	28.25
	CV (%)	5.66	7.67	10.17	8.89	10.77	4.05	28.58

Figure 1. Prevalence of diarrhoea, ARI, and ARI/diarrhoea among under-5 children at division and district level in Bangladesh with their coefficient of variations (CV) standard errors (RMSE) using direct (DIR, black lines) and SAE (ELL, red lines) estimators

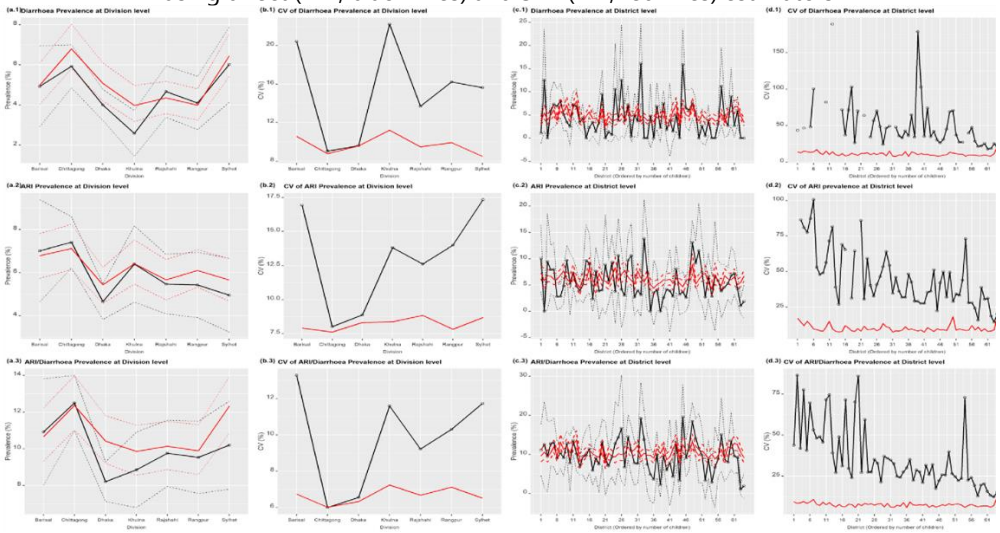
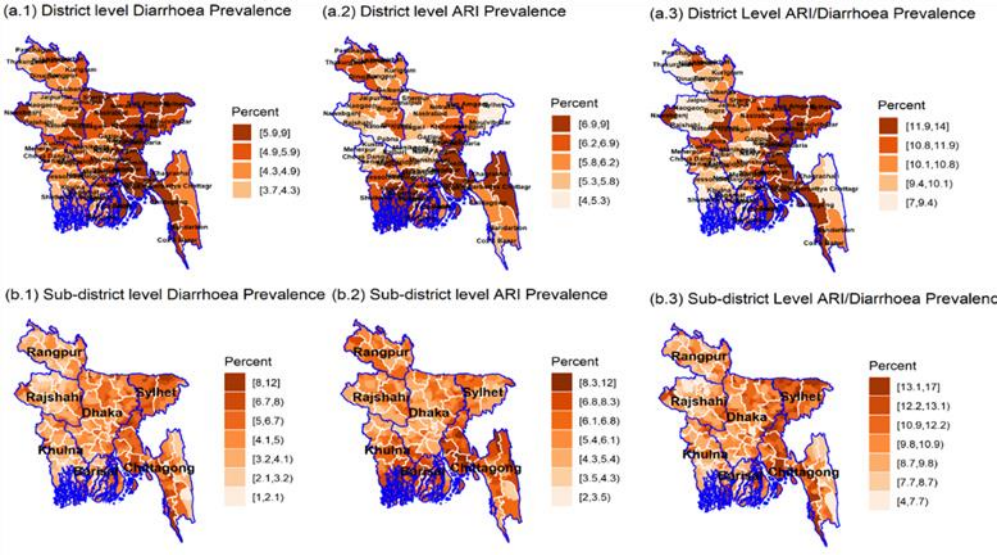


Figure 2. District and sub-district level hotspots of diarrhoea, ARI, and ARI/diarrhoea prevalence among under-5 children of Bangladesh



Mobility and state dependence in a labour market characterized by informality: the case of Morocco

Mustapha Ziroili
High Commission for Planning
LEST-Aix-Marseille University

Abstract

The informal economy plays an important role in creating jobs in the Moroccan labor market. The purpose of this research paper is to analyze and understand one of the aspects related to this phenomenon that is mobility and the degree of state dependence. To this end, I created a pseudopanel, using matching methods, of three waves of labour force survey covering the period 2013-2015.Then, to assess the extent of the state dependence to informal employment, I estimated a random effect probit model. To address the problem of initial conditions I used two specifications, Orme (1997) and Wooldridge (2005). The different estimations show that an employee who had an informal job at time t-1 has a 7.2% higher probability of being in informal employment at time t compared to an employee who held a formal job at the moment t-1. This average partial effect is greater for men (+ 7.2%) than for women (+ 5.9%).

Keywords

Informal employment; state dependence; dynamic probit with random effect; unobserved heterogeneity; initial conditions; Morocco

1. Introduction

The informal economy plays an important role in job creation, income generation and poverty reduction in many countries, particularly in developing and transition countries. In Morocco, the analysis of employment cannot be properly done without taking into account the dimension of informality, as the share of the informal employment is predominant in the total employment and the economic activity of the country. The importance of the informal sphere in the case of Morocco, as in most developing economies, stems from the opportunities it offers to the most vulnerable populations, such as the poor, women and young people.

The debate on the informal economy began in the 1970s and since then has been channelled around four schools of thought: dualistic, structuralist, legalistic and voluntarist. These schools have developed differentiated analyzes of the links between the formal and informal sectors. The first school of thought was the dualistic school, which characterized the thinking of the ILO in the 1970s (Hart, Tokman, Sethuraman); this line of thinking suggested that the informal sector included peripheral activities and was not related to

the formal sector. This theory is also known as the segmentation of the labour market or fragmentation of the labour market (Dixit, 1973), the essence of the dualistic school being the distinction between these two components of the labor market. One refers, according to the literature, to a sector called "formal", "modern" or "urban" while the other is alternatively called "informal", "traditional" or "rural" (Fields, 2007). The second school of thought on informality is the structuralist school that was defined in the 1980s through the writings of Moser (1978) and Portes (1983). The underlying idea is that the informal sector should be seen as composed of subordinate economic units and workers that serve to reduce input and labor costs and thus increase the competitiveness of large capitalist enterprises. The legalistic and somewhat bureaucratic school postulates that the informal sector is composed of micro-entrepreneurs who choose to operate informally in order to avoid the costs, time and effort of formal registration (Soto, 1989). The fourth line of thinking stems from the "voluntarist" school; and considers that entrepreneurs' informality choice is to avoid regulations and taxes in order to reduce costs (Perry et al. 2007). In this paper, I am more interested in the first school of thought related to the segmented nature of the labor market.

The paper is structured as follows. Section 2 presents the data, the construction of the pseudo-panel used in the study and declines the methodology used. Estimation results are presented in section 3, while section 4 concludes by discussion.

2. Methodology

Data and construction of the panel

Data used in this paper come from the Moroccan labour force survey (2013-2014-2015). Between 2013 and 2014 there is a rotating sample (panel with two observation). To have a panel data with more than two observations I created a pseudo-panel from the rotating sample 2013/2014 and the data of 2015. For this end, I used matching technics to find similar people between the true panel 2013/2014 and 2015. More precisely I used nearest neighbor matching method without replacement. The variables used for the matching process are educational level, gender and age of access to the labour market. After matching I obtained a balanced sample of total size of 10488 (8802 for men and 1686 for women).

For the definition of informal employment we follow the guidelines of ILO and we have taken as a proxy of informality the lack of medical coverage related to the job performed.

The model used

I estimated two specifications for the dynamic probit random effects model that allows for consideration of state dependence and unobserved heterogeneity.

Basic specification For modelling the state dependence in access to a formal employment, the analysis starts with the specification of an unobserved general effect in the form:

$$\mathrm{P}(y_{it}|y_{it-1},\ldots,y_{i0},x_i,\alpha_i) = \Phi(\mathrm{x}_{it}\beta + \lambda y_{it-1} + \alpha_i \qquad (1)$$

Where x_{it} is a vector of strictly exogenous explanatory variables and β is the vector of the associated coefficients. Under this formulation, the probability of access to a formal job depends on the unobserved heterogeneity (α_i) and past situation (dependent variable lagged by one period.y_{it-1}).Φ is the cumulative distribution function of the standard normal distribution. Testing the hypothesis of a non-zero λ is equivalent to testing for the presence of a true state dependence having controlled the unobserved heterogeneity. The model above can also be expressed as follows:

$$f(y_1,y_2,\ldots,y_{\mathrm{T}}|y_0,x,\alpha) = \prod_{t=1}^{T} f(y_t|y_{t-1},\ldots,y_1,y_0,x_1,\alpha) \qquad (2)$$

$$= \prod_{t=1}^{T}[\Phi(x_t\beta + \lambda y_{it-1} + \alpha]^{y_t}\,[1 - \Phi(x_t\beta + \lambda y_{it-1} + \alpha]^{1-y_t}$$

A model of reduced and dynamic form is given by:

$$y_{it} = 1[x'_{it}\beta + \lambda y_{it-1} + \varepsilon_i + \mu_{it} > 0]\; i = 1,\ldots,\mathrm{N}; t = 2,\ldots,\mathrm{T} \qquad (3)$$

The dependent variable y_{it} is a binary variable that is equal to unity if the individual i has a formal job in period t and zero if he has an informal job. The parameter $\beta\ and\ \lambda$ are unknown and to be estimated. The term ε_i refers to unobservable and time-invariant individual characteristics and μ_{it} is an error term that can vary according to the individuals (i) or the years (t): $\mu_{it} \sim \mathrm{N}(0.\sigma_\mu^2)$.

The estimation of a standard model with uncorrelated random effects implicitly assumes a zero correlation between the term ($\varepsilon_i + \mu_{it}$ and the set of explanatory variables. However, this hypothesis may not be sustainable in this context. For example, if we take an individual's motivation, it can be correlated with at least human capital variables such as degree level. Mundlak (1978) and Chamberlin (1984) propose another estimator that allows a correlation between $\varepsilon_i\ and\ x_{it}$ according to a linear relation between the term ε_i and the mean of the explanatory variables in the following form:

$$\varepsilon_i = \delta_0 + \overline{x_{it}}\delta' + e_i \text{ with } e_i \sim \text{iid } \mathrm{N}(0,\sigma_e^2) \qquad (4)$$

By replacing (4) in (3) we will have:

$$y_{it} = 1[x'_{it}\beta + \lambda y_{it-1} + \overline{x}_{it}\delta' + e_i + \mu_{it} > 0]\; i = 1,\ldots,N; t = 2,\ldots,T \qquad (5)$$

The correlation between errors over the different periods is approximated by: $\text{corr}(\varepsilon_i + \mu_{it}, \varepsilon_i + \mu_{it-1}) = \rho = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + 1}$, $t = 2, \ldots, \text{T}$. Assuming that $\sigma_\mu^2 = 1$ for standardization concern.

The introduction of the lagged dependent variable could be a source of endogeneity beacause of the correlation between α_i the and the y_{i0} .Specifically, it is the simultaneous presence of "initial conditions" and "unobserved heterogeneity" in the model equation, which can be correlated and, as a result, leads to an overestimation of the state dependence effect. Therefore, the problem of "initial conditions", caused by our lack of knowledge of the process determining initial labour force status, should be addressed. One of specifications for solving this problem is proposed by Wooldridge (2005)[1] .

Specification of Wooldridge (2005)

Wooldridge (2005) proposes the maximum likelihood estimator conditional on initial conditions and explanatory variables. Formally, instead of modeling the initial conditions conditionally to the observed and unobserved characteristics, Wooldridge proposes to model the unobserved individual heterogeneity conditionally to observed heterogeneity and initial conditions:

$$\varepsilon_i = \xi_0 + \xi_1 y_{i1} + \tau_i \qquad (13)$$

We introduce (13) in (5), while keeping the specification of Mundlak, we obtain:

$$y_{it} = 1[x'_{it}\beta + \lambda y_{it-1} + \overline{x}_{it}\delta' + \xi_i y_{i1} + \tau_i + \mu_{it} > 0] \; i = 1, \ldots, N; t = 2, \ldots, T \qquad (14)$$

The extent of the state dependence

The random dynamic probit model allows to calculate the transition probability for all individuals between different statuses conditional on their situations in t-1. Therefore, we can calculate the probability of staying on the same status:

$$\text{rest}_{it} = Pr\left(y_{it} = 1 / y_{it-1} = 1, Z_{it}\right) = \Phi\left(\frac{\gamma' Z_{it} + \lambda}{(1-\rho)^{1/2}}\right)$$

And the probability of changing status:

$$\text{sort}_{it} = Pr\left(y_{it} = 1 / y_{it-1} = 0, Z_{it}\right) = \Phi\left(\frac{\gamma' Z_{it}}{(1-\rho)^{1/2}}\right)$$

The extent of state dependence is assessed using the Average Partial Effect (APE) and the Predicted Probability Ratio (PPR). The starting point is to calculate the probabilities "rest" and "sort" for each individual in the sample, then we calculate the averages of these probabilities in relation to all individuals and finally, the difference between these two means is the APE and the ratio between the two is the PPR.

[1] Orme (1997) specification was estimated also.

3. Results

In addition to the lagged dependent variable, which measures state dependence, the model includes initial conditions, representing the situation encountered at the beginning of the process, as well as the gender, family situation, age, level of diploma, characteristics of employment and controls for the economic situation in the region where the individual works, in particular urban and rural unemployment.

The estimate reported in Table 2 shows a similarity between the results of the different specifications and especially in terms of the extent of state dependence to informal employment in the past. If we focus on the estimates of the Wooldridge (2005) specification, we can find the following results:

State dependency

The coefficient related to the lagged dependent variable is very significant and positive. Experiencing an informal status in period t-1 has a significant impact on the current employment situation. In fact, the probability of taking up informal employment increases by 7.2% if the individual has gone through informal employment, which confirms state dependence on this phenomenon.

State dependence and gender

The state dependence is more important for men (7.2%) than for women (5.9%). This result confirms one of the stylized facts of the Moroccan labour market, which is characterized by a strong presence of men in informal activities. The low level of state dependence among women is related to their limited participation in economic activity. Several studies (Taamouti and Ziroili (2011), Verme et al. (2014), World Bank and HCP (2017)) have analysed women's participation in the labour market and the results converge towards a negative impact of marriage, social norms and the type of growth that is unable to generate enough decent jobs that encourage women to work, especially for urban women.

Table 1 : Average partial effect and predicted probability ratio

lag_informel	Orme (1997)		Wooldridge (2005)	
	APE	PPR	APE	PPR
Total	0.0779*** (0.0130)	1,2	0.0724*** (0.0127)	1,1
Women	0.0579** (0.0287)	1,2	0.0593** (0.0290)	1,2
Men	0.0789*** (0.0142)	1,2	0.0717*** (0.0138)	1,1

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

(1): APE : Average partial effect: difference between the average probability of being in informal employment at time t conditionally to be in informal employment at time t-1 and the average probability of being in formal employment at time t conditionally to be in informal employment at the moment t-1

(2): PPR : predicted probability ratio: the ratio between the two probabilities mentioned in (1)

Table 2 : results of estimates of the probability of informal employment2

VARIABLES	Orme (1997) Total (1)	Women (2)	Men (3)	Wooldridge (2005) Total (4)	Women (5)	Men (6)
lag_informal (t-1)	0.700***	0.693**	0.697***	0.663***	0.712***	0.651***
	(0.0859)	(0.269)	(0.0919)	(0.0871)	(0.270)	(0.0937)
female	-0.231**	-	-	-0.264**	-	-
	(0.116)			(0.117)		
Age	-0.0742**	-0.0722	-0.0769**	-0.0796**	-0.0631	-0.0819**
	(0.0318)	(0.105)	(0.0342)	(0.0322)	(0.104)	(0.0348)
Age^2	0.000503	0.00126	0.000432	0.000536	0.00113	0.000456
	(0.00031	(0.0010	(0.00033	*	(0.0010	(0.00034
	4)	9)	7)	(0.00031	8)	4)
				9)		
Married	-0.331***	-0.428	-0.313**	-0.372***	-0.354	-0.351**
	(0.110)	(0.280)	(0.144)	(0.111)	(0.278)	(0.146)
Head of household	-0.0896	0.00968	-0.131	-0.112	-0.0188	-0.170
	(0.122)	(0.342)	(0.160)	(0.124)	(0.346)	(0.162)
Level of diploma						
	Modalité de Référence "Sans diplôme"					
Fondamental	-0.109	-0.352	-0.0552	-0.144	-0.321	-0.0799
	(0.0949)	(0.331)	(0.102)	(0.0958)	(0.325)	(0.104)
secondary	0.0721	0.142	0.105	-0.0141	0.0319	0.0361
	(0.279)	(0.743)	(0.313)	(0.283)	(0.760)	(0.318)
High education	-0.572**	-1.482**	-0.216	-0.682**	-1.539**	-0.328
	(0.276)	(0.651)	(0.337)	(0.278)	(0.641)	(0.343)
Technicians	-0.474***	-1.260**	-0.290	-0.600***	-1.353**	-0.407*
	(0.189)	(0.585)	(0.209)	(0.188)	(0.560)	(0.210)
Gen_residual	-0.168***	-0.0392	-0.185***	--	--	--
	(0.0558)	(0.185)	(0.0596)			
informal2013	--	--	--	-0.209**	-0.134	-0.215**
				(0..950)	(0.309)	(0.102)
Constant	5.884***	15.85	5.550***	6.372***	16.33	6.090***
	(0.847)	(615.9)	(0.894)	(0.843)	(1.028)	(0.896)
σ_μ	0.888***	1.027***	0.883***	0.916***	1.066***	0.920***
	(0.095)	(0.311)	(0.102)	(0.096)	(0.314)	(0.103)
$\rho\ (rho)$	0.440***	0.513***	0.438***	0.456***	0.532***	0.458***
	(0.0529)	(0.151)	(0.057)	(0.052)	(0.147)	(0.055)

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

[2] We kept in this table only individual and human capital variables in order to respect the 6-pages requirement. The other variables related to job as well as household characteristics will be presented in the oral presentation.

4. Discussion and Conclusion

Effect of human capital

In terms of qualification, investing in lower levels of education cannot get out of the situation of informality. Only access to a higher-level diploma or a technical diploma (vocational training) breaks the barriers to access to formal employment. Having a higher education diploma reduces the chances of access to informal employment by almost 7% and a technical diploma reduces this probability by 6.2%. It is remarkable that the holding of these diplomas is more in favour for women. In fact, the first level of education reduces their probability of having an informal job by 12% (compared to 3% for men) and for the second type of education, by 10.7% (compared to 4.2% for men). Investment in education has a greater impact on women's access to employment in general and, in particular, to decent work. On the other hand, having a secondary level of education or less does not help to get out of the trap of informality. This leads us to discuss the causes of the low level of human capital of the employed, at least from a quantitative point of view, which could be related to the inequalities of opportunity of access to school or the completion of educational path to a higher level degree.

Effect of the characteristics of jobs held

We have just shown that the degree level allows the best educated to escape from the informal sphere, but it is still necessary to have a labour market that values human capital by a greater demand for skilled workers. In my model I introduced variables that capture the quality of jobs (decent work). One of the variables that refers to the stability and sustainability of work is the type of contract. Employees working with a contract are less likely to have an informal job (with a probability of 30.2% lower). In addition, the chances of having an informal employment are becoming smaller with the size of the firm. Regularity of employment has a significant impact on the transition to informal employment, especially for women; the probability of taking up informal work is reduced by 64.5% if one is in a permanent and full-time job (compared to -19.5% for men). The quality of the job held gives a signal to future recruiters as to the qualification and performance of the employee. In addition, occupational groups and sectors of economic activity have a positive and significant influence on integration into informal employment. Compared to managers and senior managers, being employee increases the probability of informal employment by 4.2% (3.5% for women versus 4.7% for men); this probability rises to 9.9% for non-agricultural labourers (11.2% for women versus 9.8% for men). According to sectors of economic activity, Apart from work in the public administration, which reduces the probability of gaining access to informal employment, employment in other branches of economic

activity increases the risks of taking up informal employment in reference to a job in the industry sector.

References

1. Akay, A. and Khamis, M. (2012). 'The persistence of informality: Evidence from panel data', in Lehmann, H. and Tatsiramos, K. (eds.), Informal Employment in Emerging and Transition Economies, Volume 34 of Research in Labor Economics, chapter 7, Bonn: Emerald, pp. 229–256.
2. Bosch, M. and Maloney, W. F. (2010). 'Comparative analysis of labor market dynamics using Markov processes: An application to informality', Labour Economics, 17(4), pp. 621–631.
3. Caliendo, Marco & Kopeinig, Sabine. (2008). Some Practical Guidance For the Implementation of Propensity Score Matching. Journal of Economic Surveys. 22. 31-72. 10.1111/j.1467- 6419.2007.00527.x.
4. Cappellari, L and Jenkins, SP (2008) "The dynamics of social assistance receipt: measurement and modelling issues, with an application to Britain", prepared under contract JA0004519, ELS/SPD Division, Organisation for Economic Cooperation and Development, September 2008, 70 pp. Report released as OECD Social, Employment and Migration Working Paper 67, http://www.oecd.org/dataoecd/30/42/41414013.pdf.
5. Edwin Leuven & Barbara Sianesi, 2003. "PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing," Statistical Software Components S432001, Boston College Department of Economics, revised 01 Feb 2018. Annexes
6. Fields, G. (1990). 'Labour market modelling and the urban informal sector: Theory and evidence', in Turnham, D., Salome, B. and Schwartz, A. (eds.), The Informal Sector Revisited, Paris: OECD. pp. 49–69.
7. Fields, G. (2009). 'Segmented labor market models in developing countries', in Ross, D. and Kinkaid, H. (eds.), The Oxford Handbook of Philosophy of Economics, Oxford: Oxford University Press, pp. 476–510.
8. Maloney, W. (1999). 'Does informality imply segmentation in urban labor markets? Evidence from sectoral transitions in Mexico', The World Bank Economic Review, 13(2), pp. 275–302.
9. Maloney, W. (2004). 'Informality revisited', World Development, 32(7), pp. 1159–1178.
10. Moser, Caroline O. N. 1978 "Informal sector or petty commodity production? Dualism or dependence in urban development?" World Development 6(9-10). 1041-1064.
11. Mundlak, Yair (1978), 'On the pooling of time series and cross section data', Econometrica, 46:69– 85.

12. Orme, Christopher D. (1997). 'The initial conditions problem and two-step estimation in discrete panel data models', Discussion Paper No. 9633, School of Social Sciences, University of Manchester. Revised version, June 2001, retitled as: 'Two-Step inference in dynamic non-linear panel data models'
13. Perry, G., Maloney, W., Arias, O., Fajnzylber, P., Mason, A. and Saavedra-Chanduvi, J. (2007). Informality: Exit and Exclusion. Washington, DC: The World Bank.
14. Portes, Alejandro 1983 "The informal sector: definition, controversy, and relation to national development." Review 7(1): 151-174.
15. Portes, A., Castells M., and Lauren A. Benton (eds.), The Informal Economy: Studies in Advanced and Less Developed Countries. Baltimore: Johns Hopkins University Press
16. Stewart, Mark B. (2006), 'Maximum simulated likelihood estimation of random effects dynamic probit models with autocorrelated errors', Stata Journal, 6: 256–272.
17. Stewart, Mark B. (2007), 'The interrelated dynamics of unemployment and low pay', Journal of Applied Econometrics, 22: 511–531.
18. Wooldridge, Jeffery M. (2005), 'Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity', Journal of Applied Econometrics, 20: 39–54.

Short-distance and long-distance elderly migration in Indonesia year 2016: Application of multilevel logistic regression analysis

Kadek Swarniati,S.ST[1] , Dr. Alfatihah Reno Maulani Nuryaningsih S. P. M.S.ST.,M.Si[2]
[1]BPS-Statistics Indonesia of South Jakarta Municipalities,Jakarta, Indonesia
[2]BPS-Statistics Indonesia, Jakarta, Indonesia

Abstract
This research examined the migration patterns of elderly in Indonesia and the determinants associated with migration. Logistic model of migration is estimated to ascertain whether the determinants of elderly migration differ for the choice of moving short or long distances. Multilevel models are used to examine the determinants of long-distance migration of elderly in Indonesia, using individual data from the Indonesia-National Socio-Economic Survey (SUSENAS) 2016 together with area-level data from external sources. Elderly short-distance migration occurred primarily in Java Island and North Sumatera. Most long-distance elderly migrants came from the metropolitan area, DKI Jakarta provinces. East Java and Central Java Province were the most frequent destinations for elderly migrants. Age of elderly, region, educational level, and working status of elderly were associated with short-distance migration. While individual-level variables are the strongest predictors of migration, contextual variables and multilevel interactions improve the explanatory power of the models. About 11,94 percent of the diversity of elderly migration in Indonesia due to the difference in the interprovincial characteristics. Effects contextual variables, such as crime rates, and cost of food, were positively associated with long-distance migration decision of elderly. The results of this study underscore the need to look beyond the influence of individual-level factors in addressing regional alvariations in elderly migration in Indonesia.

Keywords
Elderly migrant; long-distance migration; short-distance migration; Multilevel Logistic Regression; migration pattern.

1. Introduction

Demographic aging and immigration have been dominant process in sociodemographic change. Both the processes have an impact on the age and ethnic composition of the national population. This process both underlines wider socio-economic trends concerning changes in population as well as changes in habits of an increasingly diverse population. According to Warnes et al, 2004, these social processes brings several implications: (1) the number

of older people who have been international migrants and have cultural differences from the host population have grown and will undoubtedly increase during the coming decades; and (2) the case for a more sympathetic and proactive responses to the problems and structured disadvantages of older people migrant is becoming more compelling. The implications of population aging for migration and the links between the two are becoming increasingly important. Aging has reached an unprecedented scale and will continue to increase. According to the UN, the number of people over 60 years old has been growing rapidly and is projected to grow further: from 810 million in 2012 to more than 2 billion by 2050. The number of older people is due to surpass children for the first time in history (Zaiceva, 2014).

This condition also occurs in Indonesia. The population growth causes a large social economic impact that must be faced by Indonesia. Based on the population projection results, the estimated total population of Indonesia in 2016 was 258.70 million. This shows that the total population has increased by 1 million since 2010 (238.52 million). Meanwhile, in 2035 the population is estimated to be 305.65 million (BPS, 2017). A large and quality population can be a potential economic driver. But, if the population is not qualified, it will only become a burden of development. One phenomenon that requires greater attention is the growth of the elderly population.

A country has an old structured population if it has more than ten percent of the elderly population (Adioetomo, 2013). In 2016, the percentage of elderly people in Indonesia has reached 8.69 percent of the total population. This shows that Indonesia does not include a country with an aging population yet. Indonesia is currently transitioning to a population aging process which is characterized by an increase in the number and percentage of the elderly population. The elderly population in Indonesia has increased significantly. Total elderly population increased to 18.04 million people (7.56 percent) in 2010. BPS-Statistics Indonesia projects the number of elderly people in 2020 to be 27.09 million people or 9.99 percent of Indonesia's population. In 2035 this number is predicted to increase to 48.20 million people or 15.77 percent of the total population.

Increasing the number of elderly people is one indicator of the success of achieving national human development, as well as challenges in development. In other words, this large population is predicted to become the second demographic bonus for Indonesia.

Because of changing family support systems and diversified elder care choices, an increasing number of people choose to migrate in their later life (Lu & Song, 2006; Omelaniuk, 2005; Tong & Piotrowski, 2012). When the elderly migrate, they are not only likely to find a better place to live, they also face many challenges in health care and social welfare systems.

The migration decision of elderly is being influenced by both push and pull factors that arise from their living environment. Identifying these factors are needed to help Indonesia's policy makers and government administrators to recognize and understand the issues confronting older migrants. Besides that, it will help the government in creating aging-friendly initiatives that help the elderly to achieve their goal of moving.

This study focuses on elderly migrants who had been growing old in their host region and those who return to the homeland. This study aims to analyze the migration patterns among the elderly in Indonesia. This study also aims to examine the effect of individual and contextual factors on elderly migration and to know the variation of elderly migration which is caused by differences in characteristics between provinces in Indonesia.

2. Methodology

The targeted population in this study was elderly, 60 years and above. Based on their departure and current residential locations, individuals who migrated inside the same province were defined as "short-distance migrants," and those who migrate inter-provincially were defined as "long-distance migrants". The contextual characteristics of interest included: (1) Robbery rate by province; (2) Per capita food expenditure per month in each province; and (3) Percentage of the elderly family builder in each province. The individual variables included: (1) age of elderly 5 years ago (before migrate); (2) sex of elderly; (3) region; (4) education level; and (5) working status.

This study used multilevel logistic regression. The steps of Multilevel Logistic Regression included: (1) Simultaneous significance test of parameters; (2) partial significance test of parameters; (3) Interpretation of parameter using odd ratio; and (4) calculate intraclass correlation coefficient (ICC).Multilevel analysis is an analysis to see the relationship between individual variables and variables (Hox, 2010). The multilevel model used in this study is a hierarchical model with a random intercept because this research wants to see the variations between level two units (provinces) that affect elderly migration in Indonesia.

Binary Multilevel Logit Two Levels

Two level multilevel model (Hox. 2010) :

$$\ln\left(\frac{\pi ij}{1-\pi ij}\right) = \gamma_{00} + \sum_{p=1}^{p} \gamma_{p^0} X_{pij} + \sum_{q=1}^{Q} \gamma_{0q} \, Z_{qj} + \mu_{0j}$$

γ_{00} :*intercept* (overall average)

γ_{p0} : fixed effect for pth explanatory variable at 1st level (*fixed slope*), p = 1,2,...P

γ_{0q} : fixed effect for qth explanatory variable at 2nd level, q = 1,2,...Q

X_{pij} : the pth explanatory variable at 1st level for the ith individual at 1st level in the jth group at 2nd level , p = 1,2,...,P

Z_{qj} : q[th] explanatory variable at 2[nd] level for the j[th]-group, q = 1,2,...,Q
u_{0j} : *random effect* of j[th] group at 2[nd] level

Significance Testing of Random Effect

Test statistics:

$$LR = -2 \ln\left[\frac{likelihood\ \ logistic\ \ model\ without\ \ random\ effect}{likelihood\ logistic\ model\ with\ \ random\ effect}\right] \sim \chi^2_{(1)}$$

Simultaneous Test

Chi-square with the free degree of freedom according to the number of parameters

$$G = -2 \ln\left[\frac{likelihood\ \ without\ explanatory\ variables}{likelihood\ \ with\ explanatory\ variables}\right] \sim \chi^2_v$$

Partial Test

Test statistics:

$$W_P = \frac{\hat{\gamma}_{0q}}{SE(\hat{\gamma}_{0q})} \sim z$$

The test statistics used for 1[st] levels and 2[nd] levels are W, which is assumed to follow a normal distribution.

Intraclass Correlation (ICC)

$$\rho = \frac{\sigma^2_{\mu_0}}{\sigma^2_{\mu_0} + \sigma^2_{\varepsilon_0}}, 0 < \rho < 1$$

$\sigma^2_{\mu_0}$ is the error variance at 2[nd] level, whereas $\sigma^2_{\varepsilon_0}$ is the error variance at 1[st] level. The data preparation and analysis were processed by software: SPSS 20.0, Microsoft Excel 2013, and Stata 14.0

3. Results

A. Demographic Characteristics of Migrants

In 2016, approximately 1.2 % of SUSENAS participants aged 60 and older had migrated either short or long distance. About 0.7 % had made short-distance moves and 0.5% had made long-distance moves. In all types of migration, elderly who migrate were more likely elderly in productive age (60-64 year) rather than elderly in non-productive age. Only about one-fifth of the total sample of the elderly (18.2 %) lived in urban areas in the year 2012 (before migrate).

Elderly aged 60-64 or in productive age were more likely to make short distance migration than elderly aged 65 and over. In both short-distance and long-distance migration, elderly living in urban area, with high education level, and didn't have a job more likely to migrate

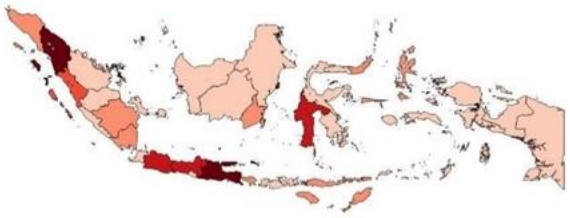


Figure 1. Short –distance elderly migration of Indonesia, 2016

Figure 2. Long-distance elderly migration (In) 2016

Figure 3. Long-distance elderly of Indonesia, migration (Out) of Indonesia,2016

B. Spatial Distribution of Elderly Migrants

For short-distance migration, the migration mobility rate is higher in Java Island, northern Sumatera Island and Southern Sulawesi Island (Figure 1). The highest short-distance elderly migration appeared in East Java Province (0.078 %), and North Sumatera Province (0.072 %), In addition, Central Java Province (0.056 %) also showed high short-distance mobility rates for elderly.

In contrast, Gorontalo province (0.001%), West Papua Province (0.0013%) and Banten Province (0.004%) showed low short-distance elderly migration. Most long-distance elderly migrants came from DKI Jakarta provinces (0.07%) (Figure 3). West Java and Central Java Province were the most frequent destinations for elderly migrants (Figure 2).

Table 1. Multilevel Logistic Regression Coefficients

Independent Variable	Short-Distance Migration	Long-Distance Migration
	Model 1	Model 2
Individual level		
Age (5 years ago/before migration)		
Productive age (< 65)[a]		
Non-productive age (65+)	-0.2352092***	-0.1285305
Sex		
Male[a]		
Female	-0.0574215	-0.1623248
Region		

Rural[a]		
Urban	0.2600165***	0.6492608***
Education Level		
Junior high school and below[a]		
Senior High School	0.26289*	0.1907098
College and above	0.5232416***	0.145435
Working Status		
Don't have a job[a]		
Have a job	-0.501196***	-0.5936791***
Regional Level		
Robbery rate by province[b]	-	0.0833767***
Per capita food expenditure per month by province[b]	-	3.53e-06**
Percentage of elderly family builder by province [b]	-	0.3973345
ICC (null model)	-	**11,94%**

[a]Designates reference category [b]Denotes continuous variable *p < .1 **p<.05 ***p<.0.1

C. Migration Impacts

The results reveal that 11.94% of the variation of elderly migration in Indonesia due to the difference in the inter-Province characteristics. In Model 1 and 2, as expected, living in the urban area was significantly related with short-and long-distance migrations, which indicated people ended up with a better neighborhood in a rural area or in other urban areas with a more comfortable environment. Besides that, as expected, working status was significantly related with short-and long-distance migrations. Elderly people who do not have jobs are more likely to migrate.

In short-distance migration, age, living region, education level and working status of elderly were proven as a strong factor to elderly migration. Model 1, indicate that among elderly, being productive elderly tended to increase the migration (OR:1.3, $p<0.01$) compared with non-productive elderly by reaching age of 65 and above. Living in urban areas tended to increase the migration (OR: 1.3 $p<0.01$) compared with elderly that living in rural areas. Have Senior high school education level tend to increase the migration (OR: 1.3, $p<0.1$) compared with having low education level. Have college and above education level tend to increase the migration (OR: 1.7, $p<0.01$) compared with having low education level. Having a job tend to reduce the migration. (OR: 1.7, $p<0.01$).

In long-distance migration, contextual variables improve the explanatory power of the models. The results indicate lower long-distance elderly migration for elderly living in areas with a low roberry rate (OR: 1.09, $p<0.01$) and for elderly living in areas with low per capita food expenditure per month (OR: 1.000004, $p<0.01$). On individual variable, the results indicate higher long-

distance elderly migration for elderly who live in the urban area (OR: 1.9, p< 0.01), and for elderly with no job (OR: 0.55, p<0.01).

4. Discussion and Conclusion

A. Discussion

Determinant 1: Individual Attributes

Compared with those who did not move, short-distance migrants tended to be younger and productive, living in urban areas, more educated and didn't have a job. While elderly migrants who migrate to another province tended to live in urban areas and didn't have a job.

Because of the barrier of distance and the cost of moving, along with potential losses that may occur, the non-productive age (65 years and over) groups may find it more difficult to migrate because their physical capacity for adapting to change might decrease with age (Dou Xiaolu, Liu Yujun, 2015).

The elderly living in the urban area tend to migrate than the elderly living in rural areas. Having a job was strongly correlated with urban residency among the elderly in this study. This phenomenon could be explained in two ways. First, big cities usually have a more significant regional discrepancy. People may have been restricted in places of work, but after retirement, they were free to move to other districts that had better social services, medical amenities, and/or living conditions. Second, urban residents have a much higher average income than their rural counterparts, which suggests they could afford to move if they desired. (Dou Xiaolu, Liu Yujun, 2015).

Education could be an explanation for migration among the elderly. As the data indicated, the elderly migrants reported have a better education than those who did not move. This occurs because more educated individuals obtain the largest absolute economic gains from migration (Grogger and Hanson, 2011).

The traditional structure of migration for better income is changing. According to the results, there was no evidence suggesting that sex of the elderly would affect elderly migration. In long-distance migration, there was no evidence suggesting that the age of elderly and educational level would affect elderly migration. Elderly may find it more difficult to migrate farther because their physical capacity might decrease with age.

Determinant 2: Regional Attributes

The spatial pattern of the Indonesian elderly migrants exhibited regional distinctions, which reflected different developing stages and their effects on migration. Most elders made short-distance migrations.

For short-distance migration, the highest mobility rate occurred in Java Island. In line whit this, for long-distance migration, In Indonesia, Central Java Province and West Java Province are the most popular destination for elderly

migrants who came from metropolitan area, as it was a relatively developed province containing lots of culture diversities (Dou Xiaolu, Liu Yujun, 2015). In Java Island, there are capital cities of Jakarta and several other metropolitan cities. So that access to food and living necessities is very close. This causes the cost of living on Java Island is relatively cheap. On Java Island, there are also so many mountains, beaches and rural areas that are very comfortable for elderly people to live in.

According to the results, there was evidence suggesting that elderly living in a province with high Robbery rate more likely to migrate. The house we live in and where we live can have a major impact on our physical and mental health; this is particularly true for groups of people who tend to spend more time in the home, including older adults (Centre for Ageing Research and Development in Ireland, 2013).

According to the results, the elderly living in a province with high Per capita food expenditure per month more likely to migrate. Many older people in residential care accommodation are undernourished, either through previous poverty, social isolation, or personal or psychological problems or due to the effects on the appetite of illness or medication (The Caroline Walker Trust, 1995).

B. Conclusion and Recommendation

In Indonesia, there is an increase in the number and percentage of the elderly population. This study highlights the importance of multilevel and non-discriminatory policies between province. Results suggested that a considerable number of Indonesian elderly migrated after retirement for several reasons, including demographic characteristics of the elderly, environmental security, and living costs especially food costs. The migration pattern also reflected social changes that related to reurbanization and economic transformation. Patterns show that elderly migrants in Indonesia usually chose to move where their children are or to the place they live from birth to adolescence so that the environment is familiar and comfortable for them in old age. Elderly migration in Indonesia demands a flexible public policy related to elderly care systems.

In order to address regional disparities in elderly migration in Indonesia, it is important to look beyond individual-level attributes. Contextual characteristics of the community must also be addressed. An aging society needs essential strategies in building appropriate amenities for the increasing older population. Responsive policies and social services are necessary not only to eliminate barriers between provinces but also to assist older migrants to adapt to new environments. Strategies to improve security from criminality and efforts to stabilize food prices in order to create a livable environment in each province also must be supported.

References

1. Adioetomo, S.M. (2013). *Ageing Monograph: Evidence from the 2010 Census*. Jakarta: UNFPA Indonesia
2. Badan Pusat Statistik. (2017). *Statistik Penduduk Lanjut Usia 2016*.Jakarta: Badan Pusat Statistik
3. BKKBN. (2012).*Bina Keluarga Lansia (BKL) Sistem Informasi Keluarga Sejahtera*.Retrieved from http://aplikasi.bkkbn.go.id/bkl/Report/DaftarLaporanBKL.aspx
4. Cuba, L., & Longino, C. F. (1991). Regional retirement migration: The case of Cape Cod. *Journal of Gerontology*, 46, S33-S42.
5. Dou Xiaolu, Liu Yujun. (2015). Elderly Migration in China: Types, Patterns, and Determinants. *Journal of Applied Gerontology* 1-21.
6. Grogger, J. and Hanson, G.H. (2011). Income maximization and the selection and sorting of international migrants. *Journal of Development Economics*, 95(1): 42-57.
7. Hox, Joop J. (2010). *Multilevel Analysis: Techniques and Applications (2nd ed.)*. New York: Routledge.
8. Lu, Z., & Song, S. (2006). Rural urban migration and wage determination: The case of Tianjin, China. *China Economic Review*, 17, 337-345. doi:10.1016/j. chieco.2006.04.007
9. Omelaniuk, I. (2005). Best practices to manage migration: China. *International Migration*, 43, 189 206. doi:10.1111/j.1468-2435.2005.00346.x
10. Tong, Y., & Piotrowski, M. (2012). Migration and health selectivity in the context of internal migration in China, 1997-2009*. Population Research and Policy Review*, 31, 497-543. doi:10.1007/s11113-012-9240-y
11. Warnes, A. M., K. Friedrich, L. Kellaher, S. Torres. (2004). The diversity and welfare of older migrants in Europe. *Ageing and Society*, Vol. 24, No. 3, pp. 307-326.
12. Zaiceva A. (2014). *The impact of Aging on the Scale of Migration-Older people migrate less than young, yet with population aging, mobility of elderly and specialized workers may increase*. IZA World of Labor

Multidimensional poverty in East Nusa Tenggara: A structural equation modelling approach

Astrid Ayu Bestari
Statistics Indonesia, Timor Tengah Selatan Regency, Indonesia

Abstract

The target of Sustainable Development Goals (SDGs) related to poverty is aimed to end poverty in all it forms everywhere. Based on data from the National Social and Economic Survey (SUSENAS) conducted by the Statistics Indonesia (BPS) in March 2018, 21,35 percent of the population in East Nusa Tenggara Province, Indonesia is classified as poor. Poverty is a complex and multidimensional problem because it is associated with limitations in acces to economic, social culture, politics and participation in society. This research aims to analyze mutidimensional poverty in East Nusa Tenggara using Structural Equation Modelling (SEM) approach. Data that used in this research are the result SUSENAS March 2018. SUSENAS is a routine survey conducted by BPS every year. The results show us that Education dimension influenced Standard of Living dimension of poor people and Standard of Living dimension affected the Health dimension of poor people. Therefore, the government needs to make policies related to these three dimensions in an effort to eradicate poverty. In addition, a existing policy needs to be monitored and evaluated so that it will provide more benefits to the poor.

Keywords

SEM; SUSENAS: Indonesia

1. Introduction

Poverty is still the biggest challenge faced by countries in the world, both developed and developing countries, Poverty is a requirements that are indispensable for sustainable development so that it requires great attention from the government in efforts to alleviate poverty in order to improve people's welfare. The target of Sustainable Development Goals (SDGs) related to poverty is aimed to end poverty in all it forms everywhere. In 2030 at least halve the proportion of men, women and children of all ages living in poverty, and implementing a national social protection system that applies to all people, including the poor and vulnerable to poverty.

Based on data from the National Social and Economic Survey (SUSENAS) conducted by the Statistics Indonesia (BPS) in March 2018, 21,35 percent of the population in East Nusa Tenggara Province, Indonesia is classified as poor. The number of poor people in SUSENAS in March 2018 increased by about

7.430 people compared to the poor population in September 2017 SUSENAS. Based on the area of residence, during the period September 2017 - March 2018, the number of poor people in rural areas increased by 4.510 and for urban areas also increased 2.910 people. In measuring poverty, BPS uses the ability concept to meet basic needs approaches. With this approach, poverty is seen as an economic inability to meet basic food and non-food needs measured by expenditure.

Poverty is a complex and multidimensional problem because it is associated with limitations in acces to economic, social culture, politics and participation in society. Poverty often define as a lack of money. Yet poor people themselves consider their experience of poverty much more broadly. A person who is poor can suffer from multiple disadvantages at the same time – for example they may have poor health or malnutrition, a lack of clean water or electricity, poor quality of work or little schooling. Focusing on one factor alone, such as income, is not enough to capture the true reality of poverty. Monitoring monetary deprivations alone can not provide a complete picture of this basic well-being. Someone may not be poor based on monetary standards but can still feel the effects of poverty if they lack access to basic needs such as health care, clean water, and education (World Bank, 2018). The main advantage of the multidimensional approach in measuring poverty is that it is possible to obtain a picture of poverty distribution based on a combination of existing dimensions, such as health and education dimensions, so that groups can be assisted based on two dimensions simultaneously, which can differ from one dimension, for example only health, or just education (Asra, 2017).

This research aims to analyze mutidimensional poverty in East Nusa Tenggara using Structural Equation Modelling (SEM) approach. Voth-Gaedert and Oerther (2014) state that The SEM Approach is best served when analyzing a multidimensional or multivariate problem like poverty. SEM allows for the incorporation and understanding of multiple relationships within a complicated reality. The use of latent variables is a concept within the area of SEM that allows the researcher to represent variables that can prove difficult to analyze through basic observations. Instead of using the idea of an index of indicators, SEM is able to avoid the errors accumulated from the summation of variables, whether weighted or not.

2. Methodology

Data used in this research are raw data of SUSENAS March 2018. SUSENAS is a routine survey conducted by BPS every year that aims to collect data about socio-economic conditions of people such as: health condition, education, fertility, family planning, employment, housing and other socio-economic condition. Since 2015, SUSENAS is held twice in a year, namely March and

September. SUSENAS in March uses 2 kind of questionnaires which are kor questionnaire and consumption expenditure (KP) questionnaire with total sample of 300.000 household while total sample of SUSENAS in September are 75.000 households.

To analyze the data from SUSENAS, The SEM approach were used. Voth-Gaeddert and Oerther (2014) state that the general SEM analysis is typically a two-step approach when working towards a confirmatory model. First define the model. There are two parts to the full model, a measurement model and a structural model. The measurement model describes the relationships between the latent variables and the observable indicator variables. Using CFA, the hypothesized model (covariance matrix) is compared to a data driven model (covariance matrix) using the Chi-Square test of model fit. If the measurement model does fit the data via several tests of model fit indices, then the structural model can be assessed. Using the same style of model fit a model is either accepted or rejected. If the model is rejected adjustments can be made (with caution) and then retested. Once the model is accepted, direct and indirect effects can be assessed between latent and independent variables. This allows for the analysis of relationships between factors such as socio economic status, education, health, etc. SEM can also be used in a purely exploratory style of analysis using EFA. If the researcher is uncertain as to which factor is described by which indicator, EFA allows for freedom amongst relationships within the measurement model. It is highly recommended within the literature that once a model is established through EFA, CFA is used with new data to test the model.

The SEM model will be made according to the 2018 Global Multidimensional Poverty Index (MPI) with a few modification to adjust with the condition in East Nusa Tenggara. There are 3 dimensions of poverty that is Health, Education, and Standard of Living. Each dimension consists of indicators which observed from people which is below the poverty line of East Nusa Tenggara. The indicators that will be used can be seen in table 1.

Table 1. Indicators and Dimensions of Multidimensional Poverty

Dimension	Indicator		ToiletType
Health	Calories	Standard of Living	Electricity
	Protein		Asset
	Fat		Roof
Education	Years of Schooling		Wall
	Diploma		Floor
Standard of Living	Water		Cooking Fuel
	TimeTakeWater		Floor per Capita
	Toilet		

3. Results

The Goodness-of-Fit statistic (GFI) was created by Jöreskog and Sorbom as an alternative to the Chi-Square test and calculates the proportion of variance that is accounted for by the estimated population covariance (Tabachnick and Fidell, 2007 as read in Hooper, Coughlan and Mullen). Goodness of Fit (GFI) includes as Absolute fit indices that are often used as a reference for assessing model fit. GFI is an indices of the accuracy of the model in explaining the model that is compiled. To determine the model fit based on GFI, the GFI value is expected to be ≥ 0.90. The GFI value ranges from 0.00 (poor fit) to 1.00 (perfect fit). From table 2, the GFI value of 0.937 is obtained, indicating that the model is fit. Adjusted Goodness of Fit (AGFI) is a fit criteria for the development indices of GFI which is adjusted to the degree of freedom ratio for the proposed model with degree of freedom for the null model. The recommended AGFI value for indication of model fit is ≥ 0.90. From table 2, an AGFI value of 0.917 is obtained, indicating that the model is fit.

Graphic 1. SEM Model

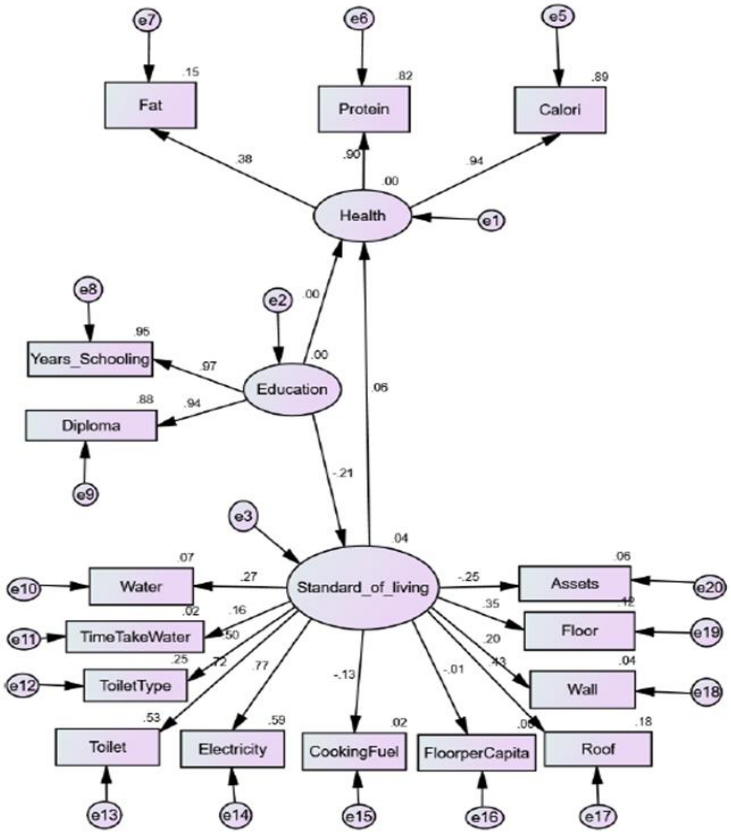


Table 2. GFI and AGFI

Model	GFI	AGFI
Default model	.937	.917
Saturated model	1.000	
Independence model	.861	.843

Table 3 shows the accuracy of the relationship between indicators with dimensions or latent variables. If the Critical Ratio (CR)> 1.96 or P <0.05, the variable relationship can be concluded as correct. From Table 3, it can be seen that the relationship between the Standard of Living with Floor per Capita has a value of P> 0.05 so it can be concluded that this relationship is not appropriate. While the relationship of other indicators with respective dimensions or latent variables are correct.

Table 3. Regression Weights of Indicators

			Estimate	S.E.	C.R.	P
Health	→	Fat	.011	.000	29.412	***
Health	→	Protein	.026	.001	49.688	***
Health	→	Calori	1.000			
Education	→	Years_Schooling	1.000			
Standard_of_living	→	Electricity	1.000			
Standard_of_living	→	CookingFuel	-.034	.004	-7.902	***
Standard_of_living	→	FloorperCapita	-.028	.057	-.497	.619
Education	→	Diploma	1.109	.029	37.661	***
Standard_of_living	→	Assets	-.144	.006	-22.582	***
Standard_of_living	→	Water	.446	.019	23.317	***
Standard_of_living	→	Toilet	1.000			
Standard_of_living	→	TimeTakeWater	34.056	3.013	11.303	***
Standard_of_living	→	Floor	.362	.015	24.780	***
Standard_of_living	→	Wall	.472	.026	17.925	***
Standard_of_living	→	Roof	.375	.013	27.955	***
Standard_of_living	→	ToiletType	-.490	.011	-42.804	***

The estimate number in table 4 shows the factor loadings of each indicator with respect to dimensions or latent variables related. The value of the factor loadings shows how strong the relationship is. Education variable consist of 2 indicators, namely the number of years of schooling and the last diploma completed, and these two indicators have a strong relationship and can explain the variable education. Health variable consist of 3 indicators, and

there are 2 factor loadings, namely protein and calories consumed last month, which have a value bigger than 0.5 so that they have a strong relationship with Health variable and can be used to explain Health variable. The Standard of Living variable consists of 11 indicators and only 2 indicators that have a strong relationship with the Standard of Living Variable namely Electricity and the existence of a toilet. While other indicators are cooking fuel, assets owned, type of water used, time to collect water, type of floor, type of wall, type of roof and type of toilet used. they have a weak relationship with the Standard of Living variable with a factor loadings value less than 0, 5. While the floor area per capita does not have a relationship with the variable Standard of Living.

Table 4. Standardized Regression Weights

			Estimate				Estimate
Health	→	Fat	.383	Standard_of_living	→	Assets	-.255
Health	→	Protein	.904	Standard_of_living	→	Water	.269
Health	→	Calori	.941	Standard_of_living	→	Toilet	.725
Education	→	Years_ Schooling	.972	Standard_of_living	→	TimeTake Water	.155
Standard_of_living	→	Electricity	.765	Standard_of_living	→	Floor	.346
Standard_of_living	→	Cooking	-.132	Standard_of_living	→	Wall	.199
			Estimate				Estimate
		Fuel					
Standard_of_living		Floorper Capita	-.010	Standard_of_living		Roof	.427
Education		Diploma	.939	Standard_of_living		ToiletType	-.498

Table 5. Estimation Result

			Factor Loading	C.R.	P
Education	→	Standard_of_living	-.210	-18.346	***
Standard_of_living	→	Health	.056	4.135	***
Education	→	Health	.001	.058	.954

Table 5 shows the results of the test of causality relationships between dimensions or latent variables. The first relationship shows the influence of Education on Standard of Living from poor people. The results of the analysis using the Amos application showed the effect of Education on Standard of Living obtained a CR value of -18,346 and p <0.05. Thus this first relationship is appropriate and acceptable in the model. The Factor Loading value of -0.210 shows a weak relationship (<0.5). The value of the negative loading factor

indicates an opposite direction relationship. This means that the higher the education of the poor, the lower the standard of living. The second relationship is the influence of Standard of Living on the health of the poor. The results of the analysis show that the influence of Standard of Living on Health obtained a CR value of 4.135 and p <0.05. Thus this second relationship is appropriate and acceptable in the model. Factor Loading value of 0.056 indicates a weak relationship (<0.5). The value of a positive factor loading shows a unidirectional relationship. This means that the higher the Standard of Living of the poor, the better their Health. While the third result of the relationship, that is the effect of Education on Health, shows the value of CR = 0.058 and p = 0.954. These result indicates that Education has no influence on the health of poor people.

4. Discussion and Conclusion

Improving education is still a challenge that must be faced by the East Nusa Tenggara government. The policy that can be carried out by the regional government is to increase infrastructure, especially in remote villages such as build schools in areas where schools are not available, increase the number of teachers, improve the quality of teachers with training for various levels of education, add reading books to school libraries and provide opportunities for people to continue their education to a higher level outside the village. Improving education is not only aimed at school-age people, but also for those who are not in school age and those who do not attend school or those whose education is low. The policy that can be done is to conduct training to increase skills and promote a program to continue education for those who drop out of school. Improving education certainly will directly relate to the job of the poor people. The higher the education/skills, the better the job they gained. Thus the welfare of the poor people will increase because of the increase in income earned.

Governments also need to make efforts to improve the standards of living of the poor. Investment promotion program for village development (improving the quality of water and sanitation) is an example to improve the quality of life of the villagers. Since 2015, the government has implemented the Village Fund program, which is the provision of funds sourced from the State's income and expenditure budget for traditional villages and villages that are transferred through the district/city regional income and expenditure budget. Village funds aim to fulfill basic needs of the community, develop village infrastructure, develop local economic potential and utilize natural and environmental resources. This program is expected to improve the welfare of the villagers, therefore monitoring from the government is needed so that this program will continue in the future.

Health is also a problem that needs full attention from the government. The high number of children suffering from malnutrition in villages in the province of East Nusa Tenggara is still a scourge for the government. The government has conducted various programs in this health dimension such as the provision of Beras Miskin/Raskin (subsidized rice) and BPJS Kesehatan PBI (insurance wihout premium) for the poor. However, there are still many problems faced by this program, such as the Raskin program which is still not right on target, villagers who have not been able to utilize BPJS Kesehatan PBI because of poor people's ignorance of how to use it and because of the closest facilities that receive BPJS Kesehatan PBI is far. For this reason, the government is expected to further review the programs that are being carried out so that the poor have more benefits.

References

1. Asra, Abuzar.(2017). Esensi Statistik Bagi Kebijakan Publik. Bogor-In Media
2. Badan Pusat Statistik. (2017). Indikator Tujuan Pembangunan Berkelanjutan Indonesia. Jakarta-BPS
3. Hair Jr., J.F., Black, W.C., Babin, B.J., Anderson, R.E.(2014). Multivariate Data Analysis.USA:Pearson.
4. Hooper, D., Coughlan, J. And Mullen, M.R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. Electronic Journal of Business Research Methods Volume 6 Issue 1
5. https://ophi.org.uk/research/multidimensional-poverty/ (Accessed November 22, 2018)
6. http://www.worldbank.org/en/news/immersive-story/2018/10/17/going-above-and-beyond-to-end-poverty-new-ways-of-measuring-poverty-shed-new-light-on-the-challenges-ahead (Accessed November 22, 2018)
7. https://www.concernusa.org/story/top-9-causes-global-poverty/ (Accessed November 22, 2018)
8. http://Multidimensionalpoverty.org/chapter-3/ (Accessed November 22, 2018)
9. http://hdr.undp.org/en/2018-MPI (Accessed November 26, 2018)
10. https://www.semestapsikometrika.com/2017/11/confirmatory-factor-analysis-cfa dengan_8.html?m=1 (Accessed December 16, 2018)
11. http://www.jonathansarwono.info/sem/sem.htm (Accessed December 16, 2018)
12. http://widhiarso.staff.ugm.ac.id/wp/distribusi-data-tidak-normal-pada-sem/ (Accessed December 16, 2018)
13. http://theconversation.com/gizi-buruk-pada-balita-di-ntt-mengapa-sulit-diakhiri-91841 (Accessed December 31, 2018)

14. Nurwati, Nunung.(2007). Kemiskinan : Model Pengukuran, Permasalahan dan Alternatif Kebijakan. Jurnal Kependudukan Padjadjaran, Vol. 10, No.1.
15. Sukati, C.W.S. Reducing Poverty: Education Planning And Policy Implications For Swaziland. Educational Planning Vol. 19 No. 2
16. UNDP.(2006). Poverty In Focus. Brazil- United Nations Development Programme
17. Voth-Gaeddert, L. E. and Oerther D. B. (2014). Utilizing Structural Equation Modeling in the Development of a Standardized Intervention Assessment Tool. Procedia Engineering, vol. 78, pp. 218-223. Elsevier Ltd.
18. Voth-Gaeddert L.E., Divelbiss D. W. and Oerther D. B. (2015). Utilizing structural equation modeling to correlate biosand filter performance and occurrence of diarrhea in the village of Enseado do Aritapera in Para, Brazil. Article in Water Science & Technology Water Supply.

Propensity score based adjustment for covariate effects on classification accuracy of biomarker using ROC curve

M. Shafiqur Rahman[1 2], Muntaha Mushfiquee*
Institute of Statistical Research and Training, University of Dhaka, Bangladesh

Abstract

The performance of a biomarker in disease-classification are generally assessed using receiver operating characteristic (ROC) curve. The actual performance may be affected by patient-level covariates and hence it requires for some adjustments. Although a number of methods have been proposed for covariate-adjusted ROC curve, all of them are able to adjust for the effect of single covariate at a time. However, availability of several potential covariates is common in practice for which adjustment is required. This study proposed a propensity score based adjustment for the covariate effects on the ROC curve, which allow for the effects of several covariates. Here propensity score, estimated from a linear combination of several covariates, is used as an alternative of several covariates in an existing non-parametric ROC method. We considered non-parametric ROC regression (Rodrguez-lvarez et al., 2011) to avoid the misspecification of the correct distributional assumption of the continuous biomarker. We described the PS based adjustment for both continuous and binary biomarker allowing for the effect of covariates. The simulation study suggest that the propensity score based adjustment perform well by producing negligible bias and MSE. Further, the method was applied to evaluate BMI to classify hypertensive and diabetic patients, adjusting for the potential covariates such as age, sex, education, socio-economic status etc.

Keywords

Generalized propensity score, non-parametric ROC regression, DHS data

1. Introduction

Given the importance of biomarker in disease classification, it is essential to evaluate their classification accuracy- the ability to provide the correct diagnosis given a subject's true disease status (discrimination). The receiver operating characteristic (ROC) curve, a graphical plot of sensitivity (true positive rate) against the 1-specificity (false positive rate) evaluated at different cutoff values of the diagnostic marker, is commonly used to evaluate the binary classification accuracy of the marker (Pepe, 2000). The potential performance of a biomarker distinguishing diseased from healthy people may

[1] Institute of Statistical Research and Training, University of Dhaka, Bangladesh
[2] Presenting author's email: shafiq@isrt.ac.bd

be strongly inuenced by the both patients and disease characteristics or features of the specimen collection or test settings. As discussed by Janes and Pepe (2008) covariate can be associated withboth biomarker and the disease outcome (acted as confounder or effect modifier), which can affect the classification accuracy. Hence an adjustment for the covariate effect in the ROC curve is necessary, otherwise the traditional pooled ROC curve may provide misleading conclusion on the performance of the marker (Janes and Pepe, 2009). To assess possible covariate effects on ROC curve, a considerable number of approaches have been discussed in literature including non-parametric (Rodrguez-lvarez et al., 2011) and semi-parametric Alonzoa and Pepe (2002); Heagerty (1999) covariate adjustment in ROC analysis. Of them the semiparametric approach is computationally more complex, especially the estimation procedure of the distribution function, in the presence of multiple covariates. The non-parametric approach is solely being able to address only a single continuous covariate. However, in practice a mixture of several continuous and categorical covariates is commonly available, for which methodological extension is required. The issues of covariate adjustment in ROC analysis is similar to the adjustment for the effects of confounders in estimating the marginal treatment effect, which is performed by using propensity score analysis (Austin, 2011) . With this motivation, this paper has provided propensity score based adjustment for the covariate effects in ROC analysis, where a set of both categorical and continuous covariates can be adjusted simultaneously DF et al. (2013). This can be performed by calculating PS from the relationship between the biomarker and the set of covariates through a multivariable regression model and using it as a single covariate, instead of all covariates, in the ROC analysis. The PS based adjustment has been implemented through the extisting non-parametric regression approach, which provides consistent estimates of the ROC curve and its area (AUC). This paper is organized as follows. Section 2 describe separately non-parametric method of covariate adjustment and the proposed PS based adjustment. The simulation study is discussed in Section 3 with necessary results. The methods were illustrated using hypertension data. Section 5 ends the paper with some discussions of main findings and concluding remarks.

2. Methodology

ROC curve

In a theoretical ROC curve, the true positive rate (sensitivity) is plotted in function of the false positive rate (1-specificity) for different cut-off points Coocicnough et al. (2003). Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC

curve that passes through the upper left corner (100% sensitivity, 100% specificity) Goncalves et al. (2014). Therefore, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test Rodrguez-lvarez et al. (2011).

Covariate-adjusted ROC curve for single covariate

Let us consider a continuous biomarker Y and a continuous covariate X. Let X_{D1} and X_{D0} denote the case and control covariate with cumulative distribution function F_{XD1} and F_{XD0} Then using the similar notations in (Janes and Pepe, 2009), the TPR and FPR for Y conditional on X are defined such that,

$$TPR = 1\text{-}F_{D_1(X)}(c) = P\left[Y_{D_1} > c|X\right],$$
$$FPR = 1\text{-}F_{D_0(X)}(c) = P\left[Y_{D_0} > c|X\right],$$

where $f_{D1(x)}$ and $f_{D0(x)}$ are their corresponding density function and p is the collection of all FPR. Then the covariate specific ROC curve can be defined as

$$ROC_X(\mathrm{p})=1-\mathrm{F}_{D_1(X)}\left[\mathrm{F}^{-1}_{D_0(X)}(1-p)\right]$$

The covariate adjusted ROC curve can be defined such that,

$$\alpha ROC(p) = \int ROC_X(p)\, dF_{X_{D1}}(X),$$

Nonparametric estimation of caviariate adjusted ROC curve

Following (Rodrguez-lvarez et al., 2011) a non-parametric regression model is assumed to test biomarker values along with a continuous covariate X for diseased and and non-disesed population:

$$Y_{D_1}=\mu_1(X)+\sigma_1(X)\epsilon_1, \quad (1)$$
$$Y_{D_0}=\mu_0(X)+\sigma_0(X)\epsilon_0, \quad (2)$$

where X is a continuous covariate, μ_1 and , μ_0 are the regression functions, and σ1 and σ0 are the variance functions. The errors 1 and 0 are assumed to be independent of the covariate X, with zero mean, unit variance. Based on this induced regression, the covariate specific ROC curve has the following form:

$$ROC_{X=x}(p) = F_{D1}\left(\frac{\mu_0(X)-\mu_1(X)}{\sigma_1(X)}+\frac{\sigma_0(X)}{\sigma_1(X)}F^{-1}_{D0}(P)\right).$$

The above ROC curve can be estimated using the following procedure. Let $\left\{(x_i^{D1}, y_i^{D1})\right\}_{i=1}^{n\,D1}$ and $\left\{(x_j^{D0}, y_j^{D0})\right\}_{j=1}^{n\,D0}$ be two independent random samples drawn from the diseased and healthy populations, respectively. Then $\widehat{\mu_{1(X)}}$ and $\widehat{\mu_{0(X)}}$ can be estimated as,

$$\widehat{\mu_{1(X)}} = \widehat{\Psi}(x,\left\{(x_i^{D1}, y_i^{D1})\right\}_{i=1}^{n\,D1}, h_1, p_1),$$
$$\widehat{\mu_{0(X)}} = \widehat{\Psi}(x,\left\{(x_j^{D0}, y_j^{D0})\right\}_{j=1}^{n\,D0}, h_0, p_0),$$

where Ψ is the local polynomial kernel estimator, h_1 and $h_0$are the bandwidth, p_1 and p_0 are the order of the polynomial. The corresponding variance function can be estimated as

$$\widehat{\sigma_{1(x)}} = \widehat{\Psi}(x,\{(x_i^{D1}, z_i^{D1})\}_{i=1}^{n\,D1}, g_1, n_1),$$
$$\widehat{\sigma_{0(x)}} = \widehat{\Psi}(x,\{(x_j^{D0}, z_j^{D0})\}_{j=1}^{n\,D0}, g_0, n_0),$$

where, $z_i^{D1} = (y_i^{D1} - \widehat{\mu_1}(xi^{D1}))^2$ and $z_j^{D0} = (y_j^{D0} - \widehat{\mu_0}(xj^{Do}))^2$ and Ψ is the local polynomial kernel estimator, g_1 and g_0 are the bandwidth, n_1 and n_0 are the order of the polynomial (Lez-Manteiga, 2011).

Now using the above estimators the TPR and FPR can be estimated as

$$\hat{F}_{D1}(z) = \frac{1}{n_{D1}} \sum_{i=1}^{n_{D1}} I\left[\left(\frac{y_i^{D1} - \hat{\mu}_1(X_i^{D1})}{\hat{\sigma}_1(X_i^{D1})}\right) \geq z\right],$$
$$\hat{F}_{D0}(z) = \frac{1}{n_{D0}} \sum_{j=1}^{n_{D0}} I\left[\left(\frac{y_j^{D0} - \hat{\mu}_0(X_j^{D0})}{\hat{\sigma}_0(X_j^{D0})}\right) \geq z\right].$$

Finally, the estimated covariate specific ROC curve can be obtained as

$$\widehat{ROC_{X=x}}(p) = \hat{F}_{D1}\left(\frac{\hat{\mu}_0(X) - \hat{\mu}_1(X)}{\hat{\sigma}_1(X)} + \frac{\hat{\sigma}_0(X)}{\hat{\sigma}_1(X)} \hat{F}_{D0}^{-1}(p)\right) \quad (4)$$

where, $\hat{F}_{D1}^{-1}(p) = Sup\{z; \hat{F}_{D1}(z) \geq t\}$. Then the estimated covariate adjusted ROC can obtained as,

$$a\widehat{ROC}(x) = \int ROC_{X_i}(p)\, dF_{X_{D1}}(X) = \frac{1}{n_{D1}} \sum_{i=1}^{n_{D1}} I\left[\left(\frac{y_i^{D1} - \hat{\mu}_1(X_i^{D1})}{\hat{\sigma}_1(X_i^{D1})} > F_{D0X}^{-1}(p)\right)\right]$$

Propensity score based covariate adjustment in ROC analysis

Here we discuss how propensity score can be used to estimate the covariate adjusted ROC curve in the presence of several covariates. First we discuss an overview of propensity score and then how can it be implemented to estimate covariate adjusted ROC curve.

Propensity score

Propensity score method is the most popular strategy for estimating marginal effect of the treatment/exposure reducing the effect of confounding in observational studies Austin (2011). The propensity score analysis is useful to address selection bias when random assignment is not feasible. Propensity score are used in such a way that the resulting case and control groups will have similar covariate values to those created through random assignment Olmos and Govindasamy (2015). The propensity score is a probability of being treated or exposed. For a binary treatment or exposure Y and a set of confounders X_1,…,X_(n)the propensity score can be estimated using logistic regression:

$$PS = P(\mathrm{Y} = 1|\mathrm{X}) = [1 + \exp(-\beta \mathrm{X})]^{-1}. \quad (5)$$

For the continuous treatment or exposure, generalized propensity score (GPS) (DF et al., 2013) are frequently used. Let $r(y,x)$ be the conditional density of the exposure/teatment given the covariates (Zhu et al., 2015), which has the following form:

$$r(y,x) = f_{Y|X}(y|x) = N(\beta X, \sigma^2).$$

Then the general propensity score can be obtained as

$$gPS = r(Y,X) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\Big[-\frac{1}{2\hat{\sigma}^2}(Y_i - \hat{\boldsymbol{\beta}}\boldsymbol{X})^2\Big].$$

The PS are result of a linear combination of several covariates. There are a number appraoches available in PS literature including matching, stratification, weighting and PS as covariate. Of them PS as covariate provide better performance in estimating marginal effect of the tratement adjusting for the effetc of several confounders.

Propensity score based covariate adjustment

The adjustment for several covariates in the ROC analysis can be achieved by replacing the single covariate X by the estimated propensity score into the non-parametric requression equations (eq:1-2). That is, the non-parametric requression equations (eq:1-2) will then take the following form:

$$Y_{D_1} = \mu_1(gPS) + \sigma_1(gPS)\epsilon_1, \quad (7)$$

$$Y_{D_0} = \mu_0(gPS) + \sigma_0(gPS)\epsilon_0. \quad (8)$$

If the biomarker Y is binary then gPS will be replaced by PS. Then PS based covariate-adjusted ROC curve ($\alpha\widehat{\mathrm{ROC}}$ (PS)) for both dichotomous and continuous biomarker can be estimated using the same non-parametric procedure described earlier, but replacing the single covariate X by PS. Alternatively, one can use the linear coefficient $\hat{\beta}X$ of the PS model instead of PS, which does not change any results in the ROC analysis because there is one-to-one relationship between PS and $\hat{\beta}X$. Using PS or$\hat{\beta}X$ as a covariate in the non-parametric regression confirm the adjustment of several covariates because the PS or$\widehat{\beta}X$ serves as a representative of the linear combination of several covariates. The corresponding PS adjusted $\alpha\widehat{\mathrm{ROC}}$(PS) curve has similar interprtation to the $\alpha\widehat{\mathrm{ROC}}(x)$ but adjusting for the effect of *PS* rather than *X*, i.e., adjusting for the effect of all covariates.

3. A simulation study

In this section, we present a simulation study to evaluate the performance of the propensity score based covariate adjustment in the ROC analysis. Separate simulation series were conducted for continuous and binary biomarker. In the both simulation series, we first generated two independent

covariates of which one is continuous and the other is dichotomuos, separately for diseased (D_1) and healthy (D_0) population as

$$X_{1D1} \sim N(\mu_{D1}, \sigma^2_{D1}); \quad X_{1D0} \sim N(\mu_{D0}, \sigma^2_{D0});$$
$$X_{2D1} \sim Bin(n, \pi_{D1}); \quad X_{2D0} \sim Bin(n, \pi_{D0});$$

We considered $\mu_{D1} = 51.54$, $\mu_{D0} = 38.19$, $\sigma^2_{D_1} = 224.13$, , $\sigma^2_{D_0} = 182.33$, $\pi_{D1} = 0.67$, $\pi_{D0} = 0.42$. Different values of the parameters associated with the covariate distribution for disease and healthy population suggest imbalance covariate distribution between disease and haelthy population, for which adjustment for covariate effect is necessary. Then the continuous biomarker values for diseased and non-diseased population were gererated from the following models:

$$Y_{D_1} = \beta_{01} + \beta_{11} X_{1D_1} + \beta_{21} X_{2D_1} + \epsilon_{D_1},$$
$$Y_{D_0} = \beta_{00} + \beta_{10} X_{1D_0} + \beta_{20} X_{2D_0} + \epsilon_{D_0},$$

where $\epsilon_{D_1} \sim N(0,1)$ and $\epsilon_{D_0} \sim N(0,1)$. The values of the parameters were set to $\beta_{01} = 29.99$, $\beta_{11} = -0.013$, $\beta_{21} = 2.06$, $\beta_{00} = 21.50$, $\beta_{10} = 0.121$, $\beta_{20} = -0.77$. Different values of the parameters of the model for diseased and healthy subjects suggest an association of covariates X_1 and X_2 with both the biomarker and the disease. Similarly the binary biomarker values were generated separately for diseased and healthy population as $Y^b_{D_1} \sim Bin\left(\pi_{D_1}(x)\right)$ and $Y^b_{D_0} \sim Bin\left(\pi_{D_0}(x)\right)$ where $\pi_{D1}(x) = Pr[Y^b = 1|X, D = 1]$ and $\pi_{D0}(x) = Pr[Y^b = 1|X, D = 0]$ obtained from the following logit model

$$\text{logit}\left[\pi_{D_1}(x)\right] = \gamma_{01} + \gamma_{11} X_{1D_1} + \gamma_{21} X_{2D_1},$$
$$\text{logit}\left[\pi_{D_0}(x)\right] = \gamma_{00} + \gamma_{10} X_{1D_0} + \gamma_{20} X_{2D_0},$$

We considered $\gamma_{01} = -1.5$, $\gamma_{11} = 0.123$, , $\gamma_{21} = 0.772$, $\gamma_{00} = -2.46$, , $\gamma_{10} = -0.013$, , $\gamma_{20} = 2.06$. For both simulation series, we considered different scenarios by varying the sample size as 150, 300, 500, 1000, with 30% for the subjects with disease and the rest of them for healthy subjects. This implies that $P(D) = 0.30$. Further simulation scenario were created by considering X_1 and X_2 correlated. For each simulation scenario, the results were summarized over 500 replicated datasets. Figure 1 shows that the whole sampling distribution of the PS adjusted ROC curves for continous biomarker over 500 replications and their mean ROC curve with superimposed true ROC curve. The results revelated that bias in the proposensity score adjusted ROC curve is negligible and the amount of bias tends to zero with the increasing sample size. The standard error (or mean squared error) decreased with increasing sample size. Similar results were observed for the binary biomarker (Figure 2).

Further simulations with correlated covariates also suggest similar results (results not shown).

4. Application

The method is applied to evaluate body mass index (BMI) to classifying the subjects with hypertension (and diabetes separately) from the healthy subjects with age over 35 years in the presence of several covariates such as age, gender, education, socio-economic status etc. Data used to illustrate the method were extracted from the 2011 Bangladesh Demographic and Health Survey (BDHS) that has been conducted through a collaborative effort of the National Institute of Population Research and Training (NIPORT), Mitra and Associates and ICF international. The biomarker measurements for both blood pressure and fasting glucose level were collected from all members of a household with age 35 years and above in every third of the selected households. A subject is classified as hypertensive if his/her systolic blood pressure (SBP) value equal to or greater than 140 mmHg or a diastolic blood pressure (DBP) value equal to or greater than 90 mmHg or he/she was taking medication for lowering the blood pressure. A subject with fasting plasma glucose level equal to higher than 7.0 mmol/L or who was taking medication is classified as a patient with diabetes. Finally, biomarker information related to both the hypertension and diabetes were collected from a total of 7328 (male 3744, female 3584) and 7593 (male 3753, female 3840) subjects, respectively who given their consent for providing blood sample. A set of individual level covariates available including age (years), sex (male, female), education (no education, primary, secondary, higher), area of residence (urban, rural), and household's socio-economic status (poorest, poorer, richer, richest). The status of both hypertension (yes, no) and diabetes (yes, no) were considered as two outcome variables and BMI (a composite index of height and weight) as biomarker. However, status of diabetes was considered as covariate when we evaluated the performance of BMI in classifying hypertensive patient and vice-versa. From the summary measures (results not shown), it can be observed that 25% and 12% of the total subjects are found to be hypertensive and diabetic, respectively. The average age of all subjects is 51.05 with almost equal number of male and female. Of the total sample, 1094 (14.97%) subjects have BMI over 18.5, 32.87% belong to urban area and 67.13% to rural area and majority have no education or primary education (approximately 44.94% and 27.71%).

Performance of BMI to classify patients with diabetes and hypertension

Separate analysis was performed to evaluate the performance of the BMI in classifying diabetes and hypertension, after adjusting for all baseline patient-specific covariates such as age, gender, area of residence and socio-

economic status. Literature and some explanatory analyses (results not shown) suggest possible association of all such covariate with both the BMI and diseases. Although the classification performance (discrimination) of BMI does not affect by the any covariates, the crude (unadjusted) performance was observed to be different from the covariate specific performance. We then evaluated the performance of BMI adjusting for each covariate and all covariates simultenously using propensity score approach. Figure 3 presents the results for PS based covariate adjustment for the different combinations of several covariates. The results revealed that there is some differences in the performance of BMI in classifying either hypertensive or diabetic patients when adjustment made for different combinations of the covariates. The greatest classification accuracy for both hypertension and diabetes was observed when all possible covariates were adjusted simultenously. However, the performance of BMI is appeared to similar for both hypertension and diabetes.

5. Discussion and Conclusion

The importance of new biomarker in biomedical research depends on how the accurately the marker classify the subjects with disease from the healthy subjects. The ROC curve is a popular method for evaluating biomarker's classification accuracy. However, the difference in patient characteristics (baseline covariates) between disease and healthy population may make it difficult to evaluate classification acucracy of biomarker and hence it is necessary to adjust for the covariate effect in the ROC analysis Huang and Pepe (2011b). Existing methods of covariate adjustment in ROC curve allow to adjust for single covariate (Rodrguez-lvarez et al., 2011). However, the adjustment of several covariates is often require in practice. This study provided a solution using propensity score to adjust for several covariates simultaneously. More specifically, the propensity score, a representative of all covariates, were used in the existing non-parametric estimator for the ROC curve that used to adjust for single covariate (Lez-Manteiga, 2011). The simulation study suggest that the PS based ROC curve appeared to be consistent Huang and Pepe (2011a). Furthermore, the method was applied to evaluate the performance of BMI in classifying the hypertensive and diabetic patient from their healthier counter parts, in the presence of several patient specific covariates such as age, gender, education, socio-economic status, area of residence. The results showed meaningful interpretation.

References

1. Alonzoa, T. A. and M. S. Pepe (2002). Distribution-free roc analysis using binary regression techniques. Biostatistics 3 (3), 421{32.

2. Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research 46 (3), 399424.
3. Coocicnough, D., R. K, and L. LB. (2003). Radiographic applications of receiver operatins characteritic (roc) curves. Radiology 229 (1), 3{8.
4. DF, M., G. BA, A. D, S. ME, R. R, and B. LF. (2013). A tutorial on propensity score estimation for mutiple treatments using generalized boosted model. Statistics in Medicine 30-32 (19), 3388{414.
5. Goncalves, L., A. Subtil, and other (2014). Roc curve estimation: An overview. Revstat Statistical Journal 12 (1), 1{20.
6. Heagerty, P. J. (1999). Semiparametric estimation of regression euantiles with application to standardizing weight for height and age in us children. Journal of Royal Statistical Society, Serices C 48 (4), 533551
7. Huang, Y. and M. S. Pepe (2011a). Assessing risk prediction models in case-control studies using semiparametric and nonparametric methods. Statistic in Medicine 29 (13), 13911410.
8. Huang, Y. and M. S. Pepe (2011b). Evaluating the predictiveness of a continuous marker. Biometrics 63 (4), 11811188.
9. Janes, H. and M. S. Pepe (2008). Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. American Journal of Epidemiology 168 (1), 89{97
10. Janes, H. and M. S. Pepe (2009). Adjusting for covariate e_ects on classi_cation accuracy using the covariate-adjusted receiver operating characteristic curve. Biometrika 96 (2), 371{382
11. Lez-Manteiga, W. G. (2011). Roc curves in non-parametric location-scale regression models. Scandinavian Journal of Statistics 38, 169{184.
12. Olmos, A. and P. Govindasamy (2015). A practical guide for using propensity score weighting in r. Practical Assessment Research and Evaluation 20 (13).
13. Pepe, M. S. (2000). An interpretation for the roc curve and inference using glm procedures. Biometric 56, 352{359.
14. Rodrguez-lvarez, M. X., J. Roca-Pardias, and C. Cadarso-Surez (2011). Roc curve and covariates: extending induced methodology to the non-parametric framework. Statistics and Computing 21 (4), 483499
15. Zhu, Y., D. L. Co_man, and D. Ghosh (2015). A boosting algorithm for estimating generalized propensity score with continuous treatment. J. Causal Inference 3 (1), 25{40.

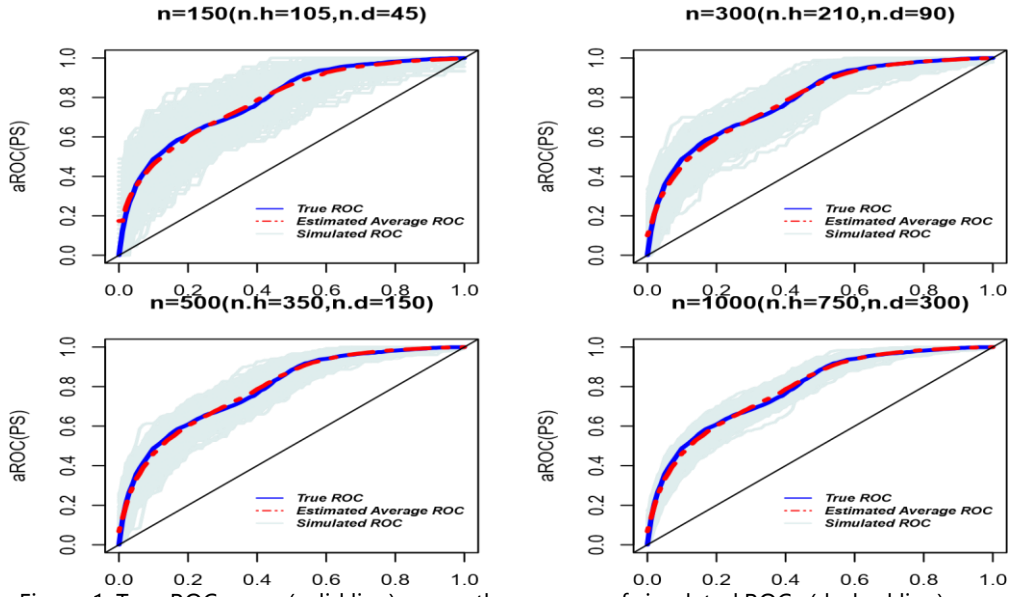


Figure 1: True ROC curve (solid line) versus the average of simulated ROCs (dashed line), along with 500 simulated ROC curves, for continuous biomarker

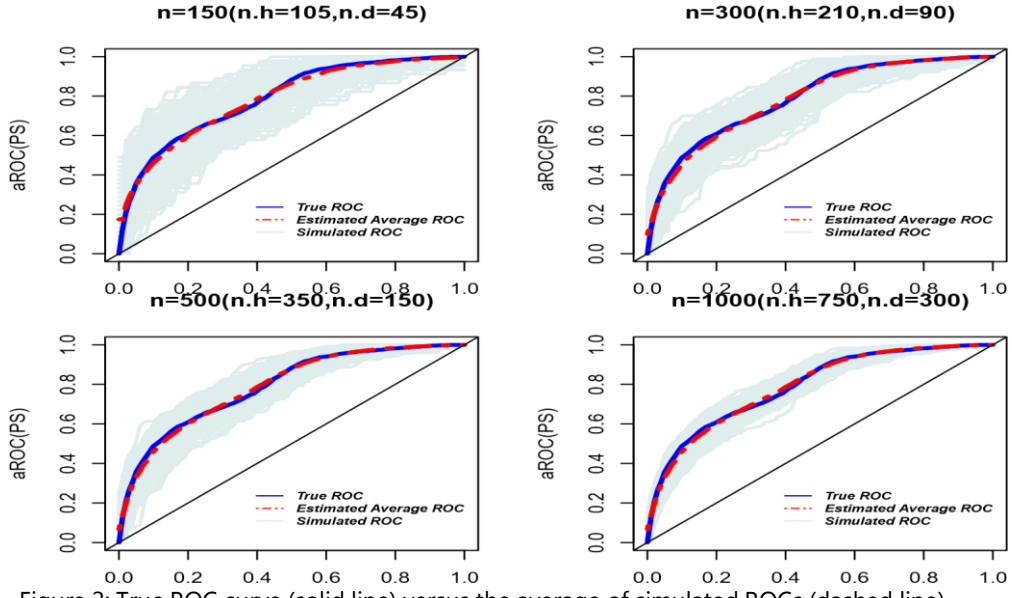


Figure 2: True ROC curve (solid line) versus the average of simulated ROCs (dashed line), along with 500 simulated ROC curves, for binary biomarker

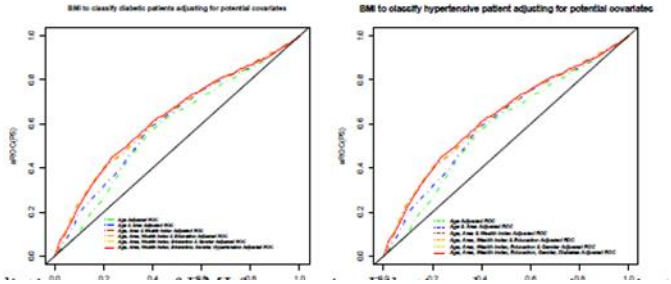


Figure 3: Checking predictive accuracy of BMI for screening Dibatic or hypertensive patients adjusting for multiple different demographic and socio-economic variable simultaneously

The construction of composite index to measure accessibility, quality and people behavior of drinking water in Indonesia

Hanin Rahma Septina[1], Aulia Dini[2]
[1]BPS-Statistics Indonesia
[2]STIS 53 Economic Division

Abstract
Indonesia is a country having the fifth largest world water reservation in forms of lakes, res-ervoirs, rivers and groundwater basins. Therefore, continuing access to clean drinking water and sanitation still needs appropriate concern. The composite index is constructed through three dimensions: accessibility, quality, and behavior based on Socio-Economic Survey pub-lished in 2016 and 2017. Weights of each variable are determined using PCA (Principal Component Analysis. Cumulative method is applied to classify thirty four provinces into three categories. The result shows that four provinces should be prioritized to improve its access, quality and behavior of drinking water. Those provinces are Nusa Tenggara Timur, Sulawesi Selatan, Sulawesi Barat, Papua.

Keywords
Water; composite index; PCA; SDGs

1. Introduction

Water has many functions in human life, one of which is for drinking. Indonesia is a country with the fifth world water reservation in forms of lake, reservoir, rivers, and groundwater ba-sins. As a fourth populated country in the world, Indonesia faces a challenge related to uni-versal access of save drinking water. Sustainable Development Goals (SDGs) as it is stated in goal six provides a guide to achieve sustainable access of drinking water. Resolution 64/292 also declared that water is a prerequisite to achieve realization of all human right.

This research aims to provide a composite index to monitor sixth goal of SDGs. Three di-mensions, accessibility, quality and behavior, are used to construct the index. Those three dimensions refer to government development plans namely RPJMN 2015-2019 as a govern-mental guide to provide universal access of save drinking water.

2. Methodology

This study is using secondary data from Socio Economic Survey in September 2016 and March 2017 published by BPS-Statistics Indonesia. The

total index (I_T) is made of three dimensions, namely accessibility, quality, and the behavior in using drinking water. Accessibility dimension (I_A) is described by three variables: households with a distance to drinking water sources are less than 1 km (X_1), households that less than 30 minutes to drinking water sources in one trip (X_2) and households with full of access to drinking water sources in last 24 hours (X_3). Quality dimension (I_K) is described by two operational variables: households having good physical quality of drinking water (X_4) and households using improved drinking water sources (X_5). Behavioral index (I_P) is described by two operational variables: households frequently cleaning water storage (X_6) and households using protected water storage (X_7).

To construct those three indexes, weights of each variable is determined using Principal Component Analysis (PCA), a mathematical procedure that is used to transform a number of variables (which may be correlate one to another) to a number non-correlated variable. These transformed-variables are called *principal component.* The number of principal component will be equal to the number of transformed variables. In this paper, the value of the weights is determined by the value of first principal component, because it gives largest contribution in explaining variation in the data.

Based on three indexes formed, an index ranking of 34 provinces is made. The results of the ranking is then used to classify the provinces into three categories: High, Moderate, and Low, using the cumulative method. Dalenius and Hogges (1959) state that cumulative method can minimize variance of units in the strata (Mukhopadhay, 2009).

3. Results

Model of Total Index:

Total Index (I_T) = 0.5402 I_A + 0.2939 I_K + 0.1659 I_P

We examined principal component of each variables then construct it into model. It can be seen that total index is most affected by accessibility index at 54.02% and less affected by behavioral index at 16.59%.

Model of each dimension:

Accessibility Index (I_A) = 0.4679 X_1 + 0.4717 X_2 + 0.0603 X_3

Quality Index (I_K) = 0.5614 X_4 + 0.4386 X_5

Behavioral Index (I_P) = 0.0165 X_6 + 0.9835 X_7

The same procedure is used to measure weights of each variable. Based on the three models, it shows that accessibility index is affected by variable households with a distance to drinking water sources are less than one kilometer at 46.79%, households that less than thirty minutes to drinking water sources in one trip at 47.17% and households with full of access to drinking

water sources in last twenty four hours by 6.03%. Quality index and behavioral index have similar interpretation to accessibility index.

Moreover, provinces are ranked from the highest index to the lowest one, each dimension is then categorized into high, moderate, and low and finally provinces are grouped into five re-gions based on a geographical main island formation (Sumatra, Jawa-Bali, Kalimantan, Su-lawesi, and other island such as Nusa Tenggara, Maluku, and Papua).

4. Discussion and Conclusion

RPJMN 2015-2019 stated that there are four principles for development of drinking water called 4K (Quality, Quantity, Continuity and Affordability). The affordability aspect is seen from the accessibility index, quality, quantity. Continuity aspect can be seen from quality index. Compo-site index built is also enriched with behavioral aspects of households in water treatment before being used for drinking.

The result shows that there is a relationship between drinking total index to region with a contingency coefficient of 0.5391 (strong and positive correlation). Based on drinking total index calculation, fourteen provinces are categorized in low index. They are Aceh, Bengkulu and Lampung in Sumatra Island. Kalimantan Barat, Kalimantan Selatan and Kalimantan Tengah are located at Kalimantan Island. Sulawesi Selatan, Sulawesi Tengah and Sulawesi Barat are located at Sulawesi Island. Maluku Utara, Maluku, Nusa Tenggara Timur and Pa-pua are categorized as other Island. It means there is a gap of total index between Jawa-Bali and non Jawa-Bali Island.

To see the relationship between categories of quality with region, the chi-square test with a significance level of 10% is used. The result shows that there is a relationship between cate-gories of quality to region with a contingency coefficient of 0.4553 (strong and positive cor-relation). Among eight provinces having high quality index, five of which are located in Java-Bali. Among sixteen provinces having low quality index, most of them come from outer Java-Bali Island. It means there is a gap of drinking water quality between Java-Bali and out-er Java-Bali. The similar pattern is also shown in accessibility index and behavioral index.

Government need to determine which provinces to be prioritized that can be seen through the composite index. Priority provinces are not only areas having low total index, but also low index in accessibility, quality and behavior. Those provinces are Nusa Tenggara Timur, Su-lawesi Selatan, Sulawesi Barat, Papua.

This result is confirmed by report of provincial snapshot published by UNICEF. It is stated that improved drinking water indicator of Nusa Tenggara Timur, Sulawesi Barat and Papua are below national average. Infrastructure development to areas with low accessibility can be applied such as building

road and bridge to accelerate access to water source. Other infra-structure development can be suggested such as building improved water storage and clean water pipe from water sources to households.

The next priority is provinces that have low total index with other two low indexes in acces-sibility or quality or behavior. Those provinces are Aceh, Lampung, Kalimantan Barat, Sula-wesi Tengah, Maluku dan Maluku Utara. This result shows that there is low access to drink-ing water in several provinces. This pattern is shown in remote areas and archipelago area such as Sulawesi Tengah, Maluku dan Maluku Utara. In this case, government intervention for the development of drinking water is needed, especially in remote areas.

In the other hand, provinces having high or moderate of access and quality index but still have a low behavioral index. Those provinces are Bali, Jawa Timur, Sulawesi Tenggara and Nusa Tenggara Barat. Suggested policy inform them about save drinking water criteria such as clear, have no color, taste, and smell. People behavior can be improved by informing them related to store drinking water before it consumes and clean water storage periodically. SDGs vision stated about universal access to drinking water, sanitation and hygiene. It also gives a mandatory message to monitor that no one is left behind.

Drinking water quality policy can be made such as educate society related to water pro-cessing before it is consumed such as purification , inform them about save drinking water criteria such as clear, have no color, taste, and smell. People behavior can be improved by informing them related to store drinking water before it consumes and clean water storage periodically.

Disclaimer

The author and co-authors declare there is not any potential conflict of interest with respect to the research, authorship, and/or publication of this article. The views expressed here are those of the individual author and co-authors and not necessarily those of BPS-Statistics Indonesia or its board, or officers, or staff.

Acknowledgment

We would like to thank Adhi Kurniawan for providing assistance during methodological research.

End Note

Next research is doing an exercise about city and district data in province Nusa Tenggara Timur, Sulawesi Selatan, Sulawesi Barat and Papua. Another plan is to do qualitative research about behavioral index.

References

1. Jolliffe, I.T., 2002. Principal Component Analysis, second edition, New York: Springer-Verlag New York, Inc.
2. Mukhophadhyay, P., *Theory and Methods of Survey Sampling*. New Delhi: PHI Learning Private Limited, 2009.
3. Organisation for Economic Co-operation and Development (OECD), *Handbook on Con-structing Composite Indicators Methodology and User Guide* OECD, France: OECD publish-ing, 2008.
4. United Nation. SUSTAINABLE DEVELOPMENT GOAL 6 : *Ensure availability and sustainable management of water and sanitation for all*. Accessed on 2 December 2018 from https://sustainabledevelopment.un.org/sdg6
5. Badan Perencanaan Pembangunan Nasional. Rencana Pembangunan Jangka Menengah Na-sional 2015-2019 pages 100-103. Accessed on 2 December 2018 from https://www.social-protection.org/
6. IUWASH. INDONESIA URBAN WATER, SANITATION, AND HYGIENE *Sulawesi Selatan, Ambon, Jayapura*. Accessed on 3 January 2019 from https://www.iuwashplus.or.id/
7. WHO Library Cataloguing-in-Publication Data. Safely managed drinking water - thematic report on drinking water 2017. Geneva, Switzerland: World Health Organization; 2017. Li-cense : CC BY-NC-SA 3.0 IGO
8. UNICEF. SDGs for Children in Indonesia Provincial snapshot: South Sulawesi. Accessed on 20 January 2019 from https://www.unicef.org/indonesia/Eng_South_Sulawesi_lowres2.pdf

Appendix 1. Rank and categories of Total Index, Accessibility Index, Quality Index and Behavioral Index

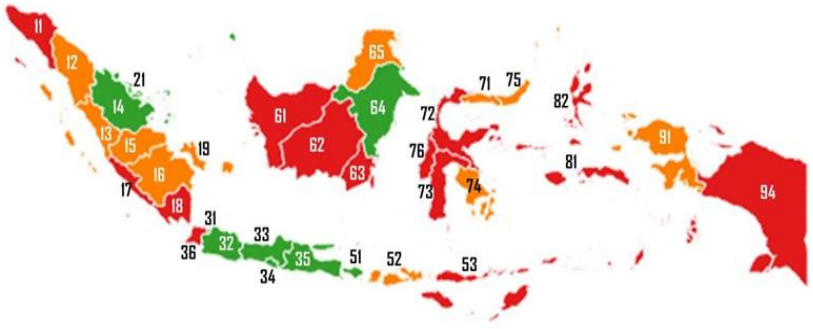
Province	Total Index (IT)		Accessibility Index (IA)		Quality Index (IK)		Behavioral Index (IP)	
	Rank	Category	Rank	Category	Rank	Category	Rank	Category
Aceh	27	Low	21	Low	30	Low	8	High
Sumatera Utara	17	Moderate	12	Moderate	20	Low	11	Moderate
Sumatera Barat	12	Moderate	15	Moderate	18	Moderate	13	Moderate
Riau	8	High	6	High	15	Moderate	19	Moderate
Jambi	20	Moderate	13	Moderate	23	Low	25	Low
Sumatera Selatan	19	Moderate	4	High	26	Low	23	Low

Bengkulu	30	Low	11	Moderate	34	Low	9	High
Lampung	26	Low	10	High	29	Low	26	Low
Kepulauan Bangka Belitung	16	Moderate	28	Low	12	Moderate	4	High
Kepulauan Riau	4	High	22	Low	3	High	7	High
DKI Jakarta	1	High	1	High	2	High	10	Moderate
Jawa Barat	9	High	14	Moderate	17	Moderate	1	High
Jawa Tengah	5	High	5	High	7	High	6	High
DI Yogyakarta	3	High	2	High	8	High	2	High
Jawa Timur	7	High	18	Moderate	6	High	20	Low
Banten	21	Low	19	Moderate	24	Low	14	Moderate
Bali	2	High	17	Moderate	1	High	24	Low
Nusa Tenggara Barat	18	Moderate	3	High	16	Moderate	32	Low
Nusa Tenggara Timur	34	Low	34	Low	27	Low	31	Low
Kalimantan Barat	22	Low	29	Low	19	Low	16	Moderate
Kalimantan Tengah	28	Low	16	Moderate	32	Low	5	High
Kalimantan Selatan	24	Low	8	High	9	Moderate	17	Moderate
Kalimantan Timur	6	High	7	High	10	Moderate	15	Moderate
Kalimantan Utara	10	Moderate	27	Low	31	Low	12	Moderate
Sulawesi Utara	13	Moderate	25	Low	4	High	21	Low
Sulawesi Tengah	29	Low	26	Low	11	Moderate	30	Low
Sulawesi Selatan	23	Low	23	Low	22	Low	34	Low
Sulawesi Tenggara	14	Moderate	20	Moderate	5	High	29	Low

Gorontalo	11	Moderate	9	High	14	Moderate	27	Low
Sulawesi Barat	31	Low	31	Low	28	Low	28	Low
Maluku	32	Low	32	Low	25	Low	18	Moderate
Maluku Utara	25	Low	30	Low	21	Low	3	High
Papua Barat	15	Moderate	24	Low	13	Moderate	22	Low
Papua	33	Low	33	Low	33	Low	33	Low

Appendix 2. Indonesian pattern of drinking water accessibility, quality and behaviour

Thematic map is made to visualize a different pattern of composite index. Red colored area indicates low composite index based on three criteria: accessibility, quality and behavior. Orange color area indicates moderate index and green color area indicates high index. The map shows us about relative gap among provinces, especially for provinces out of Java-Bali Island. Java-Bali has a better infrastructure related to drinking water. Most of low access, quality and behavior are located in eastern Indonesia especially provinces with many remote and archipelago areas. Prioritized provinces to be developed are Nusa Tenggara Timur, Sulawesi Selatan, Sulawesi Barat, Papua.



Picture 1. Visualization of composite index

Tematic map above can be seen as a region code below :

Region Code	Province
11	Aceh
12	Sumatera Utara
13	Sumatera Barat
14	Riau
15	Jambi
16	Sumatera Selatan
17	Bengkulu
18	Lampung
19	Kepulauan Bangka Belitung
21	Kepulauan Riau
31	DKI Jakarta
32	Jawa Barat
33	Jawa Tengah
34	DI Yogyakarta
35	Jawa Timur
36	Banten
51	Bali
52	Nusa Tenggara Barat
53	Nusa Tenggara Timur
61	Kalimantan Barat
62	Kalimantan Tengah
63	Kalimantan Selatan
64	Kalimantan Timur
65	Kalimantan Utara
71	Sulawesi Utara
72	Sulawesi Tengah
73	Sulawesi Selatan
74	Sulawesi Tenggara
75	Gorontalo
76	Sulawesi Barat
81	Maluku
82	Maluku Utara
91	Papua Barat
94	Papua

A solution to separation in Poisson Regression for small or sparse count data

Mominul Haque Mondol[1], M. Shafiqur Rahmanz[2], Wasimul Bari[3]
[1]University of Barishal
[2]Institute of Statistical Research and Training, University of Dhaka, Bangladesh
[3]Department of Statistics, University of Dhaka

Abstract

Separation or monotone likelihood can be observed in the fitting process of Poisson regression using maximum likelihood estimation (MLE) technique when one or more parameters diverge to infinity. The separation is very common in count data when sample size is small or there is huge number of zero count or there is sufficient number of strong predictors or mixture of two or more such conditions. The study investigates the consequence of separation in the standard Poisson models and provides a solution by incorporating Firths (Firth, 1993) type penalty term, which was originally proposed for bias reduction in MLE, in likelihood score equation. The modified score equation guaranteed an achievement of convergence and finite estimate of the regression coefficient. An extensive simulation study was conducted to assess the performance of penalized Poisson model over standard Poisson in the presence of separation. Several simulation scenarios were considered for creating complete or quasi-complete or near-to-separation by varying the sample size, proportion of event in the binary predictor which make separation, and the magnitude of beta coefficient log odds ratio relating the binary predictor to the response. The results revealed that the Firths penalized estimator, termed as `penalized Poisson', with profile likelihood based confidence interval performed well over the standard maximum likelihood based estimator in terms of bias, MSE and coverage in all simulation scenarios.

Keywords

Separation problem; Bias reduction; Sparse count data; Poisson model

1. Introduction

Separation or monotone likelihood, first reported by Albert and Anderson (Albert and Anderson, 1984), is a common case where one or more predictors have strong effects on response and hence it perfectly (or nearly perfect) predict the outcome of interest. In the case of complete separation, the responses and non-responses are perfectly separated by the predictor and in case of quasi- complete separation, the responses and non-responses are nearly separated by the predictor. The responses and non-responses are not

only separated by single predictor but also by a non-trivial linear combination of predictors. Separation commonly occurs when sample size is small and may occurs even if sample size is large but outcome is rare or there is sufficient number of strong predictors (Heinze and Schemper, 2002; Lesaffre and Albert, 1989).

In the presence of separation, maximum likelihood (ML) based binary logistic regression model faces several problems including frequent convergence failure and biased or infinite estimate of at least one regression coefficients, which are unrealistic and not-interpretable and provide misleading inference Cordeiro and Cribari-Neto (1998); Cordeiro and McCullagh (1991); Lesaffre and Albert (1989); Leung and Wang (1998); Santner and Duffy (1986); Schaefer (1983); Self and Liang (1987). Heinze and Schemper (Heinze and Schemper, 2002) showed an application of Firths Firth (1993) penalized method, which was originally proposed to reduce first order bias in the maximum likelihood estimate, to solve the problem of separation in the logistic regression. Further t Firths (Firth, 1993) method have been applied solve the problem of separation in many other models under generalized linear model (GLM) framework, which includes proportional and conditional logistic regressions and Cox PH model (Heinze and Puhr, 2010; Heinze and Schemper, 2001; Lipsitz et al., 2013)

Poisson regression, another case of GLM, is a widely used approach for modeling count data as well as for other non-negative data, which may also face similar problem of separation. The complete separation may occur in count data if, for a dichotomous predictor, there is zero cell count for all positive response (> 1) in one group of the predictor and positive cell count for response equal to zero in the other group of the predictor (Silva and Tenreyro, 2010). The quasi-complete separation may arise if there is non-zero cell (but few count) for all positive response in one group of the predictor. The separation created by a predictor often occurs in count data particularly when sample size is small, or even if large but there is a sufficient number of strong predictors, large number of zero response and/or proportion of event in binary covariate is rare. The consequences of separation in the Poisson regression with maximum likelihood based estimation are also similar to these with binary data model. Although Santos (Silva and Tenreyro, 2010) and Silvana (Silva and Tenreyro, 2011) addressed a method to avoid the problem of separation in count data model and anlyzed the data with some ad-hoc procedure, very limited number of studies have been conducted to addressed the problem of separation or monotone likehood in Posisson regression. This study investigates the problems of separation in Poisson regression and provided a solution by introducing Firth-type penalization (Firth, 1993) in the Poisson regression. The motivation is that both logistic regression and Poisson regression have the same score equation having different link function.

Therefore we followed Heinze and Schemper's (Heinze and Schemper, 2002) approach, that was used for the standard logistic model, to adapt Firth's penalty term in Poisson framework and derive a modified score equation for fixing the separation problems. The resulting regression is termed as `penalized Poisson' regression. The penalized Poisson is shown to provide finite estimate of the regression coefficient in the presence of any form of separation described earlier. A simulation studies were then conducted to assess the performance of penalized Poisson in comparison with standard Poisson in the presence of separation. We compared the performance of zero-inated Poisson to handle the excess of zero count in situation of separation.

The paper is organized as follows. Section 2 describes the methodolgy of Firth's method and its application in the Poisson regression framework and Section 3 described the simulation study and section ends the paper with a discussion.

2. Methodology

2.1 Application of Firth's method for solving separation in Poisson model

In this section, we primarily state the standard Poisson and then apply Firth-type penalization to solve the problems of separation in Poisson model. Let's have count response vector y = (y_1,y_2,...,y_n) and a n×p covariate matrix X. The count response can be modelled as E($Y_i|x_i$) = μ_i = exp($x_i^T\beta$), where i = 1,2,...,n and β^T = ($\beta 1$,...,β_p)^T. The regression coefficient β can be estimated by solving the following score equation:

$$U(\beta) = X^T(y_i - \mu_i) = 0, i = 1,2,...,n. \quad (1)$$

and the variance covariance matrix the estimated $\hat{\beta}$ can be found from the inverse of the information matrix, $V = I(\hat{\beta})^{-1} = (X^T W X)^{-1}$, where W = diag{$\mu_i$}. According to Firth's (Firth, 1993) "bias reduced" approach and Heinze and Schemper's (Heinze and Schemper, 2002) solution to separation in logistic regression, we added Firth-type penalty to Poisson score equation for estimating rth regression coefficient as follows

$$U_r^*(\beta) = U(\beta) + \Phi_r(\beta), r = 1,2,\dots,p. \quad (2)$$

With the penalty

$$\Phi_r(\beta) = \frac{1}{2} trace[I(\beta)^{-1}\{\partial I(\beta)/\partial \beta_r\}].$$

The derivative $\partial I(\beta)/\partial \beta_r$ for the *r*th coefficient can be expressed as:

$$\frac{\partial I(\beta)}{\partial \beta_r} = X^T W Z_r X, \quad (3)$$

where, Z_r = diag{x_{1r},x_{2r},...,x_{nr}} and r = 1,2,...,p. Fianlly putting equations (1) and (3) into equation (2), the modified Poisson score for the rth coefficient can be expressed as

$$U_r^*(\beta) = \sum_{i=1}^{n}\left[\left\{(y_i - u_i)(1 + {}^{h_i}/_{2}) + (1 - y_i - u_i)\,{}^{h_i}/_{2}\right\}x_{ir}\right] \quad (4)$$

$$= \sum_{i=1}^{n}\left[y_i - u_i + h_i({}^{1}/_{2} - u_i)\right]x_{ir} = 0, \ \ (r = 1, \dots, p) \quad (5)$$

where the h_i's are the ith diagonal elements of the 'hat' matrix H = $W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$. Then the Firth-type (FL) estimates of β can be obtained iteratively the usual way. The modified score equation confirm to achieve convergence and provide finite estimate of the regression cofficient.

3. Simulation study

The purpose of simulation study is to compare the performance of penalized Poisson with the standard Poisson in presence of separation. We showed the ZIP performance along side the above two method also.

3.1 Simulation design

To conduct simulation study, 1000 random samples of size n are generated from Poisson distributions with mean μ_i where

$$u_i = exp(\beta_0 + x_{i1}\beta_c + x_{i2}\beta_b), i = 1,2,\dots,n.$$

The covarite x_{i1} and x_{i2} are generated from Uniform(−1, 1) and Bernoulli(P) distributions respectively. We considered different sample size (n = 50,100,150,200,250,350,500) for fixing P = 0.4, $\beta = (\beta_0, \beta_c, \beta_b) = (-0.50, 0.30, -1.60)$. Again at fixed sample size n = 100, we varied the log odds ratio β_b = −0.40(−0.30),...,−2.20 to explore the model performance. Finally we varied the P = 0.05(0.05),...,0.95 to compare the performance. Under each simulation scenario, we generated 1000 datasets. We considered the cofficient of the regression parameters in such a way that it creates complete or quasi-complete separation according the definition discussed for count data in the earlier section. However, not all simulated data provide complete or quasi-complete separation but they can be considered as near-to-quasi-complete separation. We then summarized results from all such datasets.

3.2 Fitting the model and evaluating the properties

All the three models were fitted to all datasets. For each of the models we calculated bias $(\hat{\beta}_r)$ as $\left(\frac{1}{1000}\right)\left[\sum_{i=1}^{1000}(\hat{\beta}_{ri} - E(\hat{\beta}_r))\right]$ where $E(\hat{\beta}_r) = \sum_{i=1}^{1000}\hat{\beta}_r/1000$, the mean squared of error $MSE(\hat{\beta}_r) = Bias(\hat{\beta}_r)^2 + Var(\hat{\beta}_r)$the SimSE $(\hat{\beta}_r) = \left[\sum_{i=1}^{1000}(\hat{\beta}_{ri} - E(\hat{\beta}_r))/1000\right]^{1/2}$and the cover age as the proportion of 95% profile likelihood based confidence interval of (Lower$_{\hat{\beta}}$r,Upper$_{\hat{\beta}}$r) contained the true β_r, where r = 0, 1 and 2. The "zeroinfl" function belonging to R library "pscl" is used for fitting ZIP, "glm" for standard Poisson and self

written program in R is used for penalized Poisson. All computations were performed using R version 3.5.1.

3.3 Simulation results in case of separation

The results in Figure 1 revealed that penalized Poisson provided very lower amount of bias and MSE over standard Poisson and zero-inflated poisson (ZIP). For small sample case, the standard Poisson creates a huge amount of bias and MSE and the bias and MSE has a positive relationship with odds ratio of the corresponding binary covariate. But the proposed penalized Poisson showed better performance by reducing bias and MSE to some extents. The ZIP model is also unable to reduce the bias and MSE since the excess zero has been generated from Poisson distribution. Although the chance of separation decreases with the increment of sample size, as the large sample may face moderate amount of separation problem. On the contrary, the penalized Poisson provided small amount of bias and MSE for small sample case and it tends to zero if the sample size is large. SImilar findings can be observed for bias and MSE for the simulation scenarios with the proportion of event in binary predictor. The penalized Poisson with profile likelihood confidence interval showed better coverage performance tthan the standard Poisson and ZIP (results not shown).

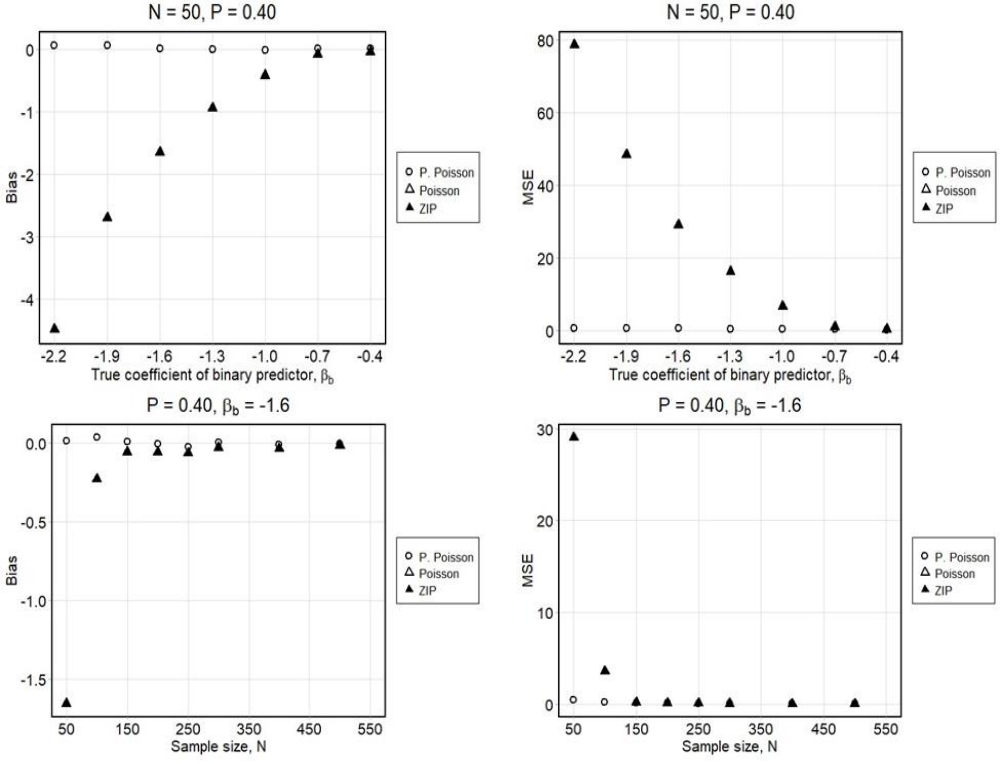


Figure 1: Behaviour of penalized Poisson (P.Poisson), standard Poisson (Poisson) and ZIP r with espect to bias and MSE.

4. Discussion

The problem of separation commonly occurs in count data for various reasons such as small sample size, large amount of zero count, availability of a set of strong or influential predictors, and unbalanced binary predictor. The standard Poisson and ZIP face problems like non-convergence of the likelihood and even if converge, they produce large amount of bias. This study investigated the problems in Poisson regression in the presence of separation and provided a solution to the problems by introducing penalized Poisson. The penalized Poisson is obtained by adding a penalty term to the standard Poisson, which is motivated by Firth (Firth, 1993) and Heinze (Heinze and Schemper, 2002) penalization to GLM and logistic regression respectively, aiming to reduce first order bias and solving the problems of separation. This penalized Poisson guaranteed convergence and finite estimate of the regression coefficient in the presence of separation, which was not possible by standard Poisson and ZIP. The simulation study creating the existence of separation in the data showed that standard Poisson and ZIP provided greater bias and MSE and inaccurate estimate of the standard error and coverage. The amount of bias and MSE increased with the increasing chances of separation in the data. Similarly coverage and SE are wrongly estimated in the presence of separation. Whereas the penalized Poisson performed much better results in all aspects in the presence of separation. The amount of bias in penalized Poisson estimates is negligible and coverage and SE were correctly estimated. The penalized Poisson provides an ideal solution to problem of separation in count data.

References

1. Albert, A. and J. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10.
2. Cordeiro, G. M. Albert, A. and J. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10.
3. Cordeiro, G. M. and F. Cribari-Neto (1998). On bias reduction in exponential and non-exponential family regression models. *Communications in Statistics-Simulation and Computation 27*(2), 485–500.
4. Cordeiro, G. M. and P. McCullagh (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 629–643.
5. Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

6. Heinze, G. and R. Puhr (2010). Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Statistics in medicine 29*(7-8), 770–777.
7. Heinze, G. and M. Schemper (2001). A solution to the problem of monotone likelihood in cox regression. *Biometrics 57*(1), 114–119.
8. Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine 21*(16), 2409–2419.
9. Lesaffre, E. and A. Albert (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society. Series B (Methodological)*, 109–116.
10. Leung, D. H.-Y. and Y.-G. Wang (1998). Bias reduction using stochastic approximation. *Australian & New Zealand Journal of Statistics 40*(1), 43–52.
11. Lipsitz, S. R., G. M. Fitzmaurice, S. E. Regenbogen, D. Sinha, J. G. Ibrahim, and A. A. Gawande (2013). Bias correction for the proportional odds logistic regression model with application to a study of surgical complications. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(2), 233–250.
12. Santner, T. J. and D. E. Duffy (1986). A note on a. albert and ja anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika 73*(3), 755–758.
13. Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine 2*(1), 71–78.
14. Self, S. G. and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association 82*(398), 605–610.
15. Silva, J. S. and S. Tenreyro (2010). On the existence of maximum likelihood estimates in logistic regression models. *Economics Letters 107*(2), 310–103.
16. Silva, J. S. and S. Tenreyro (2011). Poisson: some convergence issues. *The Stata Journal 11*(2), 207–212.

Inclusive economic growth in Java

Valent Gigih Saputri
Statistics Indonesia, Jakarta-Indonesia

Abstract

Inclusive economic growth is an intensive growth between economic sectors and without discrimination against labour (pro poor). Inclusive economic growth must be able to reduce poverty, reduce inequality and increase employment. Java Island is a regional area that has become the economic centre and the centre of government in Indonesia. Cumulative GRDP in all provinces on Java Island controls more than 50 percent of national GDP. In addition, more than 50 percent of workers in Indonesia also work on Java. However, the poverty level of each province varies considerably. Likewise with income inequality and employment that is not optimal. This study aims to determine the level of economic inclusiveness in each province on Java and what factors influence it. This study uses data on income per capita, gross fixed capital formation, gini ratio, inflation, net enrolment rates, GRDP by sector, minimum wage rate, and unemployed rate. The scope to be studied is all provinces in Java from 2010 to 2017. The measurement of inclusive economy uses the concept of Poverty-Equivalent Growth Rate (PEGR) (Klasen (2010)). Whereas to see the influencing factors used panel data regression analysis method. The results obtained are that almost all provinces on Java have not experienced inclusive economic growth. Factors influencing inclusive economic growth are gross fixed capital formation, net enrolment rate, primary and tertiary sector contribution. Indonesian government needs to do many programs to make economic growth inclusive, not just relying on high rates of growth.

Keywords

Inclusive Economic Growth; Data Panel; Poverty; Economic Inequality; Employment

1. Introduction

Inclusive economic growth is an intensive growth between economic sectors and without discrimination against labor. In addition, inclusive economic growth is also referred to as the concept of pro-poor growth. Inclusive economic growth must be able to reduce the "disadvantaged" groups in the economy. It is expected that with this inclusive economic growth, disparity between groups can be reduced (Klasen, 2010). The World Bank defines inclusive economic growth as growth that focuses on expanding

access to economic assets, successfully expanding markets and creating opportunities for the next generation. Whereas according to UNDP, inclusive economic growth defines growth based on the production side and GDP income. The process and results of growth in inclusive economic growth involve participation from all parties to produce equal benefits from that growth (Suryanarayana, 2007). According to Rusastra, inclusive growth can be said as growth that is able to synergize economic growth, increase employment opportunities, and alleviate poverty (Rusastra, 2011). Whereas according to Habito, inclusive growth is defined as GDP growth which can reduce poverty. Habito also explained that the economic structure and sectoral composition in economic growth are important factors to achieve inclusive growth, with a general statement that stronger growth in the structure of agriculture will accelerate the decline in poverty (Habito, 2009). Based on the description above, inclusive economic growth is economic growth that can reduce poverty, reduce inequality and increase employment.

Many parties consider that Indonesia's economic growth has entered an inclusive stage. During the last five years, Indonesia's economic growth has been relatively stable at 5 percent per year. However, research conducted by Sholihah (2014) shows that Indonesia's overall economic growth has not yet reached the level of inclusive economic growth. It is interesting when this inclusive growth is seen based on one of the regional areas which has become the economic centre as well as the centre of government in Indonesia, Java. Over the past five years, cumulative GRDP in all provinces in Java has more than 50 percent of national GDP. In addition, more than 50 percent of workers in Indonesia also work on Java. However, the poverty level of each province varies considerably. Likewise with income inequality and employment that is not optimal.

This study aims to determine the level of economic inclusiveness in each province on Java. In addition, this study also aims to find out what factors influence the inclusive economic growth that occurs.

2. Methodology

This study uses data on per capita income, gross fixed capital formation on the basis of constant price, gini ratio, inflation rate, investment, net enrollment rate, gross regional domestic product by sector, minimum wage rate, and unemployment. These data were obtained from the National Socio-Economic Survey, the National Labor Force Survey, the Intercensal Population Survey and other data from the Statistics Indonesia (BPS).

The scope to be examined in this study are all provinces in Java in 2010 until 2017. To find out the size of inclusive economic growth, the measurements used by Klasen (2010) were used with the concept of Poverty-Equivalent Growth Rate (PEGR).

$$IG_{ij} = \frac{E_{ij}}{E_j} * G_e$$

Where :

IG_{ij} : coefficient of inclusive economic growth
E_{ij} : growth of group i in relation to indicators j
E_j : growth indicator j
G_e : economic growth

Growth is neutral when $IG_{ij} = G_e$, every people receive benefits proportionately from growth. When $IG_{ij} > G_e$, population the poor receive more benefits of growth (pro poor growth). The condition is not yet pro poor growth when $0 < IG_{ij} < G_e$. The benefits of more growth non-poor population accepted (inequality increases) but still a reduction poverty. When $IG_{ij} < 0$, benefits of growth enjoyed the population is not poor, poverty increase (anti pro poor growth).

Table 1. Formula to Measure Inclusive Economic Growth

Inclusive economic growth in reducing poverty	Inclusive economic growth in reducing inequality	Inclusive economic growth in increasing employment
$IG_P = \frac{E_{p.eg}}{E_{p.in}} * G_e$ (2)	$IG_{disp} = \frac{E_{dis.peg}}{E_{disp.in}} * G_e$ (3)	$IG_{em} = \frac{E_{em.eg}}{E_{em.in}} * G_e$ (4)
IG_{ij}: coefficient of inclusive economic growth in reducing poverty E_{ij} : poverty elasticity of average income E_j : poverty elasticity towards economic growth G_e : economic growth	IG_{disp} : coefficient of inclusive economic growth in reducing inequality $E_{disp.peg}$: elasticity of inequality to average income $E_{disp.in}$: elasticity of inequality towards economic growth G_e : economic growth	IG_{em} : coefficient of inclusive economic growth in absorbing workers $E_{em.eg}$: labor absorption elasticity $E_{em.in}$: elasticity of employment to economic growth G_e : economic growth

Whereas to look at the factors that influence inclusive economic growth, a panel data regression analysis method is used. The term of panel data refers to the pooling of observations on a cross-section households, countries, etc. over several time periods (Baltagi, 2005). Panel data gives more informative data, variability, degrees of freedom, efficiency and less collinearity among variables. Panel data can enrich empirical empirical analysis in ways that may not be possible if we use only cross-section or time series data (Gujarati, 2004).

There are three types of panel data models: pooled regression, fixed effect model, and random effect model. The common model is :

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it} \quad (5)$$

Y_{it} is the dependent variable, α is the intercept term, β is a k × 1 vector of parameters to be estimated on the explanatory variables, xit; t = 1, …, T; i = 1, …, N.

Pooled regression model has the same intercept and slope for every individual. There's no interindividual heterogeneity (α equal). The assumptions is same with cross-section data regression. Ordinary Least Squares (OLS) produce most efficient and consistent estimators (Park, 2011).

Fixed effect model examines individual differences in intercepts, assuming the same slopes and constant variance across individual. Since an individual specific effect is time invariant and considered a part of intercept, individual effect is allowed to be correlated with other regressor. Fixed effect model is estimated by least squares dummy variable (LSDV) regression (OLS with a set of dummies) and within effect estimation methods (Park, 2011).

Random effect model assumes that individual effect isn't correlated with any regressor and estimates error variance specific to groups or time. Hence, individual effect is an individual specific random heterogeneity or a component of the composite error term. The intercept and slopes of regressors are the same across individual. The difference among individuals (or time periods) lies in their individual specific errors, not in their intercepts. Random effect model is estimated by generalized least squares (GLS) when a covariance structure of an individual is known. The feasible generalized least squares (FGLS) or estimated generalized least squares (EGLS) method is used to estimate when a covariance structure is not known. There are various estimation methods for FGLS including the maximum likelihood method and simulation (Park, 2011).

To determine the best model, used Chow Test with F test statistics to choose between pooled regression or fixed effects with the null hypothesis is pooled regression is better than fixed effects model; Breush-Pagan Lagrange Multiplier to choose between random effects or pooled regression model with the null is pooled regression model is better than random effects model; and Hausman test to choose between fixed effects or random effects with the null hypothesis for Chow test is random effects model is better than fixed effects model. The next steps after the model is selected is determine the suitable estimation methods : Ordinary Least Square (OLS) or Feasible Generalized Least Square (FGLS) based on its residual variance-covariance structure using Lagrange Multiplier (LM) test for homogeneity on variance-covariance structure and LM test for cross-sectional correlation on variancecovariance matrix.

Table 2. Variable Included in The Fixed Model

Variable		Definition
Dependent	IG_p	Coefficient of inclusive economic growth in reducing poverty
	$IGdisp$	Coefficient of inclusive economic growth in reducing inequality
	IG_{em}	Coefficient of inclusive economic growth in absorbing workers
Indepedent	Inc	Income per capita
	NER2	Net enrollment rate for junior high school
	NER3	Net enrollment rate for senior high school
	GFCF	Gross fixed capital formation
	Gin	Gini ratio
	Inf	Inflation
	Pr	Primary sector contribution
	Sc	Secondary sector contribution
	Th	Tertiary sector contribution
	W	Minimum wage rate
	Unemp	Unemployment rate

3. Results

From the results of the inclusive economic growth index calculation, most provinces in Java have not yet reached the stage of inclusive economic growth. The calculation results can be seen in table 3.

Table 3. Inclusive Economic Growth Index

Index	Jakarta								Jawa Barat							
	2010	2011	2012	2013	2014	2015	2016	2017	2010	2011	2012	2013	2014	2015	2016	2017
Ge	0.06	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.06	0.05
IGp	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.03	0.03	0.04	0.04
IGdisp	0.05	0.04	0.05	0.05	0.05	0.04	0.04	0.05	0.04	0.03	0.03	0.05	0.03	0.03	0.04	0.04
Igem	0.12	-0.03	0.04	-0.04	-0.01	0.01	0.02	-0.06	0.00	0.02	0.06	0.01	0.02	-0.02	0.02	0.06
Index	Jawa Tengah								Yogyakarta							
	2010	2011	2012	2013	2014	2015	2016	2017	2010	2011	2012	2013	2014	2015	2016	2017
Ge	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
IGp	0.06	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
IGdisp	0.05	0.05	0.05	0.03	0.04	0.05	0.03	0.03	0.04	0.03	0.04	0.04	0.04	0.03	0.04	0.04
Igem	-0.01	0.01	0.03	0.00	0.00	-0.01	0.00	0.04	-0.07	0.02	0.03	-0.02	0.04	-0.03	0.06	0.01
Index	Jawa Timur								Banten							
	2010	2011	2012	2013	2014	2015	2016	2017	2010	2011	2012	2013	2014	2015	2016	2017
Ge	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.06	0.07	0.07	0.06	0.05	0.05	0.05	0.06
IGp	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.03	0.04	0.04	0.04	0.03	0.03	0.03	0.04
IGdisp	0.05	0.05	0.06	0.04	0.05	0.03	0.05	0.05	0.04	0.04	0.04	0.04	0.05	0.05	0.03	0.04
Igem	-0.04	0.00	0.03	0.01	-0.01	0.01	-0.02	0.05	0.20	-0.05	0.02	0.00	0.03	0.00	0.05	0.00

*: yellow column to indicate the inclusive economic growth

After going through the stages of model selection, the best model to analyse what factors are influence the inclusive economic growth that occurs listed in table 4. Model I dan III are pooled regression with cross-section weight and white cross section, whereas for model II is pooled regression.

Table 4. Summary of Output Panel Data Regression

Model I (independent variable is IG_p)						
Dependent Variables	**Coeff**	**Std Error**	**Prob t-stat**	**Prob F-stat**	**R^2**	**Adj R^2**
C*	0,0336	0,0114	0,0051	0,0000	0,6057	0,5588
NER2*	-0,0008	0,0002	0,0000			
NER3*	0,0005	0,0002	0,0165			
Gin*	-0,0864	0,0307	0,0075			
LOG (INC)	0,0016	0,0010	0,1175			
LOG (GFCF)*	0,0032	0,0005	0,0000			
Model II (independent variable is IG_{disp})						
Dependent Variables	**Coeff**	**Std Error**	**Prob t-stat**	**Prob F-stat**	**R^2**	**Adj R^2**
C*	-0,0668	0,0163	0,0002	0,0000	0,7284	0,6961
INF	0,0000	0,0002	0,7615			
NER3*	-0,0006	0,0001	0,0000			
LOG (PR)*	0,0053	0,0008	0,0000			
LOG (SC)*	-0,0139	0,0018	0,0000			
LOG (TH)*	0,0158	0,0017	0,0000			
Model IIII (independent variable is IG_{em})						
Dependent Variables	**Coeff**	**Std Error**	**Prob t-stat**	**Prob F-stat**	**R^2**	**Adj R^2**
C	-3,6448	3,6557	0,3255	0,000	0,5427	0,4882
LOG (W)	-0,0248	0,0969	0,7994			
UNEMP	-0,0028	0,0105	0,8215			
LOG (GFCF)*	0,6898	0,2455	0,0079			
NER3*	-0,0038	0,0014	0,0110			

*: significance at the 5 % level

4. Discussion and Conclusion

During 2010-2017, the coefficient of inclusive economic growth in Java has not been able to reduce poverty and disparity. The coefficient value is positive even though it is still lower than economic growth. The economic growth that occurs is indeed able to reduce poverty and inequality, but the growth is not spread evenly and the benefits for the poor are not yet received. This is

different from the coefficient of inclusive economic growth in employment. In 2010, Jakarta and Banten provinces experienced inclusive growth. The same thing happened in Yogyakarta province in 2016 and West Java province in 2017. In contrast to the two previous coefficients, there were several provinces which had a negative coefficient of inclusive economic growth in employment. This can be interpreted that the economic growth that occurs is not able to absorb existing labour.

One characteristic of inclusive economic growth is it can reduce poverty. From the model I, the net enrollment rate for junior high school, net enrollment rate for senior high school, gini ratio and growth of gross fixed capital formation have an effect on the coefficient of inclusive economic growth in reducing poverty. For the variables of net enrollment rate for junior high school and gini ratio that are negatively valued in the model indicate that economic growth to reduce poverty will be more inclusive if the two variables are getting smaller. Vice versa, the higher net enrollment rate for senior high schools and the addition of gross fixed capital formation will make economic growth more inclusive.

In addition to reducing poverty, inclusive economic growth is also indicated by a decrease in inequality. The estimation results in model II show that there are 4 independent variables which influence the coefficient of inclusive economic growth which can reduce inequality. The independent variable are the net enrolment rate for senior high school, the growth of primary sector contributions, the growth of secondary sector contributions, and the growth of tertiary sector contributions. The primary sector referred to here is the agricultural sector and the mining and quarrying sector, while the secondary sector includes the processing industry sector, the electricity-gas-water sector and tertiary sectors including other sectors besides the primary and secondary sectors. The growth in contribution of the primary and tertiary sector has a positive effect on the coefficient of inclusive economic growth to reduce inequality. The greater contribution of the two sectors, the economic growth in reducing inequality is increasingly inclusive. The inflation rate variable too, even though this variable has no significant effect. The variable net enrolment rate for senior high schools has negative impact on the coefficient of inclusive economic growth to reduce inequality, even though the impact is very small. Variable growth in the contribution of the secondary sector also has a negative impact. The smaller the value of these two variables will reduce inequality because the economic growth that occurs is more inclusive.

The coefficient of inclusive economic growth in absorbing workers can be seen in model III. From the estimation results, variable unemployment and minimum wage rate growth have a negative effect on the coefficient of inclusive economic growth to absorbing workers, even though the effect is not

significant. Likewise, the variable net enrolment rate for senior high schools has a negative and significant effect on the coefficient of inclusive economic growth to absorbing workers. The smaller unemployment, the growth of the minimum wage rate and the net enrolment rate for senior high schools will cause economic growth to become inclusive so that it can absorb labour. In contrast to the independent variables in the model, the gross fixed capital formation growth variable has a positive and significant impact on the coefficient of inclusive economic growth in absorbing workers.

From all the models formed, it can be concluded that the primary and tertiary sectors that contribute greatly to economic growth also influence the inclusive growth that occurs. This condition is different from the secondary sector where the addition of the secondary sector contribution will reduce inclusive economic growth. This is due to indications of deindustrialization in Indonesia (Saputri, 2013). The increase in gross fixed capital formation growth variables also contributes to inclusive economic growth. Variable gross fixed capital formation growth, especially on physical investment can reduce poverty (Saleh, 2002). Interesting when looking at the variable net enrolment rate for senior high schools in their influence in inclusive economic growth. Increasing the variable net enrolment rate for senior high schools can have a positive effect on inclusive economic growth in reducing poverty and inequality. However, this variable has a negative effect on inclusive economic growth in absorbing workers. The increasing participation of senior high school students who attend school according to their age will cause economic growth to become inclusive. But on the other hand, the existing workforce will decrease. This indicates that workers in senior high school age (15-19 years) still occur a lot. The data from Statistics Indonesia show that from 2010 to 2018 there were 20 to 25 percent of the population aged 15-19 years who worked, while the population at that age who attended school was around 50 to 60 percent of the total population.

Based on the results of the research that has been done, the Indonesian government should focus on equitable distribution of the results of economic growth. This can be done by making pro-poor growth programs: building houses for low-income communities, decent and affordable public transportation, creating a comfortable investment climate for investors, etc. The situation will be far from inclusive if Java as the centre of economic growth does not become a concern of both central and regional governments in equitable distribution of economic growth.

References

1. Ali, Ifzal & Hyun Hwa Son. (2007). Measuring Inclusive Growth. Asian Development Review Vol. 24, No. 1, pp. 11–31. Asian Development Bank.

2. Baltagi, Badi H. (2005). Econometric Analysis of Panel Data (third ed.). Chichester: John Wiley &
3. Sons Ltd.
4. Gujarati, Damodar N. (2004). Basic Econometrics [Fourth Edition. McGraw Hill India.
5. Habito, Cielito F. (2009). Patterns of Inclusive Growth in Developing Asia : Insights from an Enhanced Growth-Poverty Elasticity Analysis. Asian Development Bank Institute (ADBI) working paper series No. 145.
6. Klasen, Stephen. (2010). Measuring and Monitoring Inclusive Growth : Multiple Definitions, Open Questions, and Some Constructive Proposals. ADB Sustainable Development Working Paper Series, Asian Development Bank.
7. Park, H. M. (2011). Practical Guides To Panel Data Modeling: A Step by Step Analysis Using Stata.
8. International University of Japan.
9. Rusastra, I Wayan. (2011). Reorientation of Paradigms and Poverty Reduction Strategies in Overcoming the Impact of the Global Economic Crisis. Agricultural Innovation Development 4 (2): 87-102. Agricultural Socio-Economic and Policy Center: Bogor.
10. Saleh, Samsubar. (2002). Factors Affecting Regional Poverty Levels in Indonesia. Journal of Development Economics (Developing Country Studies) Vol.7 No. 2, pp. 87-102.
11. Saputri, Valent Gigih. (2013). Deindustrialization in Indonesia in 1990-2011. Institute of Statistics Jakarta.
12. Sholihah, Dyah HA. (2014). Inclusive Growth : Factors that Influence and Impact on the Growth of The Middle Class. Bogor Agricultural University.
13. Suryanarayana, M.H. (2008). Inclusive Growth : What is so Exclusive about It?. Indira Gandhi Institute of Development Research.

Childhood undernutrition in Bangladesh: A policy-suggestive empirical analysis

Mohaimen Mansur, Md. Saddam Hossain
Institute of Statistical Research and Training (ISRT), University of Dhaka

Abstract

Childhood undernutrition remains a major public health problem in Bangladesh despite a decade of interventions aimed at reducing it. It duly earned ample attention in academic literature with increasing interest in socio-economic determinants of it. We argue that many such empirical studies fail to capture complex interactions among potential risk factors that may be crucial for intricate understanding of child malnutrition. Consequently, these findings provide incomplete insight into the problem and are of limited value to policy makers. In this paper we take advantage of a machine learning tool to determine interactions among potential drivers of malnutrition, and use them to identify groups of children that are at high risk of being undernourished. By using data from a nationally representative survey we demonstrate statistical importance of such interaction-induced risk classifications. We justify that our findings may be helpful for designing resource-efficient small-scale intervention programs targeted towards improving nutrition status among children.

Keywords

Undernutrition; Stunting; Interaction; Classification Tree; Odds ratio

1. Introduction

Bangladesh has made laudable progress in various health indicators including women's health, life expectancy and family planning, but the progress in improving nutrition status among children is particularly slow despite years of interventions aimed at reducing childhood undernutrition. According to the recent national health survey [1], the prevalence of stunting (low-height-for-age) declined from 51% in 2004 to 36% in 2014 while the prevalence of underweight (low-weight-for-age) reduced from 43% in 2004 to 33% in 2014. These figures are still unacceptably high compared to the developed world [2], and particularly alarming considering that malnutrition has fatal short-term consequences like child mortality [3] as well as long-term health effects including chronic illness and disabilities [4,5]. Research found that malnourished children are physically and intellectually less productive than well-nourished children [4,5].

As a long-standing public health problem in Bangladesh childhood malnutrition earned substantial attention in academic research with an increasing interest in socio-economic determinants of it. Recent studies identified several risk factors including children's low birth weight, short preceding birth interval, underweight mothers, low level of parents' education and poor socio-economic status [6-10]. In this paper, we point out a number of methodological and practical issues with existing studies. First, we argue that majority of empirical studies look at potential risk factors of childhood malnutrition in isolation and overlook complex interplay among several factors that need to be understood for deeper insight into the problem. A possible reason for this is over reliance on regression type models which are not naturally wellsuited for capturing interactions. Number of interactions increases drastically with number of variables, leading to curse of dimensionality in regression, which again results in over parametrization and estimation difficulty. Second, studies on child malnutrition in Bangladesh largely analyze data from the nationally representative Bangladesh Demographic Health Survey (BDHS), but do not adjust statistical inference for the inherent complexity of BDHS's two-stage survey design. It is well-known that multistage designs leads to higher standard errors of estimates and inability to account for this will result in false statistical significance of regression estimates. Last and most importantly, we argue that policy values of existing studies are limited in the sense that their findings and recommendations are rather generic. For example, finding that mother's education is strongly associated with prevalence of child undernutrition, and recommending to increase women's school participation is a very generic suggestion. Interventions are often more effective when small-scale programs are designed to target the most vulnerable group.

In this paper, we make an attempt to address the criticisms identified above. First, we exploit classification tree, a method borrowed from machine learning arena, to identify interaction among variables. Unlike parametric regression models this nonparametric method offers a more flexible data-driven way to fit complex data and also offers easy and intuitive interpretation to findings. Second, identified interactions are incorporated into a logistic regression in order to gauge their marginal effects on malnutrition in children. We appropriately adjust standard errors for survey design complexity in order to draw precise statistical inference on both individual factors and interactions. Finally, we discuss practical usefulness of our findings in formulating vulnerability-specific intervention programs for reducing the burden of childhood undernutrition.

2. Methodology

a. Data and variables

Data on child nutrition have been extracted from the database of the Bangladesh Demographic Health Survey (BDHS) 2014, a nationally representative survey conducted between June and November 2014. Anthropometric, demographic and socio-economic information are available on 7,131 living children of aged under five years.

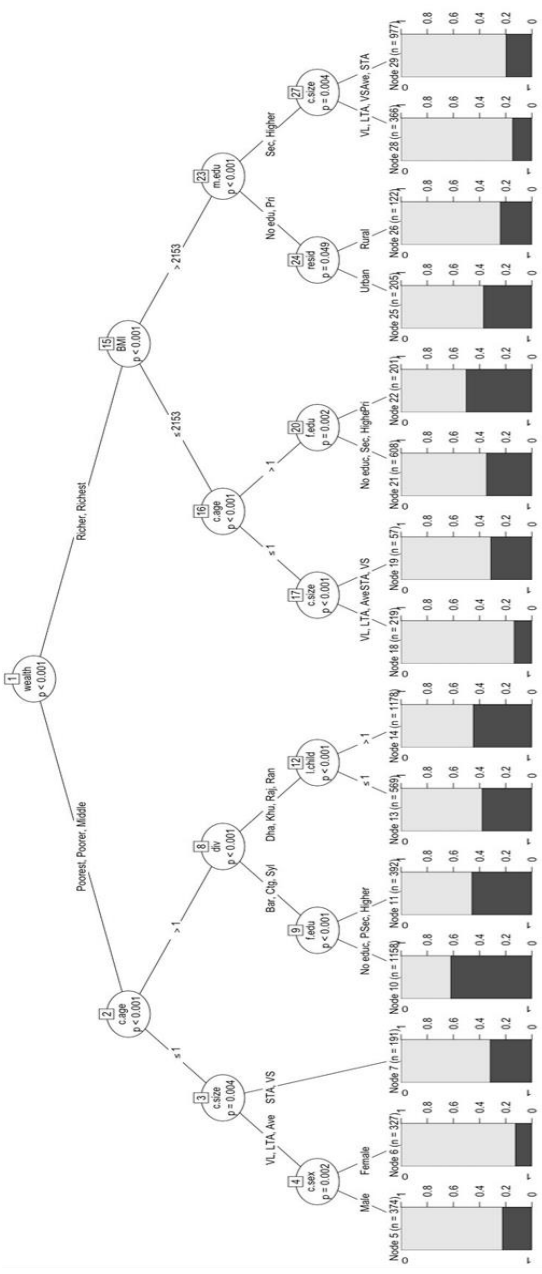
The main response variable of interest is stunting (low-height-for-age). Following recent literature we consider a number of explanatory variables that were found to be significantly associated with stunting among children. These include information on children, such as their age, gender and size at birth, and information on parents, e.g., mothers' age at first birth, number of living children, mothers' BMI and parents' education and employment status. Some household level information, e.g., wealth status, place of residence (rural or urban), type of toilet facilities and administrative divisions of residence are also considered.

b. Methods

In a two-step analysis we use a machine learning approach and a classical regression model in succession. In the first step of the analysis, the classification tree, the machine learning method first proposed in [11] is fit to the data to identify possible interactions among predictors of childhood malnutrition. In the second step, the interactions identified by the trees are incorporated in a logistic regression framework in addition to individual explanatory variables, and their likely impact on childhood undernutrition measured.

A classification tree divides the predictor space into a number of small regions based on simple rules. The rules originate from recursive binary splits on the predictors and are produced with an aim to achieving minimum error in classification of subjects into response categories, e.g., whether a child is undernourished or not. All the subjects within a region/rule are classified as the category where the majority belong. A classification tree is nonparametric in nature and holds a number of advantages over parametric regression type models including minimal distributional assumptions and easier interpretation. The most important advantage, however, is the tree's inherent capability of capturing complex, such as non-linear, interactions of variables in high dimensional settings. The standard classification tree techniques are often criticized for biased selection of variables which have many possible splits and missing values. In this paper, we opt to use a conditional inference framework proposed in [12].

Figure 1: Classification tree for stunting in children aged under five years in Bangladesh



Note: The variables appearing higher in the tree are predictively most significant. Each final node shows the size of the relevant segment of the sample and the height of the dark black bar gives the proportion of stunted children within the segment. Abbreviations for full variable names are used for better legibility. *c.age, c.size* and *c.sex* refer to age, size at birth and gender of children, respectively. *f.edu* and *m.edu* refer to highest education of father and mother. *resid* and *div* mean area of residence and division. VL, LTA, *Ave, STA* and *VS* refer to categories of size of babies at birth, namely very large, larger than average, smaller than average and very small.

While trees offer flexible approaches to reveal interactions among explanatory variables, it does not provide size of effects of these interactions on response, e.g., child malnutrition. To achieve these we follow [13] to incorporate interactions/rules as dummy variables in a logistic regression setting.

3. Results

Figure 1 shows results from the classification tree. Wealth offers the first split at the root and therefore, appears to be the most important socio-economic predictor of childhood stunting. Children born in households with middle or lower level wealth classes are placed under the left branch and are, in general, more prone to stunting. The second most important predictor on this side of the tree is children's age.

Stunting is more prominent in older children. The classification tree identifies groups of children who are the most vulnerable to stunting through stratification of the predictor space. More than 60% of the children who are older than a year, live in relatively lower-wealth households situated in either Barisal, Chittagong or Sylhet district and who have relatively lowly educated fathers are stunted. Among similar children whose fathers have more than primary level education, 45% are stunted. For children who are from other districts, number of living children appears to be more important than fathers' education. The proportion of stunting among children who are born to mothers with more than one living child is about 44% compared to 38% for children born to mothers who have only one living children. Baby girls who are born to mid or lower level wealth class households but are younger than a year old and are of average to very large size at birth are the least stunted (10%). Among boy babies with similar characteristics the prevalence of stunting is more than 20%.

The principal right hand branch originated from the root of the tree contain lower proportion of stunting cases in general, but shows considerable variation across smaller segments. Interestingly, mothers' BMI appears to be the second most important drivers of childhood malnutrition among children who are born to relatively richer households. Children born to mothers' with lower BMI (lower than 21.53) have higher percentage of stunting than those who are born to relatively high-BMI mothers. Almost half of the Children who are more than a year old, who come from richer families and are born to higher-BMI mothers and lowly educated fathers (not above primary level) are stunted. Least stunted groups of children with a stunting proportion of 10% are those who are born in rich families, have low-BMI mothers, are a year old or younger and were of average or larger size at birth. Similarly low percentage of stunting is observed among children who are born in rich families, have high-BMI and highly educated mothers and were of average or larger size at birth.

In the second stage of analysis, we identify four rules from the tree that accommodate children with the highest prevalence of stunting. The rules are:

Rule1: Household wealth status in {poorest, poorer, or middle}and child's age > 1 year and division of residence in {Barishal, Chittagong, Sylhet} and

fathers' highest level of education in {No education, primary level} (stunting percentage = 62%, relevant node = 10)

Rule2: Household wealth status in {poorest, poorer, or middle}and child's age > 1 year and division of residence in {Barishal, Chittagong, Sylhet} and fathers' highest level of education in {secondary, higher} (stunting percentage = 45%, relevant node = 11)

Rule3: Household wealth status in {poorest, poorer, or middle}and child's age > 1 year and division of residence not in {Barishal, Chittagong, Sylhet} and number of living children > 1 (stunting percentage = 44%, relevant node = 14)

Rule4: Household wealth status in {richer or richest}and mother's BMI < 21.53 and child's age > 1 year and fathers' highest level of education in {No education, primary level} (stunting percentage = 51%, relevant node = 22)

Table 1: Results of logistic regression for stunting in under-five children in Bangladesh

Variables	Estimate	Std. Error	t statistic	p-value	OR	95% CI for OR
(Intercept)	-0.773	0.778	-0.994	0.320	0.461	(0.101, 2.119)
Age	0.585	0.084	6.983	0.000	1.795	(1.523, 2.116)
Sex						
Female	-0.301	0.116	-2.594	0.010	0.740	(0.590, 0.929)
Size at birth						
Larger than average	0.685	0.514	1.332	0.184	1.984	(0.724, 5.435)
Average	1.058	0.502	2.108	0.035	2.881	(1.077, 7.704)
Smaller than average	1.363	0.509	2.676	0.008	3.908	(1.440, 10.601)
Very small	1.689	0.452	3.734	0.000	5.415	(2.231,13.143)
Mother's education						
Primary	-0.150	0.163	-0.919	0.358	0.861	(0.625, 1.185)
Secondary	-0.347	0.182	-1.907	0.057	0.706	(0.494, 1.010)
Higher	-0.761	0.331	-2.302	0.022	0.467	(0.244, 0.893)
Mother's BMI	0.000	0.000	-1.958	0.051	1.000	(0.999, 1.000)
Father's education						
Primary	-0.358	0.172	-2.088	0.037	0.699	(0.499, 0.978)
Secondary	0.024	0.189	0.129	0.897	1.025	(0.707, 1.485)
Higher	-0.577	0.359	-1.609	0.108	0.562	(0.278, 1.134)

Division of residence	-0.140	0.223	-0.629	0.530	0.869	(0.561, 1.346)
Chittagong						
Dhaka	-0.318	0.248	-1.285	0.199	0.727	(0.448, 1.182)
Khulna	-0.612	0.280	-2.187	0.029	0.542	(0.313, 0.938)
Rajshahi	-0.572	0.286	-1.996	0.046	0.565	(0.322, 0.990)
Rangpur	-0.207	0.270	-0.766	0.444	0.813	(0.480, 1.380)
Sylhet	0.086	0.220	0.392	0.696	1.090	(0.708, 1.678)
Wealth status						
Poorer	-0.590	0.227	-2.600	0.010	0.554	(0.355, 0.865)
Middle	-0.893	0.277	-3.227	0.001	0.410	(0.238, 0.704)
Richer	-0.509	0.246	-2.073	0.039	0.601	(0.372, 0.973)
Richest	-0.641	0.292	-2.195	0.029	0.527	(0.297, 0.934)
Rules from the tree						
Rule1	0.996	0.289	3.450	0.001	2.707	(1.537, 4.766)
Rule2	0.465	0.360	1.292	0.197	1.593	(0.786, 3.227)
Rule3	1.066	0.341	3.125	0.002	2.903	(1.488, 5.665)
Rule4	0.469	0.244	1.921	0.055	1.599	(0.990, 2.581)

Note: The table only reports variables whose estimates are significant at a maximum of 10% level of significant. Categorical variables are listed if estimates for at least one category is significant. The bottom panel of the table reports estimates and other statistics for dummy variables created from the four most significant rules generated in the classification tree (Figure 1). OR stands for odds ratio and CI stands for confidence interval.

Next, we convert these four rules into four dummy variables and include them in a logistic regression model alongside individual predictors. Such a dummy variable takes a value of one for individuals belonging to the associated rule, and zero otherwise. Table 1 contains results from the logistic regression.

Estimates for Rule1 and Rule3 are statistically significant at 1% level of significance and Rule4 is significant at 10%. Odds ratios associated with these rules are larger than odds ratios of all other induvial predictors, except size of children. This clearly indicates importance of including complex interactions defined by rules in studying undernutrition in children. The odds ratio of 2.7 associated with Rule1 implies that the odds of being stunted among children who are born in poorer families, are older than a year, are from Barisal, Chittagong or Sylhet division, and are born to lowly educated fathers is 2.7 times higher than the odds for children who do not belong to such strata. Odds for other rules have similar interpretations.

4. Discussion and Conclusion

The contribution of this paper is two-fold. First, it unfolds important interactions among possible sociodemographic drivers of childhood undernutrition providing intricate understanding into this long lasting public health issue in Bangladesh. Traditional parametric regressions which are widely used in similar empirical analyses are not well-suited to determine these interactions. This paper utilizes non-parametric classification trees to capture interactions in a data-driven manner. For example, we find that children's age, household wealth or parents' education do not determine undernutrition in isolation, but their combinations also contribute as important risk factors. The second and possibly more important contribution this study is that our analysis has more direct implications on policy recommendations than existing literature. In order to formulate an effective child nutrition policy it is crucial to identify socioeconomic features of children under highest risk of malnutrition and implement relevant interventions. We argue that unlike existing literature which has more generic findings, our analysis helps us determine such vulnerable groups. For example, a policy recommendation based on our findings will be to target relatively poor and lowly educated families living in Chittagong, Barisal or Sylhet and try increase education, employment and health awareness among them. These are the families with highest rate of stunted children and require the highest attention.

References

1. National Institute of Population Research and Training (NIPORT), Mitra and Associates, ICF International (2014). Bangladesh Demographic and Health Survey 2014. Dhaka, Bangladesh.
2. UNICEF East Asia & Pacific Regional Office (2003). Strategies to reduce maternal and child undernutrition. Health and Nutrition Working Paper.
3. Pelletier, D.L., Frongillo, E. A., Schroeder, D. G. & Habichit, J. P. (1995) The effects of malnutrition on child mortality in developing countries. *Bulletin of the World Health Organization,* **73**, 443–448.
4. Khanam, R., Nghiem, H. S. & Rahman, M. M. (2011) The impact of childhood malnutrition on schooling: evidence from Bangladesh. *Journal of Biosocial Science,* **43**, 437–451.
5. Gulati, J. K. (2010) Child Malnutrition: Trends and Issues. *Anthropologist*, **12**, 131–140.
6. Das, S. & Gulshan, J. (2017) Different forms of malnutrition among under five children in Bangladesh: A cross-sectional study on prevalence and determinants. *BMC Nutrition*, **3**, 1-12.
7. Hossain, M. B. & Khan, M. H. R. (2018) Role of parental education in reduction of prevalence of childhood undernutrition in Bangladesh. *Public Health Nutrition*, **21**, 1845–1854.

8. Islam, M. M., Alam, M., Tariquzaman, M., Kabir, M. A., Pervin, R. & Begum, M. (2013) Predictors of the number of under-five malnourished children in Bangladesh: application of the generalized Poisson regression model. *BMC Public Health*, 13-20.
9. Rahman, A. & Chowdhury, S. (2007) Determinants of Chronic Malnutrition among Preschool children in Bangladesh. *Journal of Biosocial Science*, **39**, 161–173.
10. Rahman, M. S., Howlader, T., Masud, M. S., & Rahman, M. L. (2016). Association of low-birth weight with malnutrition in children under five years in Bangladesh: Do mother's education, socio-economic status, and birth interval matter? *PLoS One*, **11**(6), e0157814.
11. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. CRC press.
12. Hothorn, T., Hornik, K. & Zeileis, A. (2006) Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.,* **15(3)**, 651–674.
13. Friedman, J. H. Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, **2(3)**, 916-954

Bivariate Archimedian Copula Model: An application to Customer Price Index (CPI) and Wholesale Price Index (WPI) in Indonesia

Muchtar Abdul Kholiq, Titi Kanti Lestari
BPS Statistics Indonesia

Abstract

Price stability is the main goal of every formulation of macroeconomic policies in the economy. In Indonesia, changes in prices are calculated using Customer Price Index (CPI) and Wholesale Price Index (WPI). In this study, Copula method is proposed to examine the dependencies between Customer Price Index (CPI) and Wholesale Price Index (WPI). Applications on monthly data on CPI and WPI for the period January 2013-October 2018 showed that there was a weak tail dependency based on the transformed plot, so the Copula was used. The correlation Kendall's tau produced correlation values τ = 094 and θ = 67.5.

Keywords

CPI; WPI; Copula Dependency

1. Introduction

In Indonesia, changes in prices are calculated using a variety of price indices. The price index that is often used to measure price changes is the Consumer Price Index (CPI), an index that measures the average price change in a period, from a collection of prices of goods and services consumed by residents / households in a certain period of time. The types of goods and services are grouped into 7 groups, namely food ingredients; processed food, beverages, cigarettes and tobacco; housing; clothing; health; education, recreation and sports; transport and communication. BPS Statistics Indonesia uses also published Wholesale Price Index (WPI) and Producer Price Index (PPI). It is possible that changes in one type of price have an impact on other prices. Therefore, the relationship between the Consumer Price Index (CPI) and the Large Trade Price Index (IHPB) needs to be assessed.

Copula is a multivariate function from a joint distribution. Copula is a tool that can be used to analyze the dependence of random variables in the structure described by the combined function.

This paper discusses the relationship between the Consumer Price Index and the Big Trade Price Index (IHPB). The Consumer Price Index (CPI) and the Big Trade Price Index (IHPB) are the relationships that will be studied using copula.

2. Bivariate Copula

Bivariate Copula is a distribution function of two random variables with the marginal distribution following the uniform distribution [0,1]. According to the Sklar theorem 1959, if H is a joint distribution function of random variables X and Y that have a function of marginal distribution F and G with domain R , then the joint distribution can be written as;

$$H(x, y) = C(F(x),G(y)) = C(u, v) \quad (1)$$

To simplify the joint distribution, it is assumed that two marginal distributions of F and G continue to make this C unique and can be formulated as follows:

$$C(\mathrm{u,v}) = \int_0^u \int_0^v c(\mathrm{u,v})\mathrm{du}, \mathrm{dv} \quad (2)$$

Where $C(\mathrm{u,v})$ is copula density function with α, $\beta > 0$, $x \geq 0$.

a. Gaussian Copula

The Gaussian Copula is symmetric, it does not have tail dependencies, with copula density functions as

$$\mathrm{c(u,v;\rho)} = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2(\mathrm{x}^2+\mathrm{y}^2)-2\rho\mathrm{xy}}{2(1-\rho^2)}\right. \quad (3)$$

Where $x = \Phi^{-1}(u)$ dan $y = \Phi^{-1}(v)$, Φ^{-1} denotes the inverse of the cumulative distribution function of univariate normal standard distribution.

2.3 Clayton Copula

The Clayton copula is family of Archimedian Copula. it is asymmetric, has lower-tail dependence but no upper-tail dependence. Its copula function, generator function and density function are

$$\mathrm{C}(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{\frac{1}{\theta}} \quad (4)$$

$$\varphi(t) = \frac{t^{-\theta}-1}{\theta}, \theta \in (0, \infty) \quad (5)$$

$$c(u, v) = \frac{(1+\theta)}{(u,v)^{1+\theta}}(u^{-\theta} + v^{-\theta})^{-\frac{1}{\theta}-2} \quad (6)$$

2.4 Gumbel Copula

The Gumbel copula is used to model asymmetric dependence. Copula Gumbel is able to capture strong upper tail dependence strong and weak lower tail dependence. Bivariate Gumbel Copula is defined as:

$$C^{\mathrm{Gu}} = (u_1,u_2; \theta)\exp(\ -[(-lnu_1)^{\theta} + (-lnu_2)^{\theta}]^{1/\theta} \quad (7)$$

Its generator functions is defined as

$$\varphi(t) = (-lnt)^{\theta} \quad (8)$$

Where $\theta \in [1, \infty)$, then the parameter value θ is defined as

$$\theta = \frac{1}{1-t} \quad (9)$$

2.5 Frank Copula

The Frank copula (1979) is defined as:

$$C^{Fr}(u_1,u_2;\theta) = -\theta^{-1}n\left\{1\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{(e^{-\theta}-1)}\right. \quad (10)$$

With θ = (−∞, 0) ∪ (0, ∞) , then its generator function is defined as:

$$\varphi(t) = -[\frac{e^{-\theta u_1}-1}{(e^{-\theta}-1)} \quad (11)$$

$$\tau = 1 + 4[D_1(\theta) - 1]/\ \theta^2 \quad (12)$$

Where $D_1(\theta)$ is the Debye function of order 1, which is defined as

$$D_1(\theta) = \int_0^\theta t\theta^{-1}(e^t-1)^{-1}\,dt \quad (13)$$

3. Results and Discussion

The data used in this study are Consumer Price Index data (2010 = 100) taken from International Financial Statistics (IFS) and the Large Trade Price Index (2010 = 100) taken from the Central Statistics Agency. The data used is monthly series data from January 2013 to October 2018.

3.1. Measuring Depedence

Tabel.2. Correlation Test Between CPI and WPI

Statistic Test	Correlation Value	Z-Score	P-Value
Spearman's Rho	0.9985	8.29383	0
Kendall's tau	0.9793	11.9847	0

Table.2. Shows that the results of the correlation test between Customer Price Index (CPI) and Wholesale Price Index (WPI) using the Spearman's rho test and the Kendall's tau test obtained the correlation value of the two test are equal to 0.9985 and equal to 0.9793 with p-value is equal to zero, then the decision is to reject H_0. So it can be concluded that there is a strong correlation between the CPI and WPI variables.

3.2. Fitting Copula Model and Parameter Estimation.

3.2.1 Fitting Copula Model

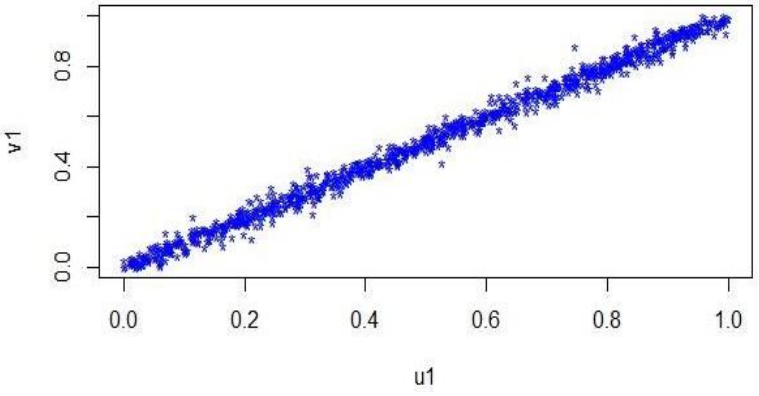


Figure.1. Scatter plot Copula

The figure shows the scatterplot the data after it is transformed to show the spread of points scattered uniformly and form a straight line. This indicates that there is no tail dependence in both the lower tail and tail upper so that Copula can explain the dependency relationship for data that does not have a tail dependence, namely Frank Copula.

3.2.2 Parameter Estimation

Table.3. Copula Model of CPI And WPI

Copula	Parameter	AIC	BIC
Gaussian	0.99	-277.24	-274.99
Clayton	16.33	-243.03	-240.78
Gumbel	13.33	-280.13	-277.88
Frank	67.51	-304.9	-302.65

Table.3 shows the results of processing CPI and WPI data that have been transformed into Uniform distributions, obtained by several copula models such as Gaussian, Clayton, Gumbel and Frank. The next step is to make a model selection. The selection of the copula model is done by looking at the smallest AIC and BIC values. Based on the Akaike Information Criterion (AIC) and Bayessian Information Criterrion values (BIC), copula model for CPI and WPI data is Copula Frank with parameter value $\theta = 67.51$, $\tau = 0.94$, pvalue = 0 is obtained.

4. Conclusion

Based on the results and discussion about the modeling of non Gaussian variable distribution dependencies using Copula along with their application in Customer Price Index (CPI) and Wholesale Price Index (WPI) in Indonesia, it can be concluded as the results of identification of the dependency structure between Customer Price Index (CPI) and Wholesale Price Index (WPI) Shows indicate the existence of dependencies. This is indicated by testing the type of Copula, namely Frank Copula, this type of Copula is characterized by Assymetric, it does not have tail dependency. This phenomenon illustrates that Customer Price Index and Wholesale Price Index(WPI) have a strong correlation.

References

1. **Akdia, Y., H. Berument and S. M. Cilasum**. 2005. *The relationship between different price indices: Evidence from Turkey*. Turkey. Retrieved from **http://berument.bilkent.edu.tr/ph04.pdf**
2. **Djehiche, B. and Hult H.** 2004. *An Introduction to Copula with Application. University of Copenhagen, Stockholm*

3. **Dowd, K.** 2008 . Copulas In Macroeconomics. Journal of International and Global Economic Studies,Vol1-26. Retrieved from https://pdfs.semanticscholar.org/2fa6/5e2f0d447e8ba8ee63a04939c43a0e72a40c.pdf
4. **Ivanov,E., A. Min, and F.Ramsauer**. 2017. *Copula-Based Factor Models for Mulltivaried Asset Returns*. Retrieved from www.mpdi.com/2225-1146/5/2/20/pdf
5. **Leskow, J. , J. Mokrzysztof and K. Krawiec.** 2011. *Modelling Stock Market Indexes with Copula Functions. Sucharskiego: University of Information Technology and Management Sucharskiego 2.*
6. **Meyer, C.** 2009. The Bivariate Normal Copula. Retrieved from *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.744.3738&rep=rep1&type=pdf*
7. **Oh, D. H.** 2014. *Copulas for High Dimensions: Models, Estimation, Inference, and Applications, Duke. Duke University*. Retrieved from https://higherlogicdownload.s3.amazonaws.com/AMSTAT/7328eb46-c57e-49af-8d4f-0460ca4a6a2c/UploadedImages/DongHwanOh_duke_Dissertation.pdf
8. **Patton, A. J.** 2007. *Copula-Based Models for Financial Time Series. United Kingdom: University of Oxford*. Retrieved from https://public.econ.duke.edu/~ap172/Patton_copula_handbook_19nov07.pdf
9. **Smith, M.S. and W. Maneesoonthorn.** 2017. *Inversion Copulas from Nonlinear State Space Models with Application to Inflation Forecasting. Melbourne Bussines Scool, Melbourne*. Retrieved from https://arxiv.org/pdf/1606.05022.pdf

Evidence based indicators for local educational monitoring -Detective work for the district Berlin-Mitte.

Ulrike Rockmann[1], Holger Leerhoff[2], Jeffrey Butler[3]
[1]Institute for School Quality Berlin-Brandenburg e.V.
[2]State Statistical Office Berlin-Brandenburg
[3]District Administration Berlin-Mitte

Abstract

As has been shown in PISA and PIAAC, the social background has an impact on educational outputs and outcomes in Germany. Efforts made in the last decades have improved the situation, but life circumstances still influence the opportunities of children and adults to develop their educational potentials, and performance gaps still remain. Considerations about developing educational policies to counteract this have focused up to now on educational institutions and kindergarten (early education and care (ECEC)) as well as on specific social groups of individuals. Especially since 2014/15 with a higher number of persons seeking asylum or recognition as refugees in Germany, in political discussions the indicator immigrant background is omnipresent and often Germany's immigration past since 1955 is widely neglected (Maaz et al. 2018b, figure 12). In the course of time the awareness also grew in non-scientific circles that, in addition to legal and constitutional necessities to identify foreigners, the value of the broad indicator immigrant background for explaining educational failure or success is very low, if at all existent. Therefore, the longstanding and on-going usage of the term in this regard can be seen in the light of lacking adequate indicator definitions and data to calculate them and represent an attempt at simplifying a complex situation. In consequence, the efforts to identify indicators that explain good or poor performance in the German educational system have been increased. In 2006 and 2016, education and immigration was the main topic in the central publication Education in Germany that is jointly commissioned by the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK) and the Federal Ministry of Education and Research. In 2017, the district government of Berlin-Mitte, a district that is home for 377.985 inhabitants with 179 nationalities, made a decision to initiate the project EduMitte to identify indicators for monitoring the educational processes on the local level covering the education time from ECEC until the end of secondary school [1]. Reasons for this initiative were the special integration policy of the district to discontinue using the term immigrant background in its planning activities due to its negative connotation and the fact that it is only a descriptive and not an explanatory characteristic [2]. Further, the fact that for many years about 11% of the

adolescents leave secondary school without a school certificate was seen as a clear political mandate to act (Rockmann, U., & Leerhoff, H. 2018, p. 18). First project results were presented in August 2018, showing for some specific family constellations that the immigrant background does not make a difference in kindergarten attendance years when controlling the results for the education of the family (ISCED level) and the languages(s) spoken at home.

Keywords
Immigrant background; indicator; education; monitoring; policy making

1. Introduction

In Germany the responsibility for primary and secondary school is highly decentralized. The KMK coordinates some broader general issues on national level. The federal states are fully responsible for providing an adequate budget for setting up sufficient primary and secondary schools, employing teachers etc. Due to the two-level administration within the Federal State Berlin the organization is further decentralized to its 12 districts, Berlin-Mitte being one of them. Within the Berlin-wide educational framework districts are able to make their own methodological decisions, for example concerning instruments used for evaluating language competences. Furthermore, schools have freedom of decision with regard to education related procedures and methods. All in all, it is no easy task to keep track of everything and comparisons between federal states and even between the districts of Berlin are challenging and bear the risk of comparing apples with oranges.

Germany has no nationwide educational register with data like participation time in educational programs, examination certificates, educational biographies etc. Therefore, the educational research community either uses data from large scale assessments, data collected by the National Education Panel Survey project (NEPS) or the Official Statistics Microcensus (M-Census). All these data sources have a limitation due to their sample size and design: it is only possible to obtain data on a national, a federal state and in rare cases a district level but not for lower local levels. Therefore, these data only provide a general framework but are not sufficient for local analysis' in Berlin-Mitte, because the social situation in this district with 378.000 inhabitants is far too heterogeneous.

For years, about 11% of the adolescents leave secondary schools in Berlin-Mitte without a school certificate – especially those with a foreign citizenship. Data from official statistics show that these adolescents have nearly no chance to enter a vocational training program and, in consequence, have a high probability to end up in the social transfer system. In 2016, 95% of the adults without a school certificate also did not complete any type of occupational training (Rockmann, U., & Leerhoff, H.2018, p. 16). The above-average rate of

unsuccessful school leavers together with the high number of refugees entering schools without sufficient knowledge of the German language were two inspirations for the project EduMitte initiated and supported by the district government. Besides identifying educational and organizational barriers and finding valid quantitative monitoring indicators, the project has the further goal of establishing sustainable administrative monitoring routines without too much additional effort. Saying that, the first project step was a review of the existing data infrastructure.

2. Methodology

Target population: The target populations of the project are children and adolescents younger than 18 years old living in the district Berlin-Mitte who attend kindergarten, primary or secondary school. Since it is not controversial that the family plays an important role in their children's education, the families belong to the target population as well. In 2016, about one third of the population in the district Berlin-Mitte lived in families (households with adults and (their) minor children).

Data sources: Data used for this first project step are:

- Administrative data: The Berlin population register and the Berlin-Mitte school enrollment health survey data (ESU-S).
- Official Statistics data: the Microcensus (1% population sample) and the Child and Youth
  Welfare Statistics (CYW-S), with a total coverage of all children attending official ECEC programs in institutions
- Survey conducted by the project: EduMitte-Q1questioning all parents of the Berlin-Mitte school enrollment cohort 2019/2020.

1st project step: Crucial factors in the educational biographies are the transitions between educational institutions. The project starts with the first transition from Early Childhood Education and Care (ECEC) to primary school. Thus, the target population of the 1st step are families with children less than 6 years old before entering primary school (about 3.500 children each year).

Input indicator(s) 1st project step: The individual input indicator are the years the children spent in ECEC. Since visiting the ECEC-institutions is not obligatory in Germany, the years spent in ECEC 01 and 02 programs vary. Most of the parents book full-time participation (Monday to Friday, 7–9 hours per day). The availability of ECEC places is not an intervening factor although the situation in Berlin is quite tense. It was not possible to include further input indicators like the quality of ECEC programs or the child-staff ratio, since

information as to which kindergarten the children visited was not available for the project.

Output indicator 1st project step: The only available output indicator is the result of the standardized language test conducted during the school enrollment health examination (ESU-S) about 10 months before entering primary school (Bettge, S., & Oberwöhrmann, S., 2017). The enrollment cohort has to participate in this examination being the basis for the school readiness decision. The language test results are assigned to four categories – no German language knowledge, single German words, speaking fluently with major mistakes, good / very good German language skills.

Immigrant background: In general, the indicator is defined in slightly different ways due to the national situation (OECD 2018, p. 17, chap. 1.1.1). Because of the strict data-protection laws in Germany, only the citizenship was available for determining the immigrant status for a long time and still many administrative data sources – like the school statistics – do not include a wider approach. Since 2005, the German Microcensus collects more detailed data including the date of immigration to Germany. A person has an immigrant background if the person itself or one of the parents does not have the German citizenship by birth [3]. Therefore, native-born children to parents with only foreign citizenships are Germans with an immigrant background. The status of a person depends on its own status and on the status of the parents and grandparents (1st, 2nd and 3rd generation). On this high aggregation level three groups are differentiated – German citizenship without immigrant background, Germans with an immigrant background and foreigners. Although the definitions of the Berlin administrative population register and the Microcensus are not harmonized, the population register allows widely to approximate the Microcensus definition. By doing this, more detailed local information about the population becomes available. Also, the ESU-S collects detailed information: place of birth of the child and the parents, the date of immigration to Germany, the citizenship of the parents. Further, the CYW-S uses another definition: available is the dichotomous information about the origin of the parents (German / not German) and the predominantly spoken language at home (German / not German).

German language: For integration and participation in the society it is of central interest whether individuals have sufficient German language skills. For a very long time only the CYW-Statistics collected rudimentary language information. Starting in 2017, the Microcensus respondents report the predominantly spoken language in the household, having 8 languages and 3 language groups for selection available. Unfortunately, the questionnaire does not consider that multi-lingual households are not singular cases anymore.

This and the assumption that respondents with an immigrant background might anticipate the socially desired answer could lead to an overestimation of German speaking households. Nevertheless, the Microcensus data give a broad idea about the situation. 2017s data show that in Germany as well as in Berlin 87% of the households predominantly speak German [4]. The figures for Berlin's districts range from 96% in the district of Treptow-Köpenick to 74% in Berlin-Mitte – clearly pointing to the special situation in Berlin-Mitte.

The more important local data source for the 1st project step is the ESU-S. As a standard since 2016, parents are asked to name all languages spoken at home. In 2018, 21 languages were reported – besides German, mainly Turkish, Arabic, English, Russian and Polish. Unfortunately, the specific language constellation within the family is not enquired. Since children have different models in German speaking parents or German speaking siblings, the additional survey EduMitte-Q1 should clarify the details.

Social background: Again, the Mircocensus provides information about the social background of the family (education, income, employment etc.) on a national, a federal state and a Berlin district level. The educational risks for children related to the social situation of their families and the lack of sufficient educational resources are described by three standard indicators and their combinations initially derived from Pierre Bourdieu's work on culture and cultural capital (1983; Maaz et al., 2018,Chap. A4, Rockmann et al. 2014).

- EduRisk: a low formal educational level of the parents, both less than ISECD 3
- PovRisk: family income under the poverty threshold
- EmpRisk: both parents unemployed

Due to the 1%-sample rate, the Microcensus is not suitable for analyzing the district Berlin-Mitte in detail. By using the data from the ESU-S it is possible to calculate the indicators EduRisk and EmpRisk for school enrollment cohort.

3. Results

Official CYW-Statistic shows that children in Berlin-Mitte start with ECEC programs quite early. 4% of the children younger than one year old attend ECEC. This rate is nearly constant since introduction of paid parental leave for 12 to 14 months after the child's birth. At the age of one year, 55% of the children are in ECEC and at the age of two years 87% (Rockmann, U., & Leerhoff, H. 2018, p. 42). Taking into account the language spoken at home, differences become obvious (fig. 1). Whereas 72% of children in Berlin-Mitte speaking mainly German at home started with ECEC before attaining the age of 2 years, the rate for children living in not-German speaking households is 43%.

The interpretation of this finding remains fragmentary: CYW-S does not deliver data about the place of birth of the child – therefore the question is unanswered as to whether families who do not speak German at home sent their children voluntarily later to ECEC or the child in question immigrated at an older age. If the general assumption is that it is beneficial for the child's development to start ECEC early, then it is of great interest to describe the group that starts later and try to identify reasons and possible barriers.

The school enrollment cohort 2018/19 started attending ECEC at the age of 2.3 years, native-born children with 2.0 years. The ECEC attendance years at the time of the ESU-S are on the average 3.8 years. Figure 2 illustrates that in case of a high family education level (ISCED 5 or more) ECEC years do not differ for children with two native-born parents or one parent being born abroad speaking only German or German and foreign languages at home. For the children living in families with an EduRisk (ISCED 0-2) neither the language(s) spoken at home nor the immigrant status of the parents make a difference and the attendance time is significantly less than for the ISCED-5-peer group.

Figure 1: ECEC starting age by language spoken at home and region 2017 (in % of children attending ECEC)

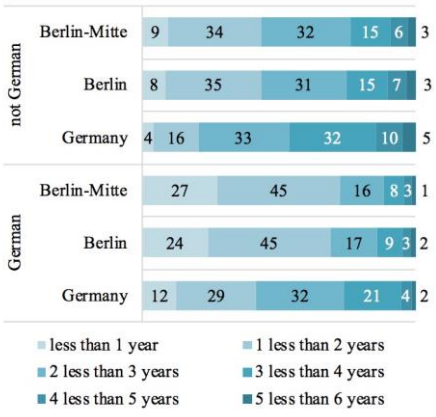
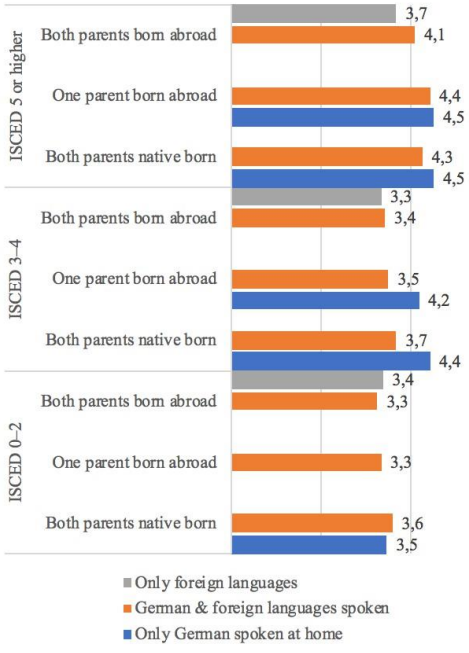


Figure 2: Attedance years (median) in ECEC of native born children at the time of ESU-S by highest ISCED level in the family and place of birth of the parents (2018)



Source (figure 1): Federal Statistical Office and statistical offices of the federal states, Child and Youth Welfare statistics, own calculations, reference date 1.3.2017
Source (figure 2): Child and Youth Health Services Berlin-Mitte, School enrolment health survey, school enrolment cohort school year 2018/19, own calculations

207 native-born children living with parents who were both born abroad and only speaking foreign languages start with ECEC later than average – at

the age of 2.5 years. Children living in Arabic, Turkish and Kurdish speaking households start at the age of 2.8 years. For this group the ISCED level and the employment status of the mother have an impact: In the case of a highly educated family (ISCED 5 and higher) and an employed mother, the children start attending at the average-age of 2.0 years.

The correlation between the years spent in ECEC and the language skills at ESU-S is well documented over many years (fig. 3). In 2018, 85% of the children attending 3.5 years and more reach the highest language level (good /very good). For native-born children the percentage is 86% – for children born abroad 66%. In regard to children attending ECEC between 2.0 and 3.5 years, the percentage of high-performers drops by 20 percentage points.

Analyzing the language results according to the family's immigrant status and ISCED level (fig. 4) shows patterns similar to the ECEC attendance years in figure 2. Like before, children with highly educated parents of which at least one is native-born, and either German or German and foreign languages are spoken in the household, are the best performers. Children living in families speaking foreign languages and having an EduRisk have more difficulties to reach a sufficient language competence.

Figure 3: Language test results ESU-S 2018 by place of birth and ECEC participation years (in %)

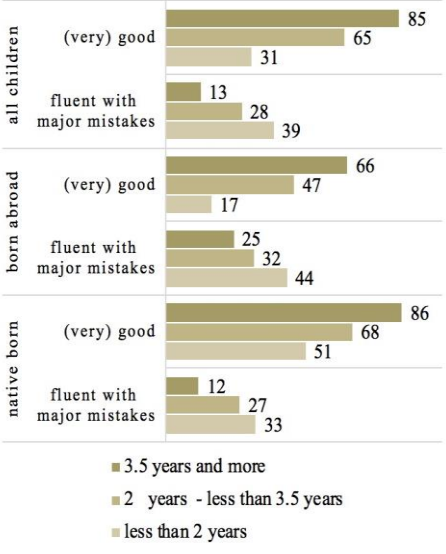
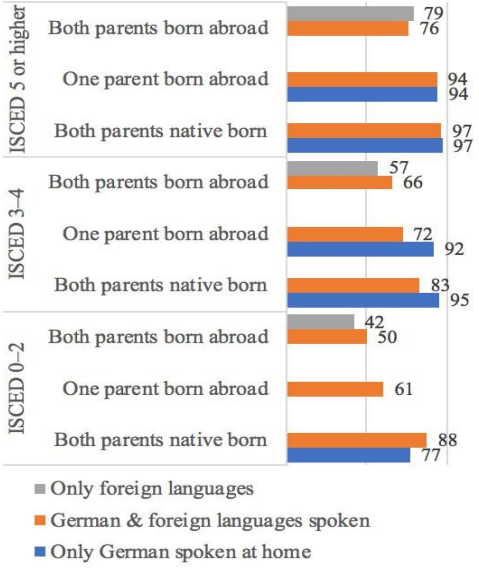


Figure 4: Language test results (very) good in ESU-S 2018 of native-born children by highest ISCED level in the family and parents' place of birth (in %)



Source: Child and Youth Health Department Berlin-Mitte, School enrollment health survey, school enrollment cohort school year 2018/19, own calculations

4. Discussion and Conclusion

The initial results obtained by analyzing the data available from administrative sources and official statistics give some indications as to whom local policies should be addressed. Although it was possible in some cases to replace the broad indicator immigrant background by the ISCED level, the spoken language and employment status, other calculations still show the indicator significant. Further, findings for special foreign languages give hints that experiences with the educational system in the former home countries are perhaps transferred to the new home country without further evaluation and therefore have an impact on the parents' decisions. The currently on-going survey EduMitte-Q1 tries to clarify the validity of this assumption and could therefore help to establish a new approach in addressing policies. Furthermore, the EduMitte-Q1 collects data concerning the educational aspirations of the parents, their knowledge of the German educational system, and the educational resources the children have at home.

References

1. Bettge, Susanne & Oberwöhrmann, Sylke (2017). Grundauswertung der Einschulungsuntersuchungsdaten in Berlin 2017.
2. Bourdieu Pierre (1983) Ökonomisches Kapital, kulturelles Kapital, soziales Kapital. In: Bauer U., Bittlingmayer U.H., Scherr A. (Eds.) (2012) Handbuch Bildungs- und Erziehungssoziologie. Bildung und Gesellschaft. VS Verlag für Sozialwissenschaften, Wiesbaden. DOI: https://doi.org/10.1007/978-3-531-18944-4_15
3. OECD/EU (2018), Settling In 2018: Indicators of Immigrant Integration, OECD Publishing, Paris/European Union, Brussels. https://doi.org/10.1787/9789264307216-en
4. Maaz, Kai, Baethge, Martin, Brugger, Pia, Rauschenbach, Thomas, Rockmann, Ulrike, Seeber, Susann & Wolter, Andrä (2018): Bildung in Deutschland – Ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung. ISBN 978-3-7639-5964-8.
5. Maaz, Kai, Baethge, Martin, Füssel, Hans-Peter, Hetmeier, H.-W., Rauschenbach, Thomas, Rockmann, Ulrike, Seeber, Susann & Wolter, Andrä (2016): Education in Germany – An indicator-based report including an analysis of education and migration, 2016, ISBN 978-3-76395742-2; DOI: 10.3278/6001820ew
6. Rockmann, Ulrike & Leerhoff, Holger (2018). Bildungsmonitoring in Berlin-Mitte. Bildungszugänge und Bildungsübergänge von Kindern im Alter von 0 bis 18 Jahren im Bezirk Berlin-Mitte. 1. Projektbericht. ISBN 978-3-9809139-9-7

7. Rockmann, Ulrike, Rehkämper, Klaus & Leerhoff, Holger (2014). Bildungskapital verringert Bildungsrisiken. In DIJ Impulse, Das Bulletin des Deutschen Jugendinstituts 3/2014, Nr. 107, S. 26-29, ISSN2192-9335

Hyperlinks

[1]: https://www.berlin.de/ba-mitte/politik-und-verwaltung/beauftragte/integration/bildungsmonitoring/; accessed 27.12.2018

[2]: https://www.berlin.de/ba-mitte/politik-und-verwaltung/gremien/migrationsbeirat/beschluss_mmmh_141119.pdf; accessed 26.12.2018

[3]: https://www.destatis.de/EN/FactsFigures/SocietyState/Population/MigrationIntegration/Methods/Migr ationBackground.html, Fachserie 1, Reihe 2.2; accessed 26.12.2018.

[4]: https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/MigrationIntegration/Migrations hintergrund.html; accessed 27.12.2018.

Optimal sample size allocation: An investigation for evaluating the behavior of a dietetic supplement

Gajendra K. Vishwakarma[1], Carlos N. Bouza-Herrera[2]
[1]Department of Applied Mathematics, Indian Institute of Technology Dhanbad, India
[2]Departamento de Matemática Aplicada, Universidad de La Habana, Cuba

Abstract
In many medical research, it is needed the determination of optimal sample size allocation in heterogeneous population. In this article, we discussed an approach for optimal sample size allocation in three investigations on the response in terms of Body Mass Index to a dietetic supplement for persons with Diabetes Mellitus, HIV/AIDS and Cancer post-operatory recovery. The algorithms for different numerical methods have been used to analyzed and evaluated in terms of computing time, number of iterations and gain in accuracy using stratification. Effects of using the algorithms for optimal allocation also presented.

Keywords
Optimal sample size allocation; Body Mass Index; Diabetes; HIV/AIDS; Cancer recovery

1. Introduction

Many investigations in medicine uses stratification for guarantying having an adequate representation of the population. In stratified sampling the main aim is obtaining estimators of the population parameters for the variable of interest with the minimum precision at a minimum cost. They both depend on the sample size. Therefore, obtaining optimal sample sizes is one of the key element of the theory. This paper deals with the development of a study of the efficacy of using mathematical programming methods for deriving optimal sample sizes for stratification sampling. Their behavior was evaluated in a research on the effect of a new dietetic medicament in the Body Mass Index (BMI) of persons suffering of diabetes mellitus, HIV/AIDS and with post chirurgic cancer treatments. See Bays et al. (2007), Koethe et al. (2011), Greenlee et al. (2016) and Temple (2013) for discussions on BMI and supplementary alimentation issues. The stratified sampling efficiency depends on several factors, but allocating optimally the sample sizes to different strata is one of the more important issues. Minimizing the variance of the estimator of the population mean, when dealing with stratification sampling, was the objective of the works of Neyman (1934). Neyman named the problem "optimal allocation". The methods of Langrage multipliers were used for

solving the problem of minimizing the variance for a fixed cost (budget). That solution may be improved by using optimization tools. Nowadays, the development of computing tools has supported the possibility of obtained a better allocation of sample sizes result using them.

2. Methodology

It is usual that stratification is used in statistical survey sampling studies. Strata are unknown in advance to the decision maker (DM). A population $U = \{u_{1,\dots,}u_N\}$ is to be analyzed. It is partitioned into *H* mutually disjoint clusters $U = \{u_{1,\dots,}u_H\}$. A stratum is considered as a cluster of population units such that the variable of interest *Y*, evaluated in each unit u_i in U_k, is close to the conditional expectation $u_k = E[Y/U_k]$. Ideally, the strata means should be very different. This fact sustains that the estimation, based on stratified random sampling (SSRS) is very accurate when compared with simple random sampling (SRS), see Singh (2003). One of the key problem in stratified random sampling is to derive the strata sample sizes optimally once the overall sample size is fixed.

A sufficiently broad optimization model is

$$\min_{\vec{n}} f(\vec{n}) = \sum_{h=1}^{H} \frac{\delta_h^2}{n_h} \qquad (1)$$

Subject to $\vec{n}^T \vec{e}$ - $\gamma = g(\vec{n}) \leq 0; g: \Re_+^H \longrightarrow \Re;$

$$\vec{n} \in \left\{\vec{b} = (b_1, \dots, b_H)^T \in \Re_+^H \middle| m_h \leq b_h \leq M_h, h = 1, \dots, H\right\}$$

The total of the variable Y

$$\mathrm{T} = \sum_{h=1}^{H} \sum_{j=1}^{N_h} Y_{hj} = \sum_{h=1}^{H} N_h \mu_n \qquad (2)$$

is commonly the parameter of interest. An unbiased estimator of T is given, when SSRS is used, by

$$\hat{T} = \sum_{h=1}^{H} \frac{W_h}{n_h} \sum_{j=1}^{n_h} y_{hj} = \sum_{h=1}^{H} W_h \bar{y}_h \; ; W_h = \frac{N_h}{N}, h = 1, \dots, H \qquad (3)$$

As the means are independent the sampling error is

$$V(\hat{T}) = \sum_{h=1}^{H} W_h^2 V(\bar{y}_h) \qquad (4)$$

Minimizing it was the objective of the works of Neyman (1934). The problem is named "optimal allocation" and it deals with the solution of

$$\min\left\{V(\hat{T}) = \textstyle\sum_{h=1}^{H} W_h^2 \, V(\bar{y}_h) \middle| (c_0 + \textstyle\sum_{h=1}^{H} c_h n_h)\right\} \qquad (5)$$

$$C = c_0 + \textstyle\sum_{h=1}^{H} c_h n_h$$

c_0: overall costs of the survey, c_h: cost of evaluating a unit in U_h, C: the budget assigned to survey.

The subsequent optimization problem is determined. The problem, as tackled in textbooks, considers solving it using Lagrange multipliers. The general optimization problem is

$$min_{\vec{n}=(n_1,\dots,n_H)^T}\left\{\sum_{h=1}^{H} W_h^2 V(\bar{y}_h) + \lambda\left(C-\left(c_0+\sum_{h=1}^{H} c_h\, n_h\right)\right)\right\}; V(\bar{y}_h)=\frac{\delta_h^2}{n_h}, h=1,\dots,H \quad (6)$$

The solution of (6) is discussed by Singh (2003) as

$$n_{(LC)h} = (C-c_0)\left\{\left(W_h\delta_h/\sqrt{c_h}\right)\Big/\left(\sum_{h=1}^{h=1} W_h\delta_h\Big/\sqrt{c_h}\right)\right\} \quad (7)$$

When the overall sample size n is fixed it may be written as

$$n_{(LC)h} = n\left\{\left(W_h\delta_h/\sqrt{c_h}\right)\Big/\left(\sum_{h=1}^{H} W_h\delta_h\,/\sqrt{c_h}\right)\right\} \quad (8)$$

In a more general context, take γ as an upper constraint. Using the Lagrangian approach is derived, under general conditions, as solution:

$$n_h = \gamma\left(W_h\sqrt{V(\bar{y}_h)}\Big/\sum_{h=1}^{H} W_h\sqrt{V(\bar{y}_h)}\right) \quad (9)$$

3. Mathematical programming and optimal allocation

1. Optimal allocations

Though optimal allocations minimize the variance of the estimator of interest, in practice some troubles may occur. For example if we use sampling without replacement the optimal allocation may fix some stratum sample sizes n_h exceeding N_h. In the other extreme case, we may obtain too small sample sizes for some strata. Thus, is desirable for the decision maker (DM) having some control for assuring adequate sample sizes or sampling fractions.

The variation of the design weights $W_h; h=1,..,H$, is measured by the Gelman-bound of the variability of the design weights

$$Max_{k=1,\dots H;k=h,h}\left(N_h n_h \,/\, N_j n_j\right) \quad (10)$$

Practitioners may analyze it when looking for adequate sample sizes for small area, see Burgard and Münnich (2012). The DM considers that the total sample size must satisfy

$$\sum_{h=1}^{H} n_h \le \gamma \quad (11)$$

Rewriting the model given in (1) in other form then we have the optimization problem as

$$\min_{(n_1,\ldots,n_H)^T}\left\{\sum_{h=1}^{H}\frac{\delta_h^2}{n_h}\right\} \tag{12}$$

subject to $\sum_{h=1}^{H} n_h \leq \gamma$; $n_h \in [m_h, M_h]$ and $h = 1, \ldots H$, where $(n_1, \ldots n_H)^T \in \Re_+^H$

Take $\vec{n}_0$ as the solution of (12). The sample survey theory recommends the use of the optimal sample sizes, when using stratified sampling, because the variances of $\hat{T}$ are considerably smaller than the sampling error generated by samples of size n for estimating T.

Consider that n is fixed and that n_h is proportional to $W_h\delta_h$. A general expression of the variance of $\hat{T}$ is

$$V\left(\hat{T}|n_h \propto W_h\delta_h; h = 1, \ldots, H\right) = V\left(\frac{N}{n}\sum_{i=1}^{n} Y_i\right) - \frac{N^2}{n}\left\{\sum_{h=1}^{H} W_h\left(\delta_h - \bar{\delta}\right)^2 + \sum_{h=1}^{H} W_h(\mu_h - \mu)^2\right\} \tag{13}$$

where $\bar{\delta} = \sum_{h=1}^{H} W_h\delta_h$

The use of $\vec{n}_0$ should provide smaller values of the variance. We will study this problem using a numerical experiment.

2. Some approaches for solving to optimal allocation problem

The main goal when using Nonlinear equation for Lagrange multiplier λ is expressing the strata sample size n_h as a function depending on it. The resulting function is continuous but not necessarily differentiable. Hence we will deal with the only basic root obtained.

The first Karush–Kuhn–Tucker (KKT) condition of the problem is equivalent to

$$0 \geq -\frac{\delta_h^2}{M_h^2} + \lambda^*; \text{ if } n_h^* = M_h, 0 = -\frac{\delta_h^2}{M_h^2} + \lambda^* \text{ if } n_h^* \in [m_h, M_h], 0 \leq -\frac{\delta_h^2}{M_h^2} + \lambda^*; \text{if } n_h^* = M_h.$$

The inequality constraint holds with equality only in the optimal solution. Then, the second condition of the KKT–Tucker is:

$$\vec{n}^{*T}\vec{e} - y = 0 \tag{14}$$

Let us consider that λ is a variable and that $\vec{n}$ depends of λ. We may set that

$$n_h(\lambda) = \left\{ \begin{array}{l} M_h \; if \; \lambda \leq \frac{\delta_h^2}{M_h^2} \\ \sqrt{\frac{\delta_h^2}{\lambda}} \; if \; \lambda \in M^* = \left[\frac{\delta_h^2}{M_h^2}, \frac{\delta_h^2}{m_h^2}\right]; h = 1, \ldots, H \\ m_h \; if \; \lambda \leq \frac{\delta_h^2}{m_h^2} \end{array} \right\}$$

From Theorem-1 of Münnich et al. (2011) we have that there is an unique solution of the optimization problem given in (1), if and only if, there exists a multiplier $\lambda^* \in \Re_+$ such that $g(\vec{n}(\lambda^*)) = 0$. Hence, solving the optimization problem is equivalent to solving $\vec{n}(\lambda)^T \vec{e} - y = 0$. As $g(\vec{n}(\lambda^*))$ is continuous but not differentiable, for solving the equation we must rely on methods which only require continuity.

The solution of the optimization problem can also be found through using a fixed point iteration. Interested in this issue may see large discussions in Berinde (2007). In this case the components $n_h(\lambda)$ have as value either M_h or m_h or their value lies in the interval $M^* = [m_h, M_h]$. Accordingly, we have a partition of the set J={1,...,H} of indices into three subsets:

$$J = \{J_{M_h} \cup J_{m_h} \cup J_{M^*}\} = \left\{\left(h \in J \middle| \lambda \leq \frac{\delta_h^2}{M_h^2}\right) \cup \left(h \in J \middle| \lambda \leq \frac{\delta_h^2}{m_h^2}\right) \cup \left(h \in J \middle| \frac{\delta_h^2}{M_h^2} < \lambda < \frac{\delta_h^2}{m_h^2}\right)\right\} \quad (15)$$

Using these sets, the equation (14) is rewritten as

$$\vec{n}(\lambda)^T \vec{e} - y = \sum_{h \in J_{M_h}} M_h + \sum_{h \in J_{m_h}} m_h + \sum_{h \in J_{M^*}} \sqrt{(\delta_h^2/\lambda)} - Y = 0 \quad (16)$$

If we solve this equation for λ, we obtain

$$\lambda = \left[\sum_{h \in J_{M^*}} \delta_h \Big/ \left(Y - \left(\sum_{h \in J_{M_h}} M_h + \sum_{h \in J_{m_h}} m_h\right)\right)\right]^2 \quad (17)$$

Solving $\vec{n}(\lambda)^T \vec{e} - y = 0$, g(λ)=0 is equivalent to solving this new equation, see Münnich et al (2012). Note that this function is discontinuous and has jumps.

Considering the fixed point λ^* is to be assumed that in J_{M_h}, J_{m_h} the inequalities are strict inequalities. Then it is conceivable that small perturbations of λ^* do not change.

$$\lambda^* = \left[\sum_{h \in J_{M^*}} \delta_h \Big/ \left(Y - \left(\sum_{h \in J_{M_h}} M_h + \sum_{h \in J_{m_h}} m_h\right)\right)\right]^2 \quad (18)$$

This means that the optimal λ^* is already delivered as an output when $\lambda \cong \lambda^*$. If the fixed point iteration converges, then the iteration terminates, after a finite number of steps, with the exact solution λ^*, because the index

sets do not change anymore. An iteration algorithm for computing the optimal allocation is described as follows:

$Step\ 1$:Input $\lambda(0)=\frac{1}{y}\sum_{h=1}^{H}\delta_h, t=0$

$Step\ 2$:t=t+1

$Step\ 3$: Calculate$J_{M_h}, J_{m_h}, J_{m^*}$

$Step\ 4$: λ(t+1)$=\left[\frac{\sum_{h\in J_{M^*}}\delta_h}{y-\left(\sum_{h\in J_{M_h}}M_h+\sum_{h\in J_{m_h}}m_h\right)}\right]^2$

$Step\ 5$: $while\ \lambda$(t+1)# λ(t) else $\vec{n}(\lambda(\text{t+1}))$ is the optimal allocation

All the arguments can also be applied to the following optimization problem:

$$\min_{(n_1,\ldots,n_H)^T}\left\{\sum_{h=1}^{H}\frac{\delta_h^2}{n_h}\right\} \tag{19}$$

subject to $\vec{n}^{*T}\vec{p}-y\leq 0$ where $(n_1,\ldots,n_H)^T\in\Re_+^H$,

where $\vec{p}$ defines a vector of penalty, cost or weighting parameters. Then the solution of the problem without box constraints can be specified as

$$n_h^*=\left(y/\sum_{h=1}^{H}\delta_h^2\sqrt{P_h}\right)(\delta_h/P_h)$$

Hence, we have in such cases that

$$n_h^*(\lambda)=\left\{\begin{array}{l} M_h\ if\ \lambda\leq\frac{\delta_h^2}{M_{hph}^2} \\ \sqrt{\frac{\delta_h^2}{\lambda}}\ if\ \lambda\,\epsilon\, M^*=\left[\frac{\delta_h^2}{M_h^2P_h},\frac{\delta_h^2}{m_h^2}\right]; h=1,\ldots,H \\ m_h\ if\ \lambda\leq\frac{\delta_h^2}{m_{hph}^2}\end{array}\right\}, \lambda(\text{t+1})=\left[\frac{\sum_{h\in J_{M^*((t))}}\frac{\delta_h}{P_h}}{y-\left(\sum_{h\in J_{M_h}}M_h+\sum_{h\in J_{m_h}}m_h\right)}\right]^2.$$

4. Numerical expriments

We will consider the use of simple random sampling for large population sizes. In such cases the distinction between with replacement and without replacement is no-important. Note that for SRS with replacement (SRSWR) $\sum_{h=1}^{H}\frac{\delta_h^2}{n_h}=\sum_{h}^{H}W_h^2\frac{\sigma_h^2}{n_h}$ and for SRSWOR $\sum_{h=1}^{H}\frac{\delta_h^2}{n_h}=\sum_{h}^{H}\left(\frac{N_h-n_h}{N_h-1}\right)W_h^2\frac{\sigma_h^2}{n_h}$,

hence for large population sizes $\left(\frac{N_h-n_h}{N_h-1}\right)\to 1$ and $\sum_h^H W_h^2\frac{\sigma_h^2}{n_h}\cong\sum_h^H\left(\frac{N_h-n_h}{N_h-1}\right)W_h^2\frac{\sigma_h^2}{n_h}$.

Accepting the approximations we have, once we determine an optimal allocation the minimum variance is given by

$$V^*\cong\sum_{h}^{H}W_h^2\frac{\sigma_h^2}{n_h^*}. \tag{20}$$

The sampling error of the corresponding SRS design is equal to

$$V_{SRS} = V\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \cong \frac{1}{Nn}\sum_{h=1}^{H}\sum_{j\varepsilon U_h}^{H}\left(Y_{hj} - \mu\right)^2 = \sum_{h=1}^{H} W_h \frac{\sigma_h^2}{n} + \frac{1}{n}\sum_{h=1}^{H} W_h(\mu_h - \mu)^2 \quad (21)$$

Comparing the variances, we may compute

$$\Delta = V_{SRS} - V^* = \sum_{h=1}^{H} W_h \sigma_h^2 \left(\frac{1}{n} - \frac{N_h}{Nn_h^*}\right) - \frac{1}{n}\sum_{h=1}^{H} W_h\left(\mu_h - \mu\right)^2 \quad (22)$$

The performance of the optimal allocation it may be measured by computing the relative measure:

$$\Delta_R = 1 - (V^*/V_{SRS}) \quad (23)$$

Large values of $\Delta_R \in [0,1]$ indicates large increases in the accuracy of the estimates due to using optimal allocation.

Body mass index (BMI) is frequently used as a reliable indicator of the health of the patients. BMI is a simple index calculated from a person's weight and height, and may be used to screen the behavior of some diseases. It is one of the factors that can be of use to assess the status of health. It serves as a quick screening tool. In particular, a low range signals that a patient is malnourished due to the improperly absorption of nutrients. The dietetic supplement to be launched should stabilize the BMI of a patient without having side effects.The selected patients were to be weighted repeatedly. to the nearest 100 grams using electronic scales after removing shoes and bulky clothing. Height was measured to the nearest millimeter using a portable stadiometer.

The numerical experiments will be conducted for establishing the magnitude of Δ_R for each population. Also were considered the time in seconds needed for obtaining the optimal allocation and the number of iterations for each algorithm. The details of data sets are given below:

1. The diabetics controlled by public medical institutions of was the population under study. Its size was N=274 349 persons. The BMI of them were measured when they started to be controlled. They were classified into 8 strata in terms of age and sex. The sample sizes. The sample size was n=2 700.
2. The person with HIV/AIDS of a national program of attention conformed a population of N=51 236 persons. Taking into account age, sex and sexual preference were defined 12 groups. The sample size fixed was n= 5 000. The BMI obtained in the last visited allowed computing the needed parameters.
3. The persons with a chirurgical operation in the last 5 years assisting periodically to the control consults were classified, considering where the tumor was present the sex and the age. N=931 945 persons and n=10 000. The number of strata was determined as 30.

The strata sample sizes were determined. They did not differ too much but the gain in accuracy with respect to SRSWR was largely different. Table 1 presents the derived results for each research by the different algorithms

For diabetics the Secant algorithm is to be recommended as it gains in accuracy is the largest and was the faster. For HIV/AIDS Regula-Falsi illinois was the best alternative and for cancer patients Regula-Falsi versions had the best results, but none of them was preferably in all the cases. In any case note that the times are similar for the secant related algorithms.

Table 1: Effect of using the algorithms for optimal allocation

Method	**DIABETICS**			**HIV/AIDS**			**CANCER PATIENTS**		
	Time in seconds	Number of Iterations	Δ_R	Time in seconds	Number of Iterations	Δ_R	Time in seconds	Number of Iterations	Δ_R
isection	10.15	18	0.19	25.09	21	0.11	11.27	25	0.12
Secant	10.20	5	0.41	10.20	6	0.19	10.25	5	0.23
Regula-Falsi normal	10.21	7	0.33	10.14	8	0.31	10.18	6	0.26
Regula-Falsi illinois	10.25	9	0.39	10.23	3	0.33	10.27	4	0.25
Regula-Falsi pegasus	10.28	9	0.31	10.24	7	0.33	10.26	10	0.29
Regula-Falsi alternative	10.32	6	0.33	10.26	9	0.30	10.36	10	0.27
Fixed point iteration	12.29	11	0.26	19.04	6	0.17	11.07	14	0.12

5. Conclusions

Evaluating the magnitude of the gain in accuracy is possible to diminish the overall samples size n for achieving the accuracy used as a goal when fixing initially. Commonly in many medical researches involving sampling, there is information in the records of the patients which may be used for designing stratified models and fixing optimal allocations subjecting to cost constraints. Having basic information from a related variable the use of mathematic programming is a good alternative for determining optimal allocations. The Secant and the Regula-Falsi algorithms behaved similarly. The increment in the accuracy found notably high using this methodology. Thus the use of suggested methodology is recommended to prefer in practice for determining the allocation of the sample in such types of studies.

References

1. Bays, H.E., Chapman, R.H. and Grandy, S. (2007). The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. *International Journal of Clinical Practice*, 61(5), 737–747.
2. Berinde, V. (2007). *Iterative approximation of fixed points*. Springer, Berlin.
3. Burgard, . and M nnich, R. (2012). Modelling over and undercounts for design-based Monte Carlo studies in small area estimation: an application to the German register assisted census. *Computational Statistics and Data Analysis*, 56(10), 2856–2863.
4. Gabler, S., Ganninger, M. and Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika,* 75(2), 151-161.
5. Greenlee, H., Unger, J.M., LeBlanc, M., Ramsey, S. and Hershman, D.L. (2016). Association between body mass index and cancer survival in a pooled analysis of 22 clinical trials. *Cancer Epidemiol Biomarkers Prev*., 26(1), 21–29.
6. Hohnhold, H. (2009*). Variants of optimal allocation in stratified sampling*. Technical report, Statistisches Bundesamt Wiesbaden.
7. Koethe, J.R., Jenkins Bryan, C.A., Shepherd, E., Stinnette, S.E. and Sterling, T.R. (2011). An optimal body mass index range associated with improved immune reconstitution among hiv-infected adults initiating antiretroviral therapy. *Clinical Infectious Diseases*, 53(9), 952–960.
8. Münnich, R., Sachs, E.W. and Wagner, M. (2011). Calibration of estimator-weights via semismooth Newton method. *Journal of Global Optimization*, 52(3), 471–485.
9. Münnich, R., Sachs, E.W. and Wagner, M. (2012). Numerical solution of optimal allocation problems in stratified sampling under box constraints. *Advances in Statistical Analysis*, 96(13), 435–450.
10. Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), 558–606
11. Singh, S. (2003). *Advanced sampling theory with applications: How Michael selected Amy.* Kluwer Academic Publishers.
12. Stenger, H. and Gabler, S. (2005). Combining random sampling and census strategies. Justification of inclusion probabilities equal to 1. *Metrika,* 61(2), 137–156.
13. Temple, N.J. (2013). The marketing of dietary supplements: a canadian perspective. *Current Nutrition Reports*, 2(4), 167–173.

A Bayesian quantile time series model for asset returns

G. Mitrodima[1], J. Griffin[2]
[1]LSE, London, UK
[2]UCL, London, UK

Abstract

We consider jointly modelling a finite collection of quantiles over time under a Bayesian nonparametric framework. Formal Bayesian inference on quantile is challenging since we need access to both the quantile function and its inverse. One commonly used approach treats the check loss function as a log-likelihood leading to an asymmetric Laplace likelihood. Alternatively, one can avoid using this mathematical link in order to build the distribution naturally. Thus, we employ a flexible Bayesian implementation of a conditional transformation model and show that this can be used to define a distribution specified by a finite collection of quantiles. The proposed model can be used to define a stationary process of distributions and can be "centred" over a parametric model. MCMC methods are employed to illustrate the use of the model on a sample of stock, index, and commodity returns in a forecasting exercise.

Keywords

Bayesian nonparametrics; conditional distribution; stationarity; efficient MCMC; predictive density

1. Introduction

The modelling of the conditional distribution of asset returns plays an important role in understanding the risk of financial decisions, and a large literature has been developed around this problem. Initial work mostly concentrated on modelling the time variation of the conditional variance. The idea was extended to modelling time variation in higher order moments. Examples of work on higher moments such as skewness and kurtosis, can be found in Hansen (1994), Harvey and Siddique (1999), Backus et al. (1997), and Christoffersen et al. (2013).

Several approaches have also been proposed to build a flexible model for a distribution using its quantiles. One example is to use quantile-based estimators, which are found to be more robust to outliers compared to the empirical moment based estimators of higher moments (see Kim and White, 2004).

In this paper, we consider jointly modelling a finite collection of quantiles over time under a Bayesian nonparametric framework. Formal Bayesian inference on quantile is challenging since we need access to both the quantile function and its inverse. We employ a flexible Bayesian implementation of a conditional transformation model (Hothorn et al., 2014). This allows us to model a finite collection of quantiles but also allows direct access to the likelihood function.

2. Methodology

Transformation models extend the data for which a particular parametric model is suitable by modeling a parametrised transformation of the data by the parametric model. We are interested in time series data, y_t, and define the transformed data

$$Z_t = G_t(Y_t), \tag{2.1}$$

where G_t is an invertible transformation with first derivate g_t. We define $\mathcal{F}_t^X$ to be the history of a generic time series X AND G up to time $t-1$. It follows that, if $p_Z(Z_t|\mathcal{F}_t^Z)$ is the transition density of a time series model for Z_t, the transition density for Y_t is $p_Y(Y_t|\mathcal{F}_t^Y) = g_t(Y_t)p_Z(G_t(Y_t)|\mathcal{F}_t^Z)$. The form of the model allows the quantiles of $p_Y(Y_t|Y_1,\dots,Y_{t-1})$ to be derived as transformation of the quantiles of $p_Z(Z_t(Z_1,\dots,Z_{t-1})$ as $Q_t^Y(\tau) = G_t^{-1}\Big(Q_t^Z(\tau)\Big)$, where $Q_t^Y(\tau)$ is the quantile function for Y_t defined by $p_Y(Y_t < Q_t^Y(\tau)|Y_1,\dots,Y_{t-1}) = \tau, 0 < \tau < 1$.

The model is very general. Consider GARCH or stochastic volatility models which assume that $Y_t/\sigma_t \sim F$ where F is a parametric distribution. This idea underlies also inversion copulas, where $F_Y(y)$ is estimated by the empirical distribution function of Y_t, $\hat{F}_Y(y)$. We will take the alternative approach where we model G_t or Q_t^Y. We believe that this has several advantages over assuming a time invariant G. We build a flexible distribution for which the quantiles can be easily calculated using a parameterised transformation of a parametric distribution.

Definition 2.1 *F has a Linearised Transformation (LIT) distribution if $F(y) = F_0\big(G(y)\big)$, where F_0 is a continuous distribution with median m and quantile function $q_0(\cdot)$, and the transformation G has parameters $0 = a_0 < a_1 < \cdots < a_k = 0.5, \theta_1^- > 0, \dots, \theta_K^- > 0,$ and $\theta_1^+ > 0, \dots, \theta_K^+ > 0,$. We define $x_0^+ = x_0^- = m$ and define $G: \mathbb{R} \to \mathbb{R}$ to be*

$$G(x) = \begin{cases} m, & x = m \\ G(x_{i-1}^-) + \dfrac{1}{\theta_i^-}(x - x_{i-1}^-), & x_i^- < x < x_{i-1}^- \\ G(x_{i-1}^+) + \dfrac{1}{\theta_i^+}(x - x_{i-1}^+), & x_{i-1}^+ < x < x_i^+, \end{cases}$$

$$x_i^- = x_{i-1}^- + \theta_i^-\big(q_0(0.5 - a_i) - q_0(0.5 - a_{i-1})\big) = \sum_{j=1}^{i} \theta_j^-(q_0\big(0.5 - a_j\big) - q_0\big(0.5 - a_{j-1}\big))$$

$$x_i^+ = x_{i-1}^+ + \theta_i^+\left(q_0(0.5+a_i) - q_0(0.5+a_{i-1})\right) = \sum_{j=1}^{i} \theta_j^+ (q_0\big(0.5+a_j\big) - q_0\big(0.5+a_{j-1}\big)).$$

This distribution will be written as $LIT(\theta^-, \theta^+, F_0, a)$ *where* $\theta^- = (\theta_1^-, \dots, \theta_K^-)', \theta^+ = (\theta_1^+, \dots, \theta_K^+)'$ *and* $a = (a_0, a_1, \dots, a_K)'$.

The LIT distribution has some useful properties. Firstly, we can obtain its density function f, where $f(x)$ will have the shape of f_0 with scale parameter θ_i^- for $x_i^- < x < x_{i-1}^-$ and θ_i^+ for $x_{i-1}^+ < x < x_i^+$. Secondly, the quantile function $q(p)$, is available analytically via $q(p) = F^{-1}(p) = G^{-1}\left(F_0^{-1}(p)\right) = G^{-1}\big(q_0(p)\big)$. There is a unique LIT distribution with quantiles $q^*(0.5 - a_i)$ for $i = 1, \dots, K$ given by the choice

$$\theta_1^- = \frac{q_0(0.5-a_{1+1}) - q_0(0.5-a_i)}{q^*(0.5-a_{i+1}) - q^*(0.5-a_i)}, \theta_i^+ = \frac{q_0(0.5+a_{1+1}) - q_0(0.5+a_{i-1})}{q^*(0.5+a_{i+1}) - q^*(0.5+a_{i-1})}.$$

If G is a linear function, then F will be the scaled version of F_0. This suggests building a prior where G is given a flexible form that is "centred" over a linear function and F_0 is interpreted as a standardised centring distribution. We assume that F_0 is a standardised distribution parameterised by Ψ and define a model with $\theta_1^-, \dots, \theta_K^-, \theta_1^+, \dots \theta_K^+$ as time varying parameters.

Definition 2.2 (B-JQTS model). *If* $Y_1, \dots, Y_T$ *is a time series of univariate observations, the time series follows a Bayesian Joint Quantile Time Series (B-JQTS) model* if $Y_T \sim F_T$, *where* $F_T = LIT(\theta_t^-, \theta_1^+, F_0(\cdot; \Psi_t), a)$ *and*

$$\theta_{i,t}^- = H\big(\theta_{i,t-N}^-, \dots, \theta_{i,t-1}^-, y_{t-L}, \dots, y_{t-1}; \lambda_i^-\big), \quad q_0(0.5-a_i) < y < q_0(0.5-a_{i-1})$$
$$\theta_{i,t}^+ = H\big(\theta_{i,t-N}^+, \dots, \theta_{i,t-1}^+, y_{t-L}, \dots, y_{t-1}; \lambda_i^+\big), \quad q_0(0.5-a_{i-1}) < y < q_0(0.5-a_i)$$

where $H(\cdot,\cdot;\lambda)$ is a function parameterised by λ and N and L are the orders of the model.

3. Results

Our data sets consist of the daily log-returns of IBM, S&P500, and crude oil (COIL). IBM and S&P500 data were obtained from the Center for Research in Security Prices (CRSP), while the data for COIL were obtained from the FRED data base. Figure 1 (panel (a)) shows the evolution of each time series and the usual characteristics of financial time series including high volatility associated with turbulent periods.

The evolution of the posterior median of the conditional quantiles for four specifications is shown in Figure 1 (plots (b)-(e)) for IBM, S&P500, and COIL. We find that the inner quantiles (25% and 75%) are less volatile over time than the outer quantiles (5%, 10%, 90% and 95%) for all assets. This is in line with

the stylised facts of asset returns. The spike that we observe in 1987 is a result of the extreme return on Black Monday. The model is also able to describe the volatile period during the end of 90's as a result of the Asian financial crisis. Finally, the model captures the financial crisis of 2008: We also observe that there are no quantile crossings, as the model addresses that by construction.

As a separate exercise we compared the forecasting performance of our model against a Bayesian semi-parametric GARCH (1,1) (DPM-GARCH) model. DPM-GARCH implies that the shape of the distribution after adjusting for the volatility (and so higher order moments such as the skewness and the kurtosis) is time invariant in contrast to our proposed B-JQTS model.

We also compared the performance of the different B-JQTS specifications for different values of K. For this predictive exercise, we considered K = 5;11 and 21 and so α= 0 : 0:125 : 0:5; α = 0 : 0:05 : 0:5; and α =0: 0:025 : 0:5 respectively. The forecasting performance of the models was assessed using the log predictive scores (LPS) (Kim et al., 1998). The findings are presented in Table 1 for IBM, but

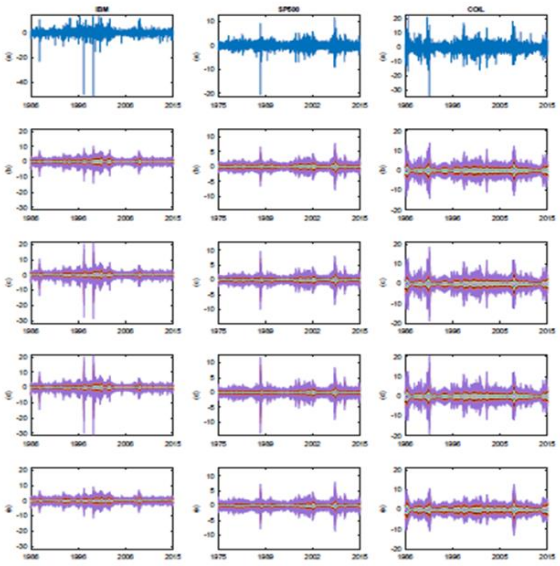


Figure 1: (a) Daily equity returns and posterior median of the conditional quantiles for (b) Bayesian Joint Symmetric Absolute Value (B-JSAV), (c) Bayesian Joint Symmetric Squared Value (B-JSSV), (d) Bayesian Joint GJR (B-JGJR), and (e) Bayesian Joint Symmetric Absolute Value L (B-JSAVL) for IBM, S&P500, and COIL. We use different colours to depict the various quantiles at K = 11 probability levels ranging from 0 : 0:05 : 0:5.

similar findings hold for the other two assets. The forecasting performance of B-JQTS is found to be superior to that of DPM GARCH as we would expect given the properties of B-JQTS. It seems that the specifications for K = 11 tend to better than those when K = 5; while they perform similarly when K = 21. B-JSAV outperforms others at any K.

LPS			
K	5	11	21
BJSAV	1.7005	1.6863	**1.6795**
BJSSV	1.7310	1.7163	1.7204
BJGJR	1.7440	1.7247	1.7190
BJSAVL	1.7542	1.7196	1.7082
DPM-GARCH	1.8305		

Table 1: IBM: The LPS for each model (with bold type used to indicate the best performing model). Under this criterion the superior model is the one which gives the smallest LPS value.

4. Discussion and Conclusion

We consider jointly modeling a finite collection of quantiles over time under a Bayesian nonparametric framework. We believe that this has several advantages. Firstly, we work with a transformation of the conditional distribution (which will usually have a known form) rather than the stationary distribution (for which it may be hard to find the quantiles). Our approach localises the departures from the parametric model. Therefore, certain parts of the conditional distribution may change more rapidly. Lastly, this framework allows us to model both the transformation and conditional distribution of the transformed data as time varying and so it allows for additional flexibility. The empirical application on a sample of stock, index, and commodity returns depicts these advantages.

References

1. Backus, D. amd Foresi, S., K. Li, and L.Wu (1997). Accounting for Biases in Black-Scholes. Technical Report 30, CRIF Working Paper series.
2. Christoffersen, P., S. Heston, and K. Jacobs (2013). Capturing option anomalies with a variancedependent pricing kernel. Review of Financial Studies 26(8), 1963–2006.
3. Hansen, B. E. (1994). Autoregressive conditional density estimation. International Economic Review 35, 705–730.
4. Harvey, C. R. and A. Siddique (1999). Autoregressive conditional skewness. Journal of Financial and Quantitative Analysis 34(4), 465–487.
5. Hothorn, T., T. Kneib, and P. B¨uhlmann (2014). Conditional transformation models. Journal of the Royal Statistical Society, Series B 76, 3–27.
6. Kim, S., N. Shephard, and S. Chib (1998). Stochastic Volatility: Likelihood inference and comparison with ARCH models. The Review of Economic Studies 65(3), 361–393.
7. Kim, T.-H. and H. White (2004). On more robust estimation of skewness and kurtosis. Finance Research Letters 1(1), 56–73.

Comparing record linkage methods for data integration on Brazilian agriculture

Andrea Diniz da Silva[1]; José André de Moura Brito[2]; Djalma Galvão Carneiro Pessoa[3]

[1]National School of Statistical Science and Brazilian Institute of Geography and Statistics
[2]National School of Statistical Science
[3]Brazilian Institute of Geography and Statistics (retired)

Abstract

In Brazil, important indicators of living conditions of the population depend on improving agricultural statistics. From the gross domestic product, that in 2017 had more than 20% of the value attributed to agriculture; to sustainable development indicators for monitoring the eradication of hunger, sustainable growth and sustainable use of natural resources, such indicators depend on continuous production of good quality agricultural data, to be effective. However, despite its importance, most Brazilian agricultural statistics are still based on decennial censuses, whose data become obsolete as time passes; in surveys over subpopulations of establishments, which are restricted to a selected set of products and do not refer to the total population; and in subjective surveys which do not provide error estimates or precision measurement. An alternative to improve agricultural statistics is integrating data from already existing surveys and administrative registers. Such integration can allow both direct production of information and building a master frame for supporting probabilistic sample surveys. Preliminary study conducted at IBGE showed that there are more than 60 sources of data on Brazilian agriculture. Nevertheless, to benefit from this opportunity, it is necessary to develop and apply a method for integrating the available data. Considering that record linkage methods are important tools for such integration, an empirical study was conducted to identify the one providing comparatively better results. Eight record linkage methods were compared, using name and address of agricultural establishments registered in two Brazilian sources: Central Business Register and records from State Tax Administrations, referring to the states of Maranhão, Paraíba and Santa Catarina. The study includes a method based on the Fellegi-Sunter decision model, two methods based on clustering techniques using the K-Means algorithm, four methods using classification trees and a method using the Support Vector Machine (SVM) algorithm. Measures of precision, recall and their harmonic mean (F-Measure) were used to compare the efficacy of competing methods.

Keywords
Record Linkage; Data Integration; Rural Statistics; Fellegi-Sunter; Machine Learning

ACKNOWLEDGMENT

The views in this paper are those of the authors and not necessarily those of either the Brazilian Institute of Geography and Statistics or National School of Statistical Science.

1. Introduction

In Brazil, important indicators of living conditions still depend on improving agricultural statistics. Ranging from the gross domestic product, that in 2017 had more than 20% of the value attributed to industry, trade and services related to agriculture (CEPEA, 2018); to sustainable development indicators for monitoring the eradication of hunger, sustainable growth and sustainable use of natural resources, all of them depend on continuous production of good quality agricultural data to be effective. The current agricultural survey system of the main producer of public statistics in Brazil, the Brazilian Institute of Geography and Statistics - IBGE, is based on decennial censuses, whose data become obsolete as it gets far from the reference date; in surveys over subpopulations of establishments, which are restricted to a selected set of products in addition to not refer to the total population; and in subjective surveys which do not provide error estimates or precision measurement. In addition, data produced on the tripod system presents distinct figures on agricultural production in the census reference years due to conceptual differences and coverage between census and surveys (Proposta..., 2011).

Due to mentioned limitations of agricultural statistics produced, a number of data sources in governmental agencies may be considered to help improving its quantity and quality. Agencies such as IBGE, the Ministry of Agriculture, Special Secretariat of Family Farming, the State Tax Administration, among others, may be used in a complementary way to cover a wide range of data domains. Thus, it is possible improving agricultural statistics, based on administrative registers and surveys, with the advantages of using data produced as part of the regular programs of government agencies. Therefore, using data that have high chances of continuity of production and, because they manage data already collected for other purposes, of relatively low cost when compared to census and new surveys. Integrating survey and register data from such sources is a way to enable

producing information on producers and agricultural production at national and subnational levels, with better geographic and thematic coverage than the one achieved nowadays. In addition, the integrated data can be used to build and update a sample frame and enable designing a complete agricultural statistics system.

However, despite the existence of data sources, the simple aggregation of content does not allow producing statistics at a quality level as required for the Brazilian official statistics. This occurs mainly because some records are incomplete and because some producers have more than one record, what affects the general quality of the data. Thus, it is necessary identifying the records referring to the same producer to then complete the information with data from more than one source and remove the duplicates. Not all data sources have a common unique identifier associated with each establishment, which would greatly facilitate the identification of the records belonging to the same establishment, within and between datasets. However, there is now a variety of record linkage methods that do not depend on a unique identifier to find records belonging to the same unit or individual and enable data integration and deduplication. Nevertheless, despite the wide range of methods, there is no one whose efficacy is superior, except in specific applications. This is because the adequacy of each method depends, at least, on the structure and quality of the data used. Therefore, ad hoc methods have often been used for data integration (Winkler, 1995). At this point, proposing a record linkage method that fits the available data and, concomitantly, allows achieving a required quality of linked data, can increase supply of data and, consequently, contribute to the improve quantity and quality of Brazilian agricultural statistics.

2. The Empirical Study

The data used come from two lists: one corresponding to producers registered in a Business Register – CEMPRE, maintained by the Brazilian Institute of Geography and Statistics - IBGE, dating from 2014; and another to producers in the State Tax Administration - SEFAZ system, in 2016. In both lists, geographic coverage is three Brazilian States: Maranhão, Paraíba and Santa Catarina.

Record linkage followed the three-step approach described in Silva (2018). Step one – Preparing, was dedicated to data preparation for record linkage and corresponded mainly to standardization, expanding abbreviations and acronyms and parsing names. In step two -Preparing, records in CEMPRE dataset were compared to records in SEFAZ dataset. All comparisons were restricted to State. The number of comparisons varied over State due to different number of producers (Table 1).

Table 1: Number of record per source and number of comparisons, by State

State	State		Comparisons
	CEMPRE	SEFAZ	
Maranhão	820	750	615.000
Paraíba	334	348	116.232
Santa Catarina	2.173	235	510.655

Name of the producer and street name of the business address were divided into single parts and first four parts of each were used as comparison variables, together with address number. The similarity scores for these nine variables, on a continuous scale ranging from 0 to 1, were computed using a Jaro-Winkler distance metric. Therefore, a vector with nine similarity scores was generated for each pairwise record comparison. Procedures in step one and two were performed only once, whereas the step three were repeated eight times, corresponding to eight classification algorithms as follows.

These similarity scores were then used for classification into link or non-link based on four approaches: clustering, weighting, classification tree and maximal margin classification. K-means and Bagged Clustering (Bclust) are the two algorithms under first approach. Weighting approach used Fellegi-Sunter (FS) decision model (Fellegi and Sunter, 1969) and Expectation Maximization algorithm to calculate comparison weights used to classify records pairs as a link or non-link. The tree-based approach used Recursive Partitioning Tree (Rpart), Bagging, Bumping and Adaptive Boosting (ADA); and the maximal margin approach used Support Vector Machine algorithm.

R Language was used to perform the three steps. Packages stringr and dplyr (Wickham, 2018; Wickhan et. al., 2017) for first step and RecordLinkage Package (Borg e Sariyar, 2016) for steps two and three.

A gold standard link was defined using a unique identifier, which was used only for checking linkage quality. True positives and negatives as well as false positives and negatives were counted, allowing computing precision and recall, thus its harmonic mean F-measure. Values associated to clustering and weighting based methods correspond to a single classification while for the other methods 100 classification were performed applying cross validation technic. Those quantities were used to compare performance of eight record linkage methods.

3. Comparison of Methods

In Maranhão State, ADA method is the one that stands out for the high **precision**. The other methods show similar precision, except for Kmeans and

Bclust, whose precision is close to zero. In the Satates of Paraíba and Santa Catarina, SVM shows higher precision than the other methods. Therefore, the one with the highest proportion of true pairs among the predicted ones. In these states, the tree-based methods, Rpart, Bagging, Bumping and ADA result in higher more precision than Kmeans, Bclust and FS. The difference appears only in the second decimal when the comparison is made with the FS but when it comes to Kmeans and Bclust the difference is expressive. Note that the low precision observed is not due to absence of true pairs, but rather to the very small number of true positives combined with huge number of false positive (Table 2).

Table 2: Precision of Record Linkage Method by State

Record Linkage Method	State		
	Maranhão	Paraíba	Santa Catarina
FS	0,82	0,68	0,64
Kmédias	0,00	0,00	0,00
Bclust	0,00	0,00	0,00
Rpart	0,81	0,71	0,83
Bagging	0,81	0,68	0,79
Bumping	0,80	0,71	0,82
ADA	**0,88**	0,71	0,86
SVM	0,83	**0,75**	**0,94**

Because **recall** is the proportion of true pairs that have Link as the predicted class, there is sometimes an inverse relation between precision and recall, i.e., highest precision associated with lower recall, and vice versa. This occurs in the state of Paraíba, especially with the methods Kmeans and Bclust, whose precision is almost nil, and the recall is the highest compared to that achieved by the other methods. The FS, in the same state, has higher recall than other methods, although it is not the most precise method. In the states of Maranhão and Santa Catarina, the tree-based methods are those with the highest recall, i.e., those that result in the largest number of true pairs classified as Link (Table 3).

Table 3: Recall of Record Linkage Method by State

Record Linkage Method	State		
	Maranhão	Paraíba	Santa Catarina
FS	0,46	0,75	0,48
Kmédias	0,66	0,84	0,61
Bclust	0,17	0,77	0,33
Rpart	0,67	0,63	0,66
Bagging	0,78	0,68	0,69
Bumping	0,73	0,65	0,68
ADA	0,76	0,70	0,69
SVM	0,57	0,57	0,56

Empirical results do not indicate a method with higher precision and recall, concomitantly, and for this reason, using a synthetic measure, such as the **F-Measure**, to compare the overall efficacy of the record linkage method, becomes even more important. It should be noted that general efficacy here refers to a consensus measure, since the F-Measure is the harmonic mean of the two others. In specific cases, choosing to maximize precision or recall may be preferable. In the states of Maranhão and Santa Catarina, ADA, followed by Rpart, Bagging, Bumping provided highest F-measure. In Paraíba, the measure associated with this method is lower in 0.01 when compared to the one associated with the FS. Therefore, one can say that ADA method provide F-Measure higher or equivalent to the other methods (Table 4).

Table 4: F-Measure of Record Linkage Method by State

Record Linkage Method	State		
	Maranhão	Paraíba	Santa Catarina
FS	0,59	0,72	0,55
Kmédias	0,00	0,00	0,00
Bclust	0,00	0,00	0,00
Rpart	0,73	0,67	0,73
Bagging	0,80	0,68	0,74
Bumping	0,77	0,68	0,74
ADA	0,81	0,71	0,76
SVM	0,68	0,64	0,70

4. Some Final Remarks

In realistic scenarios, choosing the record linkage method may take into consideration specific project requirements, such as the interest of maximising

the number of correct links over the predicted ones (precision); maximising the number of correct links over the existing ones (recall); the fastest; or what works well in equipment with reduced memory capacity. In this work, the criterion used to choose the method was efficacy, represented by the measures of precision, recall and its harmonic mean, the Fmeasure. As already mentioned, tree-based methods similar results regarding their efficacy taking into consideration associated F-measure. Therefore, it is expected to find similar results when using any of these methods to perform the record linkage to integrate data on agricultural establishments using name and address. The suggested methods are those comparatively better, considering the selected set of methods; and the evaluation was based on experimental results. Therefore, those results should be seen with some caution. Different levels and efficacy patterns can be obtained in other applications using other data and other methods.

The experiment involves a variety of approaches, both the traditional Fellegi-sunter decision model, as well as recent ones such as the ones of machine learning area. However, the present study does not exhaust the set of record linkage methods present in the literature and already applied, although with synthetic data. Methods based on Bayesian classifiers or on artificial neural networks, for example, are popular in the literature and have not been compared. In addition, there are several variations of the algorithms such as Adaptive Bumping, Random Forest, Bayesian SVM algorithms, among others, that can be considered in future works.

Considering the record linkage methods presented, there is still a myriad of possibilities for combinations to be explored. For example, in all methods the same comparison function for the calculation of the similarity values (Jaro-Winkler) was used. Despite the use of the most popular function in the literature, in specific cases, other functions may work better. In addition, parameters of classification algorithm can be adjusted. The use of default parameters does not necessarily guarantee the best possible performance of all methods for the data used. Examples of possible adjustments include increasing the number of samples used to implement the Bclust, Bagging, Bumping and ADA algorithms; replacing the linear function used in SVM by a nonlinear function; the modification of the value of the allocation cost of the comparison in each class. In the same sense, the limit that distinguishes link and non-links in Fellegi-Sunter model can also be established based on a few bootstrap samples, using or not modelling to find the most appropriate value.

Developing computational tools in the R language to perform the first step for Portuguese language text data can foster the use of record linkage to solve a number even greater of problems, such as imputation, for example. Although there is already a set of commands and functions, especially in Stringr and plyr packages, a complete package to prepare data for record

linkage does not exist. Just RecordLinkage that is suitable for steps two and three, i.e., comparing and classification.

Last but not least, the experiment is limited to register-to-register linkage. The extension to register-to -survey and survey-to-survey, is certainly an important contribution to the data integration on agriculture.

References

1. BORG, A.; SARIYAR, M. Package 'RecordLinkage': Record Linkage in R. 2016. Available in: https://r-forge.r-project.org/projects/recordlinkage/, accessed 14 Nov 2018.
2. CEPEA. (2018) Centro de Estudos Avançados em Economia Aplicada. PIB do Agronegócio e PIB total – Brasil. Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo. Available in: https://www.cepea.esalq.usp.br/br/pib-do-agronegocio-brasileiro.aspx, accessed 14 Nov 2018.
3. FELLEGI, I.; SUNTER, A. (1969) A Theory for Record Linkage, Journal of the American Statistical Association, Vol. 64, 1183-1210. Available in: http://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf, accessed 14 Nov 2018.
4. PROPOSTA de Sistema Nacional de Pesquisas por Amostragem de Estabelecimentos Agropecuários – SNPA: Concepção geral e conteúdo temático - 2ª versão. (2011). Coordenação de Estatísticas Agropecuárias - COAGRO, Diretoria de Pesquisas - DPE, Instituto Brasileiro de Geografia e Estatística - IBGE.
5. SILVA, Andréa Diniz. (2018). Proposta de método de pareamento para integrar dados sobre a agropecuária. Escola Nacional de Ciências Estatísticas. 93p. Available in: http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2018/tese/AndreaDinizSilvaTese 2018.pdf, accessed 21 Nov 2018.
6. WICKHAM, Hadley. (2018). Package 'stringr': Simple, Consistent Wrappers for Common String Operations. Disponível em: https://cran.r-project.org/web/packages/stringr/stringr.pdf, accessed 21 Nov 2018.
7. WICKHAM, Hadley; FRANCOIS, Romain; HENRY, Lionel; MU¨LLER, Kirill. (2017). Package 'dplyr': A Grammar of Data Manipulation. Disponivel em: https://cran.rproject.org/web/packages/dplyr/dplyr.pdf, accessed 14 May 2018.
8. WINKLER, W. E. Matching and Record Linkage. U.S. Bureau of the Census. Washington D. C. 1995. Available in: https://www.census.gov/srd/papers/pdf/rr93-8.pdf, accessed 14 Nov 2018.

The Impact of ageing population, unemployment, obesity, inflation, out-of-pocket expenses, and income per capita on Malaysia's health expenditures: A linear regression analysis

Abd Aziz Arrashid Abd Rajak, Humaida Banu Samsuddin
School of Mathematical Sciences, Faculty Science & Technology, Universiti Kebangsaan Malaysia

Abstract
The health expenditure contribute almost one sixth of a country gross domestic product (GDP). The GDP plays an important role as a benchmark to a country's economic stance. Therefore, the aim of this study is to analyse the factors that affected the GDP of health expenditure in terms of ageing population, unemployment, obesity, inflation, out-of-pocket expenses, and income per capita. The data conduct is based on annual data from the year of 2000 until 2016. Using the analysis of linear regression, it shows the relationship between these variables and how it affect the GDP of health expenditure. It can be shown that the ageing population and income gives positive relationship based on the model while the other variable shows negative relationship. In conclusion, the variables used in this analysis has a high relationship in determining the impact it has towards the fluctuation of GDP especially in health expenditure.

Keywords
Health expenditure, ageing population, unemployment, obesity, inflation

1. Introduction

Over a couple of decades, it has been acknowledge all over the globe that the healthcare performance is strongly dependent on the economy. The relationship between these components plays a vital role and should not be underestimated, especially in the growth of population's welfare. Just as growth, income, investment, and employment are a function of the performance and quality of the economic system, its regulatory frameworks trade policies, social capital and labour markets depend not just on standards of living, but on the actual performance of health system. However, healthcare system face a tough and complex challenges in part derived by new pressures such as the ageing population, growing prevalence of chronic illnesses, and intensive use of expensive yet vital technology. It can be assumed that wealthier countries have healthier populations as the health performance is interlinked with economic performance. This assumption however is disputed in research by [1]. According to the research, when the economy temporarily improves, the rate of mortality also increases.

Nonetheless, economists argue on the increment of healthcare system during high economic cycle. This is because, it creates job and in turn, people will spend more on the health system. Considering the case in the United States of America (USA) in 2009, Obama stated that healthcare contributed to one sixth of USA gross domestic product (GDP). In 2010, the Patient Protection and Affordable Care Act (Obamacare) was passed and a few years later, healthcare spending increased by 3.9% and produced around 500,000 jobs in the sector. Meanwhile, in the United Kingdom (UK), healthcare spending has increased dramatically since 1950 and the GDP rate has dropped significantly over the years. Health care experts says that it all comes down to money and shortage of staff. The country is not allocating enough cash to pay for the care required for a growing and ageing population. Besides that, innovations in the health sciences have resulted in dramatic changes in the ability to treat disease and improve the quality of life [2]. On the other hand, Malaysia has risen to its immense potential in the development of healthcare system. One of the main challenges in Malaysia is the continuation of the provision of healthcare, especially when Malaysia's population is still growing although increasingly ageing over time with higher demands expected with greater longevity. This leads to higher healthcare expenditure which required extensive health product and expensive technology.

2. Literature Review

Due to changes in population characteristics, the global demand for health services has increased hugely over the past few years [3]. The ageing population is considered to be one of the most worrying aspect that affected the health expenditures. According to [4], the increased in ageing population is because of decreasing in fertility among the youth community and the increase in life expectancy. While research by [5] and [6] shows that age is a predisposing determinant that is not directly responsible for utilisation of health care. Meanwhile, research by [7] in Japan find the solution on the increase in health expenditure especially among the elderly. Research by [8] have used a pseudo-panel of health care expenditure data for Germany to demonstrate that per capita health care expenditure are significantly influenced by the age composition of the population. In recent interview, Tun Dr. Mahathir Mohamad emphasise on the vitality of research to handle the rising health expenditure, especially in Malaysia as it is about to enter the ageing population growth [9].

Another factor that may affect the health expenditure is the unemployment rate. Research by [10] study the relationship between increased unemployment in European Union (EU) states during global economic crisis. They conclude that initiatives that bolster employment and maintain total health care expenditure may help minimise increases in

mortality during economic crises. Meanwhile, study by [11] in USA show a significant slowdown in health care spending and decreased utilisation. Meanwhile, obesity is a risk to health and longevity of life. Based on study by [12] and [13] shows that obesity is associated with an increase in drug prescription. The study indicate that obesity have stronger association with the occurrence of chronic medical conditions and increase in health care and medication spending.

Besides that, the major concern is increases in health expenditure can affect major economic indicators such as GDP, employment and inflation. Study by [14] in Turkey shows that total health expenditure growth rate has a statistically positive affect on the inflation rate. However, research based on health expenditure and inflation rate is still limited around the globe. In literature, studies mostly focus the relationship between health expenditures and income. Such as study conduct by [15] investigates the long run economic relationship between healthcare expenditure and income in the world using data on 167 countries over the period 1995-2012. The study suggests that size of income elasticity depends on the position of different countries in the global income distribution, with poorer countries showing higher elasticity. Meanwhile, study by [16] conduct in China finds that, the income elasticity of China's urban residents indicate that health care is a luxury rather than necessities. All of the factors that seems to contribute in the spike of health expenditure has been carried out all around the globe. The main objective for most of the study is to find the association of said factors towards growth of GDP especially in health expenditure. Despite that, these research is still obsolete and rarely been conduct by Malaysian scholars. Therefore, this paper will focus on analysing these factors in Malaysia.

3. Data and Methodology

The data used in the analysis is collected from the data world bank from the year 2000 until 2016. The ageing population is consider as one of the popular factor in health expenditure. This is because, an increase in the proportion of the aged group is associated with an increase in the prevalence of ill health and therefore increase the health expenditure. In this study, the age population considered is the population for age 65 years and above. The unemployment rate is perhaps the economic variable that worries people the most. An obvious point is that unemployment causes great suffering because people who are out of work, experience economic hardship and unhappiness. Hence, increasing the possibility to health problems.

Obesity is a multifaceted problem with wide-reaching medical, social and economic consequences. The obesity rates is based on Body Mass Index (BMI). For the obese person, excess weight denotes an increased risk of disabling chronic diseases, lowered quality of life and loss of earnings since obesity-

related health burdens carry staggering financial implications. The causes of inflation are often ascribed to the level of demand or supply. Consider an economy running at close to full capacity and suppose that the demand for good rises. This extra demand could come from consumers, from the government wanting to spend more money or even from overseas residents demanding more of the country's exports. However, inflation can be a sign of a healthy, growing economy. It is often part of the process of recovery from recession.

An out-of-pocket expense is the direct payment of money that may or may not be later reimbursed from a third-party source. In India, huge out-of-pocket expenses pushes nearly 7% of the households across the country into poverty every year [17]. Meanwhile, Malaysia high out-of-pocket is due to the small number of people who have health insurance or medical benefits at work. Income per capita measures the average income earned per person in a given area (city, region, country, etc.) in a specified year. It is calculated by dividing the area's total income by its total population. This is used to see the wealth of the population with those of others.

This paper, will focus on regression methods that fall under the rubric of ordinary least squares (OLS) regression. In OLS regression, a quantitative dependent variable is predicted from a weighted sum of predictor variables, where the weights are parameters estimated from the data. The aim of this paper is to predict the response variable on the explanatory variable using OLS regression. OLS regression fits models as in equation (1):

$$\widehat{Y}_i = \beta_0 + \beta_1 \hat{X}_{1i} + \cdots + \beta_k \hat{X}_{ki} \qquad (1)$$

where $i = 1 \cdots n$, n is the number of observations, k is the number of explanatory variables. OLS regression serves three major purposes in an analysis that is description, control and prediction.

The F-statistic is a test to describe how well a response variable related to the explanatory variables. Most F-statistic arise by considering a decomposition of the variability in a collection of data in terms of sums of squares. The test statistic in an F-statistic is the ratio of two scaled sums of squares reflecting different sources of variability. These sums of squares are constructed so that the statistic tends to be greater when the null hypothesis is not true. In order for the statistic to follow the F-distribution under the null hypothesis, the sums of squares should be statistically independent, and each should follow a scaled χ^2-distribution. To test whether there exist relationship between response variable, $\hat{Y}$ and explanatory variable, $\hat{X}_i$, consider the following hypotheses:

$$H_0: X_0 = X_1 = \cdots = X_i = 0$$
$$H_1: X_0 \neq X_1 \neq \cdots \neq X_i \neq 0$$

The F-test is shown in equation (2):

$$F^* = \frac{MSR}{MSE} \tag{2}$$

where MSR is the regression mean square and MSE is the mean square error. Both MSR and MSE can be computed by using formula in equation (3) and (4) respectively:

$$MSR = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}_i\right)^2}{k} = \frac{SSR}{k} \tag{3}$$

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-k} = \frac{SSE}{n-k} \tag{4}$$

Multiple R is the correlation between Y (observation values) and $\hat{Y}$ (estimated value). Multiple R^2 is simply a measure of R-squared (R^2) for models that have multiple predictor variables. Therefore it measures the amount of variation in the response variable that can be explained by the predictor variables. R^2 is given as in equation (5):

$$R^2 = \frac{SSR}{SSR + SSE} \tag{5}$$

Meanwhile, adjusted R^2 controls against this increase, and adds penalties for the number of predictors in the model and is given in equation (6):

$$Adjusted\ R^2 = 1 - \left[(1 - R^2)\frac{n-1}{n-k-1}\right] \tag{6}$$

4. Results

In this section discuss the results from the OLS regression based on the predictor variables (age 65+, unemployment, obesity, inflation, out-of-pocket expenditure, and income) and response variable (GDP). The results reported provide some interesting insights on factors that affect the GDP for health expenditure in Malaysia. Firstly, Table 1 reports the descriptive statistics of the response and explanatory variables. Based on the measure of central tendency in Table 1, it can be seen that variables GDP, age 65+, obesity, inflation , and out-of-pocket appears to be skewed to the right, which explains the higher value of mean compare to the value of median. Meanwhile, the variables for

unemployment and income appears to be skewed to the left. Based on the dispersion of the data, it can be seen that from the value of standard deviation and the range, the data spread around the central tendency.

Table 1 Descriptive Statistics

Variables	Mean	Median	Variance	Std Dev	Min	Max	Range
GDP	3.22	3.17	0.16	0.40	2.56	3.90	1.33
Age 65+	4.84	4.78	0.43	0.66	3.91	6.08	2.17
Unemployment	3.30	3.33	0.06	0.24	2.87	3.69	0.82
Obesity	11.06	10.90	7.24	2.69	7.00	15.60	8.60
Inflation	2.23	2.03	1.33	1.15	0.58	5.44	4.86
Out-of-pocket	35.76	35.53	3.41	1.85	32.51	41.00	8.49
Income	4.28	5.57	14.99	3.87	-4.56	8.67	13.23

Secondly, the correlation between each variables is displays in Table 2. From the correlation table, the data can be analysed into two part that is correlation between response and explanatory variable and correlation within the explanatory variables itself. It can be seen from the table that variables age 65+, obesity, and inflation gives positive relationship to GDP. This means that as one variable gets larger the other gets larger. Meanwhile, unemployment, out-of-pocket and income gives negative relationship to GDP which means that as one variable gets larger, the other gets smaller. Age 65+ give the highest correlation towards GDP that is 0.977. This is parallel to study conduct by [18], [19], and [20]. Based on their findings, ageing population reacts positively to the health expenditure. The ageing population gives the highest correlation towards GDP compare to other variables. This is because, the elders require a long-term care and prone to illnesses and diseases which require constant health care and require higher health expenditure.

While, the unemployment gives a high negative correlation that is -0.433. This results contradict to research such as [21] and [22] which stated that the higher the rate of unemployment, the higher the health expenditure. As it can be seen from the correlation value of unemployment is negatively correlate with GDP health expenditure. The lower rate of unemployment means that job opportunities are high. These high employment reduce the stress associated with economic insecurities which results in lower health expenditures. Research by [1] acknowledge the negative correlation of unemployment and health expenditure. Simultaneously, variables age 65+ and obesity gives the highest positive correlation between explanatory variables that is 0.994. This variable is highly correlated because as the age increase, the rate of obesity among the population might increase. This is mainly because as aged population does not have the time to exercise and control the food consumption which may leads to obesity.

Table 2 Correlation between Variables

Variables	GDP	Age 65+	Unemployment	Obesity	Inflation	Out-of-Pocket	Income
GDP	**1.000**	0.977	-0.433	0.976	0.121	-0.374	-0.223
Age 65+	0.977	**1.000**	-0.44	0.994	0.176	-0.244	-0.222
Unemployment	-0.433	-0.44	**1.000**	-0.445	-0.284	-0.183	-0.131
Obesity	0.976	0.994	-0.445	**1.000**	0.187	-0.294	-0.237
Inflation	0.121	0.176	-0.284	0.187	**1.000**	0.103	0.381
Out-of-Pocket	-0.374	-0.244	-0.183	-0.294	0.103	**1.000**	0.299
Income	-0.223	-0.222	-0.131	-0.237	0.381	0.299	**1.000**

Thirdly, the model summary is displays in Table 3. The regression coefficients indicate the increase in the dependent variable for a unit change in a predictor variable, holding all other predictor variables constant. In Table 3, it shows that the intercept for the model is 1.976. The coefficient estimate for the age population above 65 years is 0.765, unemployment is -0.127, obesity is -0.049, inflation is -0.019, out-of-pocket expenses is -0.041, and income is 0.004. Therefore, the model can be summarise as equation (7):

$$Y_{GDP} = 1.976 + 0.765_{age65+} - 0.127_{unemployment} - 0.049_{inflation} - 0.041_{out-of-pocket} + 0.004_{income}$$

Table 3 Model Summary

Coefficients	Estimate	Std. Error	t-Value	Pr(> |t|)
Intercept	1.976	0.692	2.854	0.017
Age 65+	0.765	0.294	2.600	0.026
Unemployment	-0.127	0.093	-1.371	0.200
Obesity	-0.049	0.075	-0.663	0.522
Inflation	-0.019	0.018	-1.063	0.313
Out-of-Pocket	-0.041	0.013	-3.228	0.009
Income	0.004	0.005	0.804	0.440

Residual Std Error: 0.07107
Multiple R-Squared: 0.9806, Adjusted R-Squared: 0.969
F-Statistics: 84.43, p-value: 5.525e^{-08}

From equation (7), the predicted GDP will increase 0.765 for every 1% increase in ageing population, decrease by -0.127 for every 1% increase in unemployment rate, decrease by -0.049 for every 1% increase in obesity rate, decrease by -0.019 for every 1% increase in the rate of inflation, decrease by -0.041 for every 1% increase in out-of-pocket spending, and increase by 0.004 for every 1% increase in income per capita. RSE is measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an

error term, ε. Due to the presence of this error term, the response variable will deviate from the true regression line. In this paper, the actual GDP expenditure affected by the predictor variables deviate from the true regression line by approximately 0.07107.

R^2 statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. R^2 is a measure of the linear relationship between our response variable from the predictor variables. In this research, the R^2 value is 0.9806 or roughly 98% of the variance found in the response variable from the predictor variables. Finally, F-statistic is a good indicator of whether there is a relationship between our explanatory and the response variables. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis. In this paper, the F-statistic is 84.43 which is relatively larger than 1 and indicate that it is sufficient to reject the H_0 (no relationship between response and predictor variables).

5. Discussion and Conclusion

The age population gives the largest impact on the increasing of health care expenditure in Malaysia. As Malaysia is entering the ageing population, the increasing of the health expenditure based on the age population factor is expected. Followed by the ageing population is the income per capita factor. The result shows the increase in health expenditures is caused by the rising income. Despite that, previous studies carried out in different countries expect different result on this contributing factor. In Malaysia, as the income per capita increase, the health expenditure also increases. This is because, as the income increases indicate that less Malaysian had the time to spend on curricular activities such as exercising which leads to stressful environment. This will lead to high health expenditure. Based on the increasing of income and health expenditure, it can be relate to the rate of unemployment. From our result it can be indicate that as rate of unemployment increase, the health expenditure will decreases and vice versa. Lower unemployment rate means a high employed population. Based on the income factor, high employed rate gives high income rate and will increase the health expenditure in Malaysia. Based on the result, it shows that there are several factors that may affect the fluctuation of health expenditure in Malaysia.

References

1. Ruhm C J 2003 Good times make you sick J. of Health Economics **22** 637-658
2. DiMasi J A, Hansen R W, and Grabowski H G 2003 The price of innovation: new estimates of drug development costs J. of health economics **22(2)** 151-185.

3. Mousnad M A, Shafie A A, and Ibrahim M I 2014 Systematic review of factors affecting pharmaceutical expenditures Health Policy **116(2-3)** 137-146.
4. Zweifel P, Felder S, and Meiers M 1999 Ageing of population and health care expenditure: a red herring? Health economics **8(6)** 485-496.
5. Kildemoes H W, Christiansen T, Gyrd-Hansen D, Kristiansen I S, and Andersen M 2006 The impact of population ageing on future Danish drug expenditure Health Policy **75** 293-311.
6. De Meijer C, Wouterse B, Polder J, and Koopmanschap M 2013 The effect of population ageing on health expenditure growth: a critical review European Journal of Ageing **10(4)** 353-361.
7. Tamakoshi T and Hamori S 2015 Health-care expenditure, GDP and share of the elderly in Japan: a panel co-integration analysis Applied Economics Letters **22(9)** 725-729.
8. Breyer F, Lorenz N, and Niebel T 2015 Health care expenditures and longevity: is there a Eubie Blake effect? The European journal of health economics **16(1)** 95-112.
9. Bernama 2018 Malaysia needs to study practices in other countries, for ageing population New Straits Time Retrieved from https://www.nst.com.my/news/nation.
10. Maruthappu M, Watkins J A , Waqar M, Williams C, Ali R, Atun R, Faiz O and Zeltner T 2014 Unemployment, public-sector health-care spending and breast cancer mortality in the European Union: 1990-2009 The European Journal of Public Health **25(2)** 330-335
11. Maeda J L K , Henke R M, Marder W D, Karaca Z, Friedman B S and Wong H S 2014 Association between the unemployment rate and inpatient cost per discharge by payer in the United States, 2005-2010 BMC Health Services Research **14(1)** 378
12. Vandegrift D and Datta A 2006 Prescription drug expenditures in the United States: the effects of obesity, demographics, and new pharmaceutical products Southern Economic Journal 515-529.
13. Sturm R 2002 The effects of obesity, smoking, and drinking on medical problems and costs. Health affairs **21(2)** 245-253.
14. Baltagi B H, Lagravinese R, Moscone, F and Tosetti E, 2017 Health care expenditure and income: A global perspective Health economics **26(7)** 863-874.
15. Kim T J, Vonneilich N, Lüdecke, D and von dem Knesebeck O 2017 Income, financial barriers to health care and public health expenditure: A multilevel analysis of 28 countries Social Science & Medicine **176** 158-165.

16. Zeng Z, Qi Y and Ge M 2018 Is health care a necessity or a luxury? Evidence from urban China Applied Economics Letters **25(17)** 1204-1207.
17. DNAIndia 2018 Out-of-pocket health expenses push 7% households into poverty every year Daily News and Analysis https://www.dnaindia.com/
18. Lopreite M and Mauro M 2017 The effects of population ageing on health care expenditure: a Bayesian VAR analysis using data from Italy Health Policy **121(6)** 663-674.
19. Pascual-Saez, M, Cantarero-Prieto D and Castañeda D 2017 Public health expenditure, GDP and the elderly population: a comparative study International Journal of Social Economics **44(10)** 1390-1400.
20. Yin J D C and He A J 2018 Health insurance reforms in Singapore and Hong Kong: How the two ageing asian tigers respond to health financing challenges? Health Policy
21. Fattahi M 2015 The role of urbanization rate in the relationship between air pollution and health expenditures: a dynamic panel data approach International Letters of Social and Humanistic Sciences **53** 68-72.
22. Chamberlain D and Menclova A K 2015 The effects of unemployment rate fluctuations on private health insurance coverage in New Zealand New Zealand Economic Papers **49(2)** 157-170.

New Zealand crime and victims survey: Filling the gap

Dr Michael Slyuzberg, Tianying Chu, James Swindells
New Zealand Ministry of Justice, Wellington, New Zealand

Abstract

The New Zealand Crime and Victims Survey (NZCVS) is a nationwide, face-to-face annual survey. It asks 8,000 randomly selected New Zealanders aged 15 years and over, about incidents of potential crime that they experienced in New Zealand over the last 12 months. This includes both incidents reported to the Police and unreported incidents. These incidents are later assessed and categorised by expert coders. The NZCVS consists of core crime and victimisation questions that repeat every year, and additional modules on different topical subjects that change from year to year. It provides information for researchers, policy makers and the public about the nature and extent of crime and victimisation in New Zealand. The NZCVS was first conducted in 2018.

Keywords

Statistical surveys; Crime statistics; Survey design

1. Introduction

There is very little information about the amount of crime and the number of victims in New Zealand. It is something of an information "black hole" and it is well established internationally that not all crime is reported to police. The NZCVS estimates that only 23 percent of crime is reported. The NZCVS fills in this information void by providing unique valuable information and data for researchers and policy makers in the Ministry of Justice, Statistics New Zealand (Stats NZ), the Ministry of Social Development, the Ministry of Māori Development, New Zealand Police, the Department of Corrections, the Ministry for Children, and the Ministry for Women. It is also of great value to universities, NGOs working in the justice sector, media and general public.

The NZCVS replaces the New Zealand Crime and Safety Survey (NZCASS) following a 2015 review by Stats NZ. A key recommendation of that review was to explore options to redevelop NZCASS in order to report crime volume data annually, expand the crime type coverage, allow more comprehensive data analysis and improve the cost efficiency of running the survey and delivering the results.

The NZCVS consists of a core crime and victimisation questions that repeat every year, and additional in-depth modules on different topical subjects that

change from year to year. A family violence in-depth module was selected for 2018, the first time that the NZCVS was conducted. The survey design was developed after extended consultations with key stakeholders.

2. Methodology[1]

Sampling. Nationwide, random probability sampling, with one respondent selected per household, using multistage cluster sampling methods. Primary sampling units (PSUs) were drawn from Stats NZ's Household Survey Frame[2]. Houses were selected within each PSU. A single respondent was selected from within each dwelling. Each respondent then answered questions about incidents they had experienced. Two samples were drawn for NZCVS: the main sample and a Māori[3] booster sample. The purpose of the Māori booster sample was to ensure that the survey collected sufficient data from Māori to produce reliable results for this group.

Modes of interviewing. Interviews as part of the NZCVS were conducted using:

- computer-assisted personal interviewing (CAPI), where interviewers enter respondents' answers into a laptop; and
- computer-assisted self-interviewing (CASI), where respondents are handed the laptop and can enter their own responses.

There are three key advantages to this mode of interviewing in relation to the NZCVS:

- computer-assisted interviewing software ensures that survey logic is adhered to;
- the selection of victim forms can be automated; and
- respondents can answer sensitive questions confidentially using CASI and reduce bias.

The questionnaire. A comprehensive questionnaire was designed to meet all key survey objectives and, as much as possible, specific user requirements identified during stakeholder consultations. The questionnaire includes: initial demographics, victim screener questions, victim form questions, in-depth module questions, main demographics, and exit and re-contact questions.

Selection of incidents. During the screener questions, respondents were asked how many incidents of each type of potential crime they had

[1] More details are available from https://www.justice.govt.nz/assets/Documents/Publications/NZCVS-2018-Methodology-Report-Year-1-fin.pdf

[2] See http://archive.stats.govt.nz/survey-participants/survey-resources/hes-resource.aspx

[3] Māori are the indigenous people of New Zealand. Māori make up about 15 percent of the New Zealand population.

experienced in the past 12 months. The NZCVS consists of 29 screener questions and 17 follow-on clarification questions. The questions do not ask directly about a particular type of crime. Rather, they describe different situations that respondents might experience. The follow-on questions collected additional information about the incident which enabled a provisional incident code to be assigned.

Individual and cluster victim forms. In order to collect as much information about as many incidents as possible, similar incidents (where a similar thing was done, under similar circumstances and probably by the same person / people) were grouped together, and the respondent was asked the questions about the group of incidents as a set. These were termed 'cluster' victim form questions. Where two or fewer incidents were recorded for a particular incident scenario, the respondent was asked about each incident separately. These were termed 'individual' victim form questions and related to a single incident.

Offence codes. As part of the design process, we identified a list of offences from the Australian and New Zealand Standard Offence Classification (ANZSOC) database, that were to be considered in-scope for the survey. However, consultations showed there was no need amongst stakeholders for offence data at this level of granularity. Therefore, offences were aggregated into more-general classifications, that aligned with the categories expected to be used in the reporting. These broader classifications were designed to also maintain consistency with the Police coding practice. Overall, we identified 18 groups of offences.

The NZCVS is focussed on the victimisation and experiences of the survey respondent, not third parties. This means the survey does not include offences when:

- there is no victim or the victim is unidentifiable (e.g. drug offences);
- the victim is another family member (e.g. child);
- the victim is not alive (e.g. murder and manslaughter); and
- the victim is a commercial entity or public sector agency (e.g. shoplifting, benefit fraud, etc.).

Automated and manual coding. A key objective for NZCVS is to reduce the amount of post-hoc manual offence coding, produce data and publish the results in a timely manner. To meet this objective, an automated offence coding algorithm was developed and programmed into the survey. The algorithm took the inputs from the screeners and follow-on questions to automatically assign an offence code. The algorithms used, were reviewed by Police to ensure they reflected the Police coding practice.

In addition, each incident was manually coded by a trained coder who did not know the automatic coding outcomes. If the automated and manual coding results were different, the incident was escalated to the coding team manager and, if necessary, to the Police registrar for a final decision. This approach proved its effectiveness.

Weighting. The project team worked with Stats NZ to design a weighting methodology for NZCVS that was robust and clearly defined. The following weights were used to adjust collected data for factors such as differential selection probabilities, non-response patterns and sample skews relative to population figures: household selection weights, adjustments for non-response, person weights, incident weights and replicate weights (used to calculate standard errors for estimates derived from NZCVS data).

Imputation. The new design of NZCVS has to a large extent eliminated this requirement, although some imputation is still needed so that all the information collected in the survey can contribute to the analysis of results. In particular, imputation has been used for missing income data (using the R package "hotdeckimputation" based on the nearest neighbour technique) and for assigning some final offence codes when a victim form was not completed. The imputation methods employed for NZCVS were designed in consultation with Stats NZ.

Heavy victimisation cut-off (capping). Within each offence code there might be respondents that report a large number of incidents. In crime and victims' surveys in some other countries the incident counts have been censored so that in any single survey the very heavily victimised do not contribute to the data analysis. A cut-off of 98 percent has typically been employed. In NZCVS incident frequencies have been "capped" at the 98th centile for each offence type.

Testing. Before field work, the survey tool moved through multiple phases of testing. This included internal and external (stakeholders') peer-reviews, extensive cognitive testing of the questionnaire, CAPI / CASI programmes testing, a survey pre-pilot trial (testing on highly victimised respondents), and a full-scale pilot study.

As a result of each testing, a number of changes were made to improve the survey / methodology.

Interviewers' training. The field team was formed from a pool of experienced interviewers who had a proven track record working on other large government surveys. All interviewers completed baseline training modules

including public sector surveying, maximising response rates, cultural awareness, enumeration and safety management. On the next training phase interviewers were required to complete a set of specific online training modules: the purpose of the survey and use of the data, survey methodology and fieldwork procedures, survey content and areas to pay attention to, and orientation of the NZCVS Sample Manager. Finally, all interviewers in the launch team attended a face-to-face training day. At the end all interviewers were assessed to confirm that they were ready to begin interviewing as part of the NZCVS. The assessments included an examination of recruitment technique, interview delivery and incident description recording.

Field work. Field work was conducted seven days a week between the hours of 9:00am to 8:00pm. Occasionally, respondents requested an appointment time outside of these hours.

In order to increase the likelihood of finding a resident at home, interviewers visited households on a mixture of weekdays and weekends and at different times of the day. There were no differences in visiting days or times between urban and rural areas.

Up to a maximum of 10 calls were made in person to selected dwellings.

A robust quality assurance process consisting of monitoring key interview statistics, assessing the quality of collected data and random telephone audits was in place to ensure high quality interviews and data recording.

3. Results[4]

Field work statistics. The key field work statistics results presented in *Table 1* below.

	Main	Māori booster	Overall
Dwellings visited	6,633	3,481	10,114
Estimated eligible	6,528	3,441	9,968
Projected number of interviews	5,400	2,880	8,280
Number of interviews achieved	5,273	2,757	8,030
Interview yield from dwellings visited	79%	79%	79%
% of projected completed (interviews achieved/projected)	98%	96%	97%
% of total sample	66%	34%	100%
Response rate	81%	80%	81%
Data linking consent	93%	92%	92%
Consent for future research	93%	92%	93%

Table 1. Summary of key fieldwork statistics by sample

[4] More details are available from https://www.justice.govt.nz/assets/Documents/Publications/NZCVS-2018-Topline-report.pdf

Victim forms. Respondents could complete up to eight victim forms during the interview. *Table 2* presents the distribution of victim forms completed per respondent.

Victim forms completed	Number of respondents	%
0	5,026	62.6
1	1,733	21.6
2	694	8.6
3	277	3.4
4	147	1.8
5	94	1.2
6	32	0.4
7	24	0.3
8	3	0.0
Total	**8,030**	**100.0**

Table 2. Distribution of victim forms completed per respondent

The total **average interview duration** was 21.5 min varying from 16.7 min for those not reported crime incidents to 91.5 min for those filled 8 victim forms (the highest amount of forms available for reporting).

Summary of key findings.

A. The extent and nature of crime

- A significant majority of New Zealanders (71 percent of adults and 80 percent of households) experienced no crime over last 12 months.
- 1,777,000 offences were identified over last 12 months, where personal offences make up the majority (68 percent of total offences).
- On average, there were 32 household offences per 100 households and 30 personal offences per 100 adults.
- The three most common offences were burglary (17 offences per 100 households), harassment and threatening behaviour (8 offences per 100 adults) and fraud and deception (7 offences per 100 adults).

Percentages of adults and households experiencing one or more incidents of crime (prevalence of crime) presented below on *Figure 1* and *Figure 2* accordingly.

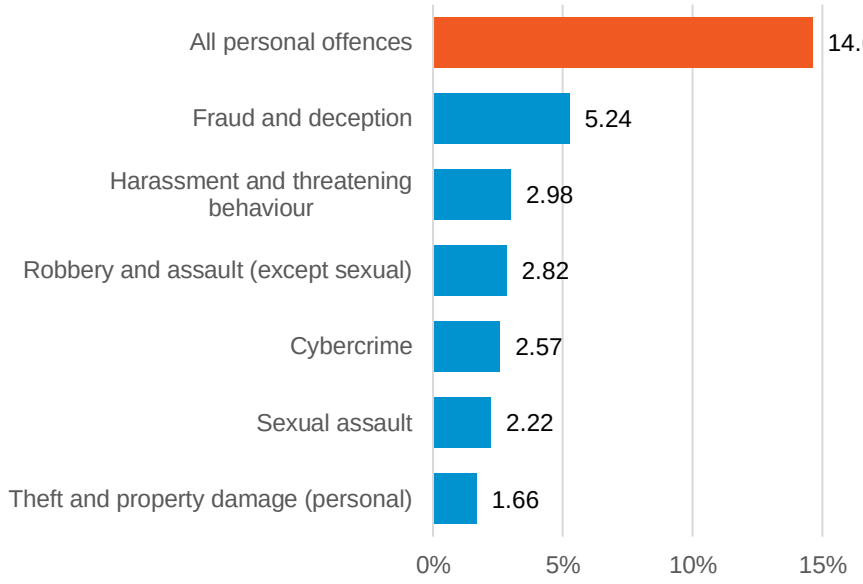


Figure 1 Percentage of adults victimised once or more by personal offence type

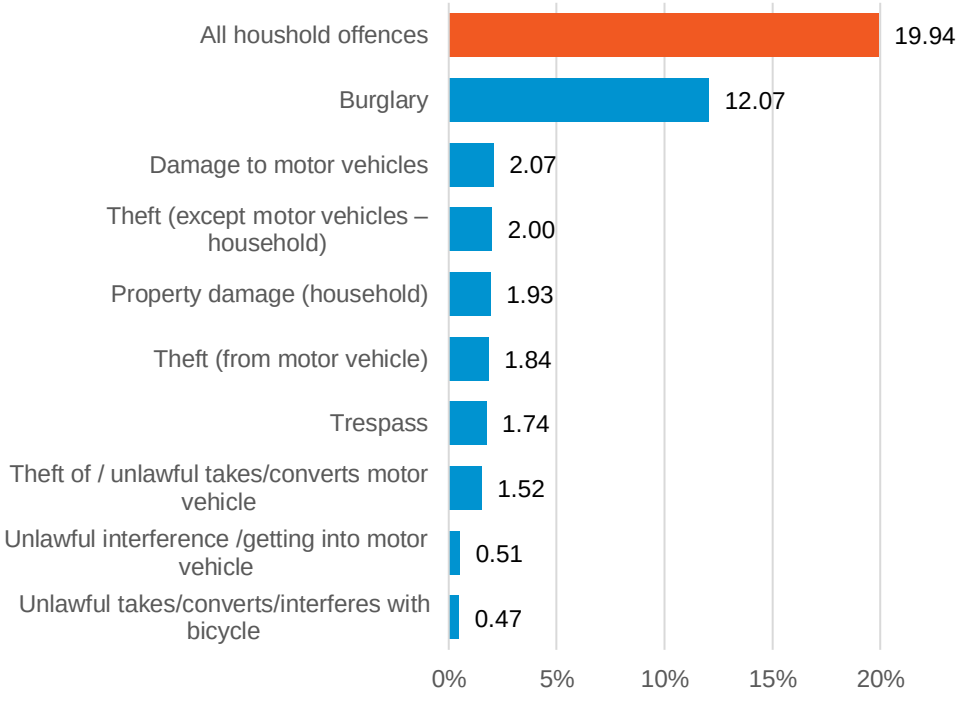


Figure 2. Percentage of households victimised once or more by household offence type

B. Who experiences crime

- Māori (37 percent) were more likely to be victims of crime over 12 months.
- Men (29 percent) and women (29 percent) were equally likely to be victims of crime over last 12 months.

- Women (21 percent) were more likely than men (10 percent) to have experienced one or more incidents of partner violence at some point during their lives.
- Women (34 percent) were more likely than men (12 percent) to have experienced one or more incidents of sexual violence at some point during their lives.
- People aged 65 and over (18 percent) were less likely to be victims of crime and people aged 20 to 29 (40 percent) were more likely to be victims of crime.
- People living in the three major centres (Auckland 29 percent, Wellington 33 percent, Canterbury 29 percent) had about the same chance to be victims of crime as an average New Zealander (29 percent).

C. Reporting of crime

- Less than a quarter (23 percent) of all crime was reported to the Police over last 12 months.
- More household offences (34 percent) as opposed to personal offences (17 percent) were reported to Police.

4. Discussion and Conclusion

Wider picture of crime in New Zealand. NZCVS delivers a more comprehensive picture of crime in New Zealand compared with administrative data because only about one quarter of offences are reported to Police. Therefore, the survey provides unique data for policy makers and the entire Justice Sector allowing to improve services and make better decisions.

Key benefits. The key benefits that NZCVS provides are:

- an increased ability to quantify the underlying level of crime;
- an improved ability to monitor crime trends over time by delivering annual reports;
- an ability to collect particular aspects of victimisation or types of crime and to learn about victims' experience related to the selected prioritised topic;
- an improved ability to support performance monitoring for the wider Justice System; and
- an improved ability to analyse survey results by linking victimisation to other outcomes by bringing the NZCVS into Stats NZ's Integrated Data Infrastructure (IDI) in order to better inform conversations and decision-making.

Limitation: comparability with previous surveys. NZCVS is a new survey with some significant differences in design compared with its predecessor NZCASS. In particular, the NZCVS:

- has a larger annual sample (8000, versus 7000 for NZCASS)
- uses different approach to offence coding (more consistent with Police approach)
- applies much lower level of data imputation as compared with NZCASS
- covers additional offence types (e.g. fraud, cybercrime, trespass)
- employs different approach for collecting data from highly victimised people (allowing similar incidents to be reported as a cluster).

These differences, especially the different approach to offence coding and to data imputation make direct comparison with NZCASS highly questionable, even within similar offence types.

However, consistent annual reporting provides significantly better opportunity to build a reliable time series and analyse victimisation trends. The NZCVS will produce a much greater range and depth of information than the previous survey, and this data will be current.

Future plans. In December 2018 we published a high-level analysis of information collected by the NZCVS and the methodology report. We intend to gradually provide other reports and resources on the NZCVS pages of the Ministry of Justice website. In particular, we expect to publish a full report in the first quarter of 2019 and then to release a series of follow-up reports on specific topics, such as family violence, Māori victimisation, regional offences, victims' experience and more using confidentialised data in the Stats NZ's ID

Clustering planar shapes combined with multidimensional scaling

Moua Golalizadeh
Department of Statistics, Tarbiat Modares University, Tehran, Iran

Abstract

One way to analysis the objects statisticallyis to rely on the statistical shape analysis, which deals with all geometrical information remained after filtering the translation, rotation and scale effects out from the objects. To distinguish objects from each other and then cluster them in terms of their geometrical structures, i.e. shapes, received a great attention in recent years. This paper considers clustering planar shapes combined with applying multidimensional scaling. Variant shape distances are utilized for evaluating the performance of the proposed methods. The idea is applied to a real-life example in which the gorilla skulls are to be separated regarding their shapes.

Keywords

Shape analysis; Clustering; Multidimensional scaling; Shape distances; Gorilla skulls

1. Introduction

The statistical shape analysis is a new field of science which studies, geometrically, objects ignoring the translation, rotation and scale effects already existed on the objects. This field was first introduced to the statistical communities by Kendall (1977). He later advocated it in more details with some rigorous mathematics concepts (Kendall, 1984).

There is an interest to cluster the objects using their shapes. It can also be done in terms of their sizes. However, there is a challenge to deal with such problem in the context of shape analysis. The objective here is to preserve the geometry throughout any statistical analysis including object clustering. In another word, the data in the context of the statistical shape analysis are the shape of the objects (Bookstein, 1986). Precisely, the data are in fact members of the shape space, which is curved rather than flat (Dryden and Mardia, 2016). Mathematically, the topological structure of the objects should be taken into account while obtaining, for instance, the mean of the shape and then deriving Within Sum of Squares of the Groups (WGSS), as a suitable measure in cluster analysis. Hence, extra care is needed in clustering shape data because many statistical tools, already available for the data on the Euclidian space, do not directly work here.

To motivate the problem tackeled in this paper, considerthe pictures of the gorilla skulls, displayed in the Figure 1. We shall recall it later while the aim is on implementing the methodology described in this paper. As seen from the Figure 1, the skulls have the same shape, although they are different in their position, direction and scale. An initial statistical analysis of these skullswas done by O'Higgins (1989). Other shape analysis views to this data are given in Dryden and Mardia (2016).

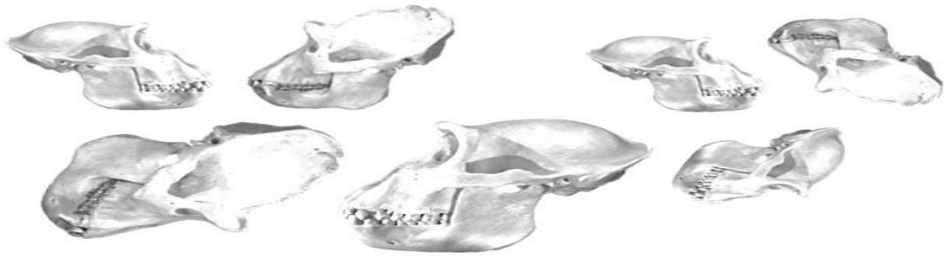


Figure 1: The pictures of some typical skulls of gorilla. The objective is to cluster them in terms of their sizes and shapes.

Multidimensional scaling view to shape analysis data has also been considered in the literature. Kent (1994) suggested the multidimensional scaling to the shape data for which a particular case holds. Since the classical multidimensional scaling to the shape data leads to a biased estimate of mean shape under normal errors, Lele and Richtsmeier (2001) proposed the method of Euclidean distance matrix analysis to correct such bias. Dryden et al. (2008) highlighted the connection between the embedded copy of the shape space and multidimensional scaling and obtained a central limit theorem, enabling them to use the standard statistical techniques to analysis the shape data. However, there is a rare study on using multidimensional scaling for clustering the shape data. This paper will address such view.

The main purpose of this paper is to present a procedure on adopting classical multidimensional scaling for clustering of the pre-labelled planar objects. To highlight this, a brief background of the statistical shape analysis and the related definitions are first presented. Then, after providing some materials on clustering the shape data, a procedure to implement the multidimensional scaling of such data is provided. We also report the results of applying the proposed procedures on gorilla skulls data. Finally, we give some general conclusions and hints on possibility of further research in this field.

2. Methodology

One of the well-known definitions of the shape has been introduced by Kendall (1977) as follows: Shape is any geometrical information of object after removing translation, scale and rotation. Based upon this definition, all

homologous images are considered to be in one class and so can be represented as a single entity on a curved space. From this point of view, one encounters with the equivalent classes and so particular mathematical theories are needed to further analysis of the data. Dryden and Mardia (2016) can be consulted for reviewing those theories and other related topics for more details.

As known, the cluster analysis attempts to reveal the inter-relationships among the features into the final allocation. A function with high success to combine the features in such a way that its misclassification rate is low will be a suitable choice for clustering purpose. In another word, any clustering procedure is attempting to discover the same labelling of the individuals as it was known before initiating the clustering analysis. To do this, one can undergo through either parametric or non-parametric procedures. See, for example, Mardia et al. (1979) for more details. Although the implemented algorithm is a crucial part of selecting a candidate cluster analysis, the function with an optimized extracted features, as the inputs, plays a vital role as well (Huang et al., 2015).

Suppose the landmarks set on the Figure 1 are represented in a configuration matrix, say A. Then, it is common to remove the location from A, via pre-multiplying it with Helmert sub-matrix H. Scale and rotation effects are omitted via invoking Euclidean norm and rotation matrix, respectively. The ultimate quantity, as a matrix, is called the shape of the raw configuration. Note that taking out location and rotation effects but not scale leads to pre-shape which is one step before getting shape.

The inner product of the pre-shape matrices are main sources to start our multidimensional scaling procedure. In particular, we can derive various shape distances from pre-shape objects and then apply multidimensional scaling to them. On the other hands, we are seeking to find clusters with most similar objects. This can also be done through invoking many cluster procedures on the shape distances. However, proper adoption needs to be done here as the shape space is curved while methods to perform cluster analysis is defined for the flat spaces.

From the practical point of view, one of the most common algorithms to cluster the data is the algorithm proposed by Hartigan and Wong (1979). Similar to other contexts in statistics, it was implemented in the statistical shape analysis by Amaral et al. (2010) via comparing the performance of various shape distances in clustering the objects from oceanography. Nabil, and Golalizadeh (2016) evaluated performance of clustering objects via comparing different shape distances and various clustering criteria. Two vital facts on invoking the Hartigan and Wong's algorithm in the shape analysis case have been ignored in the most research conducted so far. If one measures some versions of the Euclidean distances among the data, she could then

utilize it in this algorithm. Otherwise, one should change the steps of this algorithm to check how far an individual is from the centre of the mass data. Bearing this in mind, we adopted the Hartigan and Wong's algorithm to cope with the non-Euclidean property of the shape space. Another critical issue in implementing the Hartigan and Wong's algorithm, which mostly arises in the real application, is concerned about the starting value of the algorithm. To investigate this, we also repeated our clustering studies via changing the initial points and selected those values with optimal performances.Moreover, to choose the starting points for initiating the K-means cluster analysis, we followed a simple trick. First, we computed the distance matrix for all paired of observations. Then, those observations with the largest distance was considered as the starting values. Thereafter, the clustering algorithmwas followed based upon the common procedures described in many multivariate textbooks.

Similar to the situation in Euclidean space, one can invoke the multidimensional scaling to have an idea of the numbers of clusters in clustering the shape data. Although most of the shape distances take the differential geometry of the shape space into account, the multidimensional scaling technique applied to those distances leads to the points represented on the Euclidean spaces (Dryden et al., 2008)

3. Results

We applied the multidimensional scaling to the gorilla skulls data. The data contain 59 samples (29 male and 30 female) for gorilla skulls and are available in the package shapes; freely available in the software R. The Figure 2 shows the result. As seen, the male gorilla numbered 28 is far from the group of males in the plane. Despite its ambiguity, the gorilla numbered 11, which is a male gorilla, is among the female group. So, a shortcoming result of this experiment is that there are some problems with implementing the multidimensional scaling on shape data. It might be either due to invoking incorrect algorithm or to using an inappropriate distance in the procedure of clustering. Consequently, extra caution is needed in utilizing, directly, the multivariate techniques to analyse the shape data.

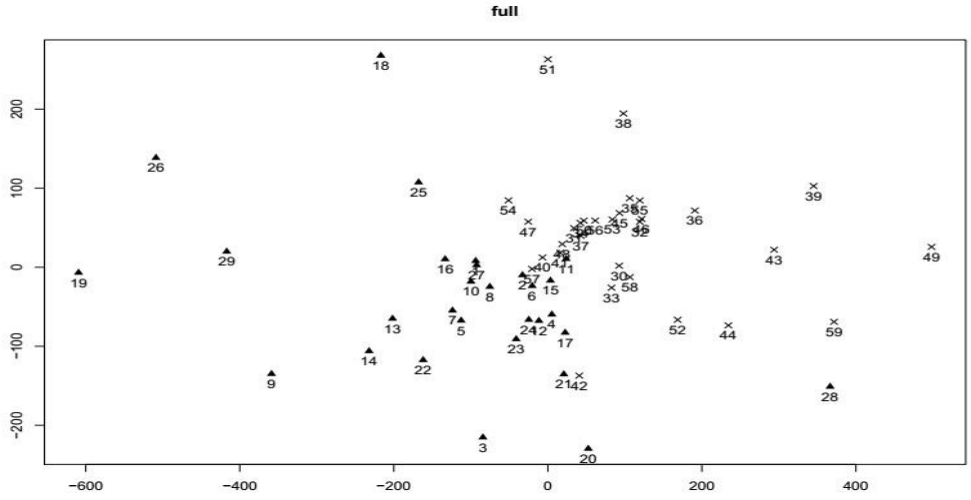


Figure 2: Multidimensional scaling applied to the full Procrustes distances for the gorilla skulls. The triangle and cross signs indicate the male and female gorilla skulls, respectively.

One of the important distances in the context of the statistical shape analysis is shape-and-size distance (Dryden and Mardia, 2016). It takes both the shape and size of the objects into account while attempting to derive a measure of similarity among the objects. We are interested in using this distance for our gorilla skulls data. The results of applying the multidimensional scaling on the shape-and-size distance for the gorilla data is given in Figure 3. As seen, a promising separation between female and male gorillas has occurred in this case. Hence, it sounds the size is a key covariate in assigning the correct label for separating the gorilla skulls.

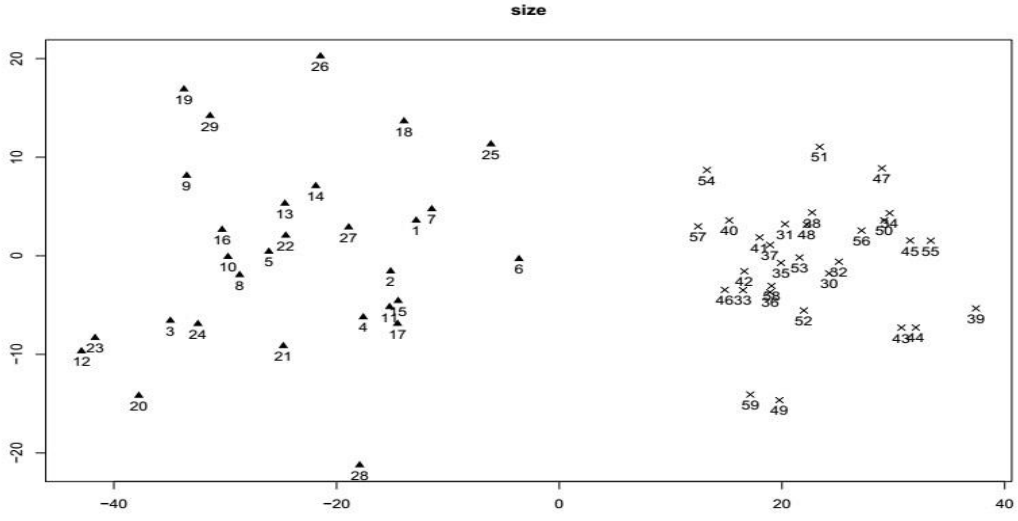


Figure 3: Multidimensional scaling applied to the size-and-shape distances for the gorilla skulls. The triangle and cross signs indicate the male and female gorilla skulls, respectively.

It is common to evaluate the clustering procedure after its implementation on any data set. Different techniques can be taken into account based on the

main feature of data under study. This is also the case for the shapes data where there are various measures for this purpose. To recall the concept of the shape analysis as well as the shape space, it is of great important to invoke some appropriate criteria to validate the cluster analysis. To cope with the literature in the statistical shape analysis, we used Root Mean Square RMS of various shape distances. In particular, we considered all shape distances along with the common Euclidean distance to evaluate the clustering of gorilla skulls data. The RMS of distances are reported in Table 1. As seen, the full Procrustes distance outperformed all alternative distances. Amazingly, the Euclidean distance did a reasonable job despite its weakness in taking the shape structure into account.

Table 1: The RMS criterion for gorilla data in each group before and after clustering in terms of different shape distances as well as common Euclidean distance. Before and after refers to before (after) clustering.

Distance	Step	Male Gorilla	Female Gorilla
Full procrustes	Before	265.89	147.99
	After	249.08	145.81
Riemannian	Before	0.25	0.22
	After	0.25	0.22
Tangent space	Before	71.39	54.20
	After	54.20	70.97
Partial procrustes	Before	0.25	0.22
	After	0.25	0.22
Size-and-shape	Before	89.04	63.91
	After	89.04	63.91
Euclidean	Before	212.57	163.79
	After	199.23	178.69

4. Discussion and Conclusion

Shape as the data play an essential role in various scientific fields. To cluster the shapes, registered based on the tools from statistical shape analysis, is also of interest in some real-world applications. Since the shape data are members of non-Euclidean space, to invoke standard clustering methods require particular care. To fill in the gap in this topic, we suggested implementing the multidimensional scaling as input for clustering the shape in this paper.

In real application case, we observed that to consider the size-and-shape distance is superior to other shape distances while applying clustering procedure combined with multidimensional scaling for the gorilla skulls, it is better . The size can, sometimes, play a key role on clustering the objects.

Unlike other distances, the size-and-shape distance has an ability to take this measure along with geometrical structure into account.

The procedure described in this paper can be followed from model-based clustering view. Although there is an activity (Huang et al., 2015) in this area by using the offset-normal distribution, it might be feasible to use another density for clustering purpose. In particular, a well-known distribution in the shape analysis context is the size-and shape offset normal. So, one can consider it for clustering the shape in a model-based clustering.

References

1. Amaral, G. J., Dore, L. H., Lessa, R. P. and Stosic, B. (2010) K-Means Algorithm in Statistical Shape Analysis, Communications in Statistics: Simulation and Computation, 39, 1016-1026.
2. Bookstein, F. L. (1986) Size and Shape Spaces for Landmark Data in Two Dimensions, Statistical Science, 10, 181-222.
3. Dryden, I. L. Kume, A. Le, H. and Wood, A. T. A. (2008) A multi-dimensional scaling approach to shape analysis, Biometrika, 95, 779-798.
4. Dryden, I. L. and Mardia, K. V. (2016) Statistical Shape Analysis, Chichester, Wiley.
5. Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136: A K-means Clustering Algorithm, Applied Statistics, 28, 100-108.
6. Huang, C., Styner, C., Zhu, H. (2015) Clustering High-Dimensional Landmark based Two-Dimensional Shape Data, Journal of the American Statistical Association, 110, 946-961.
7. Kendall, D. G. (1977) The Diffusion of Shape, Advances in Applied Probability, 9, 428-430.
8. Kendall, D. G. (1984) Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces, Bulletin of the London Mathematical Society, 16, 81-121.
9. Kent, J. T. (1994). The Complex Bingham Distribution and Shape Analysis, Journal of the Royal Statistical Society, Series B, 56, 285-299.
10. Lele, S. R., & Richtsmeier, J. T. (2001). An invariant approach to statistical analysis of shapes. Chapman and Hall/CRC.
11. Mardia, K.V., Kent, J. T. and Bibby J. M. (1979), Multivariate Analysis. London: Academic Press.
12. Nabil, M. and Golalizadeh M. (2016), On Clustering Shape Data. Journal of Statistical Computation and Simulation, 86, 2995-3008.
13. O'Higgins, P. (1989) A Morphometric Study of Cranial Shape in the Hominoidea, Ph.D. Dissertation, University of Leeds, Leeds

Index

A

Abd Aziz Arrashid Abd Rajak, 399
Abubakar S. Asaad, 31, 38, 55, 188
Ahmad Risal, 266
Alfatihah Reno Maulani Nuryaningsih S. P. M.S.ST.,M.Si, 301
Amal El Sharnoby, 179
Andrea Diniz da Silva, 391
Astrid Ayu Bestari, 310
Aulia Dini, 330

B

B. K. Hooda, 101
Bappi Kumar, 284
Barbara Bogacka, 124

C

Carlos N. Bouza-Herrera, 377
Chang-Yun Lin, 9
Chin Wen Cheong, 236
Christine Ablaza, 145

D

D. S. Hooda, 101
D. Stephen Coad, 124
Desiree Robles, 188
Divina Gracia L. del Prado, 31, 38, 55
Djalma Galvão Carneiro Pessoa, 391
Du Jin Zhu, 201

E

El Mostafa TOGUI, 1
Ema Tusianti, 226
Erniel B. Barrios, 31, 38, 55

F

Fethi Şaban Özbek, 78
Filali A. Fatine, 115
Firano Zakaria, 115

G

G. Mitrodima, 386
Gajendra K. Vishwakarma, 377
Gholamhossein Mosalmani Nooshabadi, 169
Guixiang Zhao, 259

H

Hanin Rahma Septina, 330
Henning Omre, 70
Holger Leerhoff, 368
Hong Liu, 108
Humaida Banu Samsuddin, 399

I

Ibarra Aaron R. Poliquit, 55
Indrani Basak, 92

J

J. Griffin, 386
James Swindells, 409
Jason Hsia, 259
Jay M. Manlapaz, 38
Jay R. Manlapaz, 55
Jeffrey Butler, 368
Jing ShengZhe, 236
Jitendra Kumar, 84
Joao Pedro Delgado, 45
José André de Moura Brito, 391

K

Kadek Swarniati,S.ST, 301
Kakoli Rani Bhowmik, 274
Kelvin Hii Chee Yun, 209

L

Liu ChengZhi, 236
Liu Jia, 201

M

M. Iftakhar Alam, 124
M. Shafiqur Rahman, 319
M. Shafiqur Rahmanz, 338
Machell Town, 259
Mae Abigail O. Miralles, 188
Manik Awale, 241
Manoj Chacko, 63
Md. Atiqul Islam, 274
Md. Saddam Hossain, 354
Melanie C. Estrada, 38
Michael Slyuzberg, 409
Mohaimen Mansur, 354
Mominul Haque Mondol, 338
Moua Golalizadeh, 418
Muchtar Abdul Kholiq, 363
Muntaha Mushfiquee, 319
Mustapha Ziroili, 292

N

Ni Wendong, 156

O

Ordak Michal, 23

P

Paulo Jorge Gomes, 45
Peng Ding, 249

Q

Quindale E. Caraos, 31

R

Rabeh Morrar, 217
Rahul Mukherjee, 249

S

Saeed Fayyaz, 169
Sagaren Pillay, 15
Sahda Ratnasari, 226
Sarah B. Balagbis, 31, 38

Index

Selamawit Moja, 70
Siti Muchlisoh, 266
Song Xue, 108
Sumonkanti Das, 274, 284
Symala Krishnannair, 194
T
Teppei Ogihara, 134
Tianying Chu, 409
Tiffany Tan Shi En, 209
Titi Kanti Lestari, 363
Tsung-Jen Shen, 164
U
Ulrike Rockmann, 368
V
Valent Gigih Saputri, 345
Valerie Mercer-Blackman, 145
Varun Agiwal, 84
Venkataraman Sivakumar, 194
W
Wang Chun Zhi, 201
Wasimul Bari, 338
Willard Zvarevashe, 194
Y
Ye ZhiQing, 236
Youhua Chen, 164
Yu Jin, 156

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 978-967-2000-67-9

9 789672 000679

#ISIWSC2019