62nd ISI World Statistics Congress

ISI 2019 Kuala Lumpur

18 - 23 August 2019

PROCEEDING

CONTRIBUTED PAPER SESSION

VOLUME 3

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 3)

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 3, 2019. 444 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 3

Preface i

Scientific Programme Committee ii

Local Programme Committee iii

CPS1923: Test for exponentiality against renewal increasing mean residual life class 1

CPS1931: Every Policy Is Connected (EPIC): A generic tool for policy-data integration 8

CPS1934: Prediction of daily respiratory tract infection episodes from prescription data 16

CPS1941: A new model selection criterion for finite mixture models 25

CPS1943: Player selection strategy: A quantitative perspective 34

CPS1944: Time-lagged variables and incidence of pneumonia in wet-dry tropical North Australia 46

CPS1946: Assessment of condominium occupancy rate in Bangkok and its vicinity from electricity meter data analytics 56

CPS1947: Semi-parametric single-index predictive regression 65

CPS1951: Comparative analysis of R&D statistical systems between China and major developed countries 72

CPS1952: A statistical modelling framework for mapping malaria seasonality 80

CPS1954: Multiclass classification of growth curves using random change point model with heterogeneity in the random effects 89

CPS1955: Master's Programme in Data Analytics for Government: The UK Experience 97

CPS1956: Detecting life expectancy anomalies in England using a Bayesian Hierarchical Model 104

CPS1965: The modified Lee-Carter model with linearized cubic spline parameter approximation for Malaysian mortality data 111

CPS1970: The Lee-Carter Model: Extensions and applications to Malaysian mortality data 118

CPS1972: Multi-Aspect permutation methods for cytomorphometric data under multivariate directional alternatives with application to comparative neuroanatomy 125

CPS1973: Structural breaks in nonparametric models via atomic pursuit methods 134

CPS1979: Investigating dissimilarity in spatial area data using Bayesian inference: The case of voter participation in the Philippine national and local elections of 2016 142

CPS1982: Traditional and newly emerging data quality problems in countries with functioning Vital Statistics: Experience of the Human Mortality Database 150

CPS1983: Data analytics for better statistics 159

CPS1985: Trusted official smart statistics - challenges for official statistics in using data sources coming from private data producers 165

CPS1988: Unreliable retrial queue with two types customers, two delays and vacations perspective 173

CPS1994: Improving energy access for Africa through regional integration 181

CPS1996: The relationship between income inequality and disparity education and effect to achieve SDGs in Egypt 190

CPS1997: Integration of Economic Establishments Data into a uniquely identified comprehensive frame in Egypt 200

CPS1999: Effect of the ocean heat content on the global sea ice extent using fuzzy logic approach 209

CPS2002: The impact of weather risk on the estimation of yield-based agricultural losses and value at risk using Copula models 216

CPS2003: Structured additive regression modeling of Pulmonary Tuberculosis infection 225

CPS2010: Spatial and temporal trends in non- monetary wealth in Latin America (1990-2010) 234

CPS2011: Using SOM-based visualization to analyse the financial performance of consumer discretionary firms 241

CPS2012: Forecasting conditional covariance matrices in high-dimensional data: A general dynamic factor approach 251

CPS2016: Spatial multivariate outlier detection in the water quality of Klang River basin, Malaysia 258

CPS2018: Using Google trend data as an initial signal Indonesia unemployment rate 266

CPS2019: Comparing rainfall curves between climatological regions using functional analysis of variance 274

CPS2025: mpcmp: Mean-Parametrized Conway- Maxwell Poisson (Com-Poisson) regression 282

CPS2026: Trends in the extremes of environments associated with severe US thunderstorms, and signals in their spatial dependence 290

CPS2037: Comparing the household final consumption expenditure in national accounts to the household budget survey - or vice versa? 299

CPS2041: Contribution and growth of selected economic activities in the non-oil real GDP in the Emirate of Abu Dhabi 2007-2018 308

CPS2045: Let the PDEs guide you to new insight into and fast inference for complex models in space and space-time 314

CPS2052: An analysis of the contribution of women in Abu Dhabi 319

CPS2054: Epidemiology of acute kidney injury in critically ill patients in a South African intensive care unit 327

CPS2060: Monitoring unit root in sequentially observed autoregressive processes against local- to-unity hypotheses 335

CPS2061: An investigation into parametric and non-parametric modelling of LGD to estimate extreme percentiles of the loss distribution with respect to defaulted loans 342

CPS2064: Experimental statistics: A hub for data innovation in official statistics? 350

CPS2071: Robust estimation of treatment effects in a latent-variable framework 359

CPS2072: A new additive index number system with maximum characteristicity for International Price Comparisons 367

CPS2074: Estimating measurement errors in mixed-mode surveys using a Multitrait-Multierror Model 374

CPS2079: Improving statistical literacy in Albania, the role of the National Statistical Institute 383

CPS2081: Composite indicator of Food insufficiency 390

CPS2082: Monitoring population strategies in GCC: opportunities and challenges 398

CPS2083: Robust estimation of multi - input transfer function model with structural change 406

CPS2085: Competing risk analysis of lifetime data using Inverse Maxwell Distribution 413

CPS2087: Outlines of SCAD's pilot test for the 2020 register based census 420

CPS2088: Advances in maintenance of critical plant machinery equipment, frequency optimization and minimization of breakdowns perspective 427

Index 434

Test for exponentiality against renewal increasing mean residual life class

Deemat C Mathew[1]; Sudheesh K Kattumannil[2], Anisha P[3]
[1]St. Thomas College Palai, Kottayam, India
[2]Indian Statistical Institute, Chennai, India
[3]Institute of Public Health, United Arab Emirates University, UAE

Abstract

In this paper, we develop an exact test for testing exponentiality against renewal increasing mean residual life class. Asymptotic properties of the test statistic are studied. Numerical results are presented to demonstrate the performance of the testing method and we illustrate the test procedure using a real data.

Keywords

Exponential distribution; Renewal increasing mean residual life; U-statistics

1. Introduction

Due to its importance in the analysis of age replacement model, testing the null hypothesis that a lifetime is exponential (ageless) against the alternatives that it has decreasing (increasing) mean residual life (MRL) class has received considerable attention during the last two decades. Hollander and Proschan (1975) developed a test for exponentiality against DMRL alternative. Chen et al. (1983) extended Hollander and Proschans test to the case of randomly right censored data. Bergman and Klesfjo (1989) developed a family of test statistics for testing exponentiality against DMRL when the data is both complete and censored. For recent review of different test procedure we refer to Henze and Meintanis (2005).

If a device is experiencing a random number of shocks governed by a homogeneous Poisson process, then renewal increasing mean residual life is useful concept to study age replacement model. In this context testing exponentiality against renewal increasing mean residual life shock class can be used to determine whether to adopt a planned replacement model over unscheduled one.

Sepehrifar et al. (2015) developed a non-parametric test against RIMRL_{shock} class and obtained a critical region based on the asymptotic theory of U-statistics. Motivated by Sepehrifar et al. (2015), we develop an exact test for testing exponentiality against RIMRL_{shock} class. We also obtain the critical region of the asymptotical test proposed by Sepehrifar et al. (2015) and then find the Pitman's asymptotic efficacy to find the performance of the asymptotic test.

The rest of the paper is organized as follows. In Section 2, we propose an exact test for testing exponentiality against RIMRL_{shock} class and then calculate the critical values for different sample sizes. The asymptotic normality and the consistency of the proposed test statistic are proved in Section 3. The Pitman's asymptotic efficacy value is also given in this section. In Section 4, we report the result of the simulation study carried out to assess the performance of the proposed test. Finally, in Section 5 we give the conclusions of our study.

2. Exact Test

Let X be the lifetime of a device which has absolutely continuous distribution function $F(.)$. Suppose $\bar{F}(x) = P(X > x)$ denotes the survival function of X at x. Also let $\mu = E(X) = \int_0^\infty \bar{F}\,(t)dt < \infty$. Assume that the device under consideration is experiencing a random shock. Suppose $N(t)$ denotes the total number of shocks up to time t with probability mass function $P(N\,(t) = j) = F^j(t) - F^{j+1}(t), j = 0, 1, 2$ .... Suppose that the random variable $W_j, j = 0, 1, 2$, quantify the amount of hidden lifetime absorbed by the jth shock with $W_0 = 0$ and having common distribution function $G(x) = P(W_j \leq x)$. The total cumulative life damage up to time t is defined as $Z(t) = \sum_{j=0}^{N(t)} W_j$ with the cumulative distribution function $Q(x) = P(Z(t) \leq x) = \sum_{j=0}^{\infty} G^{(j)}\left[F^j(t) - F^{(j+1)}\,(t)\right]$. It is assumed that the unit fails when the total life-damage exceeds a pre-specified level $x > 0$. We refer to Glynn and Whitt (1993), Roginsky (1994) and Sepehrifar et al. (2015) for discussion related this framework. Let $X^* = X - Z\,(t)$ be the residual lifetime of an operating device with cumulative damage $Z(t)$. Note that the realizations of X^* is available to us for further analysis. Consider a device subjected to $N(t)$ number of shocks up to time t. Given that such a device is in an operating situation at time instant t after installation, the MRL function of X^* denoted by $m^*(t)$ is defined by $m^*(t) = E(X^* - t|X^* \geq t)$. Note that the total life-damage will not exceed the threshold level x. From the definitions it is evident that the random variables X^* and $Z(t)$ are independent.

Definition 2.1. *The mean residual life of a device under shock model* (MRL_{shock}) *at time t is defined as*

$$m^*(t) = \frac{1}{\bar{r}(t)} \int_t^\infty \bar{r}(z)\,dz,$$

where $\bar{r}(z) = \int_0^x \bar{F}\,(z + w)dQ(w)$.

Definition 2.2. *The random variable* X *belongs to the* $RIMRL_{shock}$ *class if the function* $m^*(t)$ *is a non-decreasing function for all* $t > 0$.

Next, we develop a family of test for testing exponentiality against $RIMRL_{shock}$. We are interested to test the null hypothesis

$$H_0 : F^* \text{ is exponential}$$

against the alternatives

$$H_1 : F^* \text{ is } RIMRL_{shock} \text{ (and not exponential).}$$

on the basis of a random sample $X_1^*, X_2^*, \dots, X_n^*$: from an absolutely continuous distribution function F^*.

For the above testing problem Sepehrifar et al. (2015) proposed a non-parametric test based on the departure measure $\Delta^*(F^*)$ defined by

$$\Delta^*(F^*) = \frac{1}{\mu^*} E_{f*}\left(min(X_1^*, X_2^*) - \frac{1}{2}X_1^*\right) = \frac{\Delta(F^*)}{\mu^*}$$

where $\Delta(F^*) = E_{f*}\left(min(X_1^*, X_2^*) - \frac{1}{2}X_1^*\right)$ and $\mu^* = E(X_1^*)$. Based on U-statistics theory Sepehrifar et al. (2015) obtained the following test statistic

$$\hat{\Delta}^* = \frac{\hat{\Delta}}{\bar{X}^*}, \qquad (1)$$

where $\bar{X}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$ and $\hat{\Delta} = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i;j=1}^{n} h(X_i^*, X_j^*)$ with $h(X_1^*, X_2^*) = \min(X_1^*, X_2^*) - \frac{1}{2}X_1^*$. Hence the test procedure is to reject the null hypothesis H_0 in favour of H_1 for large values of $\hat{\Delta}^*$.

Next we develop an exact test based on the test statistics $\hat{\Delta}^*$ and then calculate the critical values for different sample size. We use a result due to Box (1954) to find the exact null distribution of the test statistic.

Theorem 2.1. Let X^* *be continuous non-negative random variable with* $F^*(x) = e^{-\frac{x}{2}}$. *Let* $X_1^*, X_2^*, \dots X_n^*$ *be independent and identical samples from* F^*. *Then for fixed* n

$$P(\hat{\Delta}^* > x) = \sum_{i=1}^{n} \prod_{j=1, j\neq i}^{n} \left(\frac{d_{i,n} - x}{d_{i,n} - d_{j,n}}\right) I(x, d_{i,n}),$$

provided $d_{i,n} \neq d_{j,n}$, *for* $i \neq j$, *where* $I(x,y) = \begin{cases} 1 & if \quad x \leq y \\ 0 & if \quad x > y \end{cases}$ *and* $d_{i,n} = \frac{(n-2i+1)}{2(n-1)}$.

The critical values of the exact test for different n are tabulated in Table 1.

Table 1. Critical values of the exact test

n	90% level	95% level	97.5% level	99% level
2	0.4000	0.4500	0.4750	0.4900
3	0.2764	0.3419	0.3883	0.4292
4	0.2189	0.2678	0.323	0.3693
5	0.1883	0.2383	0.28	0.325
6	0.1679	0.2131	0.2508	0.2927
8	0.1413	0.1799	0.2125	0.2492
10	0.1243	0.1586	0.1877	0.2208
20	0.0852	0.109	0.1295	0.1531
30	0.0689	0.0882	0.1049	0.1241
50	0.0529	0.0679	0.0808	0.0957
100	0.0373	0.0477	0.0569	0.0675

3. Asymptotic properties

In this section, we investigate the asymptotic properties of the proposed test statistic. Making use of the asymptotic distribution we also calculate the Pitman's asymptotic efficacy.

3.1. Consistency and asymptotic normality. As the proposed test statistic is a U-statistic, we use the asymptotic theory of U-statistics (Lee, 1990) to discuss the limiting behaviour of $\hat{\Delta}^*$.

Theorem 3.1. *The* $\hat{\Delta}^*$ *is a consistent estimator of* $\Delta^*(F^*)$ under the alternatives H_1.

Theorem 3.2. *The distribution of* $\sqrt{n}(\hat{\Delta} - \Delta(F^*))$*, as* $n \to \infty$*, is Gaussian with mean zero and variance* $4\sigma_1^2$*, where* σ_1^2 *is the asymptotic variance of* $\hat{\Delta}$ *and is given by*

$$\sigma_1^2 = \frac{1}{4} V\,ar(2X^*\bar{F}^*(X^*) + 2\int_0^{X^*} ydF^*(y) - \frac{1}{2}X^*) \tag{2}$$

Proof: Since $\hat{\Delta}$ is a U-statistic it is a consistent estimator of $\Delta(F^*)$ (Lehmann,1951). Hence $\hat{\Delta}$ converges in probability to $\Delta(F^*)$. Note that $\bar{X}^*$ converges in probability to μ^*. As we can write

$$\hat{\Delta}^* = \frac{\hat{\Delta}}{\Delta(F^*)} \cdot \frac{\Delta(F^*)}{\mu^*} \cdot \frac{\mu^*}{\bar{X}^*},$$

the proof of the theorem is immediate.

Corollary 3.1. *Let* X^* *be continuous non-negative random variable with* $\bar{F}^*(x) = e^{-\frac{x}{\lambda}}$*, then the* distribution of $\sqrt{n}(\hat{\Delta}^* - \Delta^*(F^*))$, as $n \to \infty$, *is Gaussian with mean zero and variance* $\sigma_0^2 = \frac{\lambda^2}{12}$.

Corollary 3.2. *Let X^* be continuous non-negative random variable with $\bar{F}^*(x) = e^{-\frac{x}{\lambda}}$ then the distribution of $\sqrt{n}(\hat{\Delta}^* - \Delta^*(F^*))$, as $n \to \infty$, is Gaussian with mean zero and variance $\sigma_0^2 = \frac{1}{12}$.*

Apart from the exact test we can construct an asymptotic test based on the asymptotic distribution of $\hat{\Delta}^*$. Hence in case of the asymptotic test, for large values of n, we reject the null hypothesis H_0 in favour of the alternative hypothesis H_1, if

$$\sqrt{12n}\left(\hat{\Delta}^*\right) > Z_{\propto},$$

where $Z_{\propto}$, is the upper a-percentile of $N(0,1)$.

3.2. Pitman's asymptotic efficacy. In our case, the Pitman's asymptotic efficacy (PAE) is given by

$$PAE\left(\Delta^*(F^*)\right) = \frac{\left|\frac{d}{d\lambda}\,\Delta^*\,(F^*)\right|\lambda \to \lambda_0}{\sigma_0} = \sqrt{12}(W'\,(\lambda_0)\; -\; W(\lambda_0)\mu_a^{*\prime}\,(\lambda_o)),$$

where $W = E(min(X_1^*, X_2^*))$ and μ_a^* is the mean of X^* under the alternative hypothesis and the prime denotes the differentiation with respect to λ. We calculate the PAE value for three commonly used alternatives which are the members of $\mathrm{RIMRL}_{\mathrm{shock}}$ class

Next we compare the performance of the proposed test with some other tests available in the context of age replace model. The Table 2 gives the PAE values for different test procedures. From the Table 2, it is clear that our test is quite efficient for the Weibull and linear failure rate alternatives.

Table 2. Pitman's asymptotic efficacy (PAE)

Distribution	Proposed test	Li and Xu (2008)	Kayid et al. (2013)
Weibull	1.2005	1.1215	0.4822
Linear failure rate	0.8660	0.5032	0.4564
Makeham	0.2828	0.2414	2.084

Table 3. Empirical type 1 error of the test

n	Type 1 Error (5% level)	Type 1 Error (1% level)
10	0.0635	0.0123
20	0.0540	0.0115
30	0.0518	0.0107
60	0.0516	0.0102
80	0.0511	0.0100
100	0.0504	0.0101

4. Simulation and data analysis

Next, we report a simulation study for evaluating the performance of our asymptotic test against various alternatives. The simulation was done using R program. Finally, we illustrate our test procedure using a real data.

4.1. Monte carlo study. First we find the empirical type 1 error of the proposed test. we simulate random sample from the exponential distribution with cumulative distribution function $F(x) = 1 - \exp(-x), x \geq 0$. Since the test is scale invariant, we can take the scale parameter to be unity, while performing the simulations. The empirical type 1 error for different values of n and is reported in Table 3. From the Table 3 it evident that the empirical type 1 error is a very good estimator of the size of the test even for small sample size.

For finding empirical power against different alternatives, we simulate obse,rvations from the Weibull, linear failure rate and Makeham distributions with various values of λ where the distribution functions were given in the Section 3. As pointed out earlier these are typical members of the RIMRL_{shock}class. The empirical powers for the above mentioned alternatives are given in Tables 4, 5 and 6. From these tables we can see that empirical powers of the test approaches one when the λ values are going away from the null hypothesis value as well as when n takes large values.

Table 4. Empirical Power: Weibull distribution

	$\lambda = 1.2$		$\lambda = 1.4$		$\lambda = 1.6$		$\lambda = 1.8$	
n	5% level	1% level	5% level	1% level	5% level	1% level	5% level	1% level
60	0.50	0.23	0.93	0.76	0.99	0.97	1.00	0.99
70	0.55	0.27	0.96	0.84	0.99	0.99	1.00	1.00
80	0.60	0.31	0.98	0.89	0.99	0.99	1.00	1.00
100	0.69	0.41	0.99	0.95	1.00	0.99	1.00	1.00

5. Conclusion

Testing exponentiality against RIMRL_{shock} class enables reliability engineers to decide whether to adopt a planned replacement policy over unscheduled one. To address this issue, an exact test for exponentiality against RIMRL_{shock} class was introduced. We obtained the critical values for different sample sizes. Using the asymptotic theory of U-statistics, we showed that the test statistic was consistent and has limiting normal distribution. Making use of asymptotic distribution we obtained the PAE values and this shows that our asymptotic test has high efficiency for some of the well-known alternatives.

Table 5. Empirical Power: Linear failure rate distribution

	$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
n	5% level	1% level	5% level	1% level	5% level	1% level	5% level	1% level
60	0.49	0.22	0.68	0.38	0.79	0.51	0.87	0.65
70	0.55	0.27	0.74	0.46	0.84	0.61	0.92	0.74
80	0.60	0.32	0.80	0.53	0.89	0.68	0.94	0.81
100	0.69	0.41	0.87	0.65	0.94	0.80	0.98	0.90

Table 6. Empirical Power: Makeham distribution

	$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
n	5% level	1% level	5% level	1% level	5% level	1% level	5% level	1% level
60	0.37	0.14	0.49	0.22	o.65	0.36	0.87	0.63
70	0.42	0.17	0.55	0.26	o.72	0.43	0.92	0.72
80	0.46	0.20	0.60	0.31	0.78	0.49	0.94	0.79
100	0.55	0.27	0.70	0.40	0.86	0.62	0.98	0.90

References

1. Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification, Annals of Mathematical Statistics, 25, 290-302.
2. Chen, Y.Y., Hollander, M. and Langberg, N.A. (1983). Tests for monotone mean residual life, using randomly censored data, Biometrika, 39, 119-127.
3. Glynn, P.W. and Whitt, W. (1993). Limit theorems for cumulative processes, Stochastic Processes and their Applications, 47, 299-314
4. Hollander, M., Proschan, F., (1975). Tests for mean residual life, Biometrika, 62, 585-593.
5. Henze, N. and Meintanis, S.G. (2005). Recent and classical tests for exponentiality: a partial review with comparisons, Metrika, 61, 29-45.
6. Kayid, M., Ahmad, I.A., Izadkhah S. and Abouammoh, A.M. (2013). Further results involving the mean time to failure order, and the decreasing mean time to failure class, IEEE Transactions on Reliability, 62, 670-678.
7. Lee, A.J. (1990). U-Statistics, Marcel Dekker Inc., New York.
8. Lehmann, E.L. (1951). Consistency and unbiasedness of certain nonparametric tests, Annals of Mathematical Statistics, 22, 165-179.
9. Li, X and Xu, M. (2008). Reversed hazard rate order of equilibrium distributions and a related ageing notion, Statistical Papers, 49,749-767.
10. Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. Technometrics, 5, 375-383.
11. Roginsky, A.L (1994). A central limit theorem for cumulative processes, Advances in Applied Probability, 26, 104-121.
12. Sepehrifar, M.B, Khorshidian, K. and Jamshidian, A.R. (2015). On renewal increasing mean residual life distributions: An age replacement model with hypothesis testing application, Statistics and Probability, 96, 117-122.

Every Policy Is Connected (EPIC): A generic tool for policy-data integration

Arman Bidarbakhtnia, Christopher Ryan, Sharita Serrao
United Nations ESCAP

Abstract

The lack of data is blamed forthe absence of appropriate policies due to insufficient evidence, whilst the lack of demand is seen as the main challenge for producing relevant data. The root cause analysis begins with "lack of demand" or "lack of supply" for data, depending on which the analyst is - the data producer or the policymaker. This paper is introducing features and application of a tool called EPIC (Every Policy Is Connected) that facilitates policy-data dialogue aiming to break this vicious cycle by identifying policy priorities as well as data needs. The tool is integrating all four development dimensions (Economic, Environmental, Institutional and Social) in every policy plan and develops a comprehensive indicator framework for policy monitoring. The outs from the application of EPIC are two types; data and disaggregation needs (for immediate action), and policy formulation recommendations and indicator development (for future considerations).

Keywords

Indicator framework; sustainable development; policy-data integration; disaggregation; vulnerable groups

1. Background

Lack of data is often blamed for the absence of appropriate policies due to insufficient evidence, whilst the lack of demand is seen as the main challenge for producing relevant data. It is a case of the chicken or the egg dilemma. The root cause analysis begins with "lack of demand" or "lack of supply" for data, depending on which the analyst is - the data producer or the policymaker. Bidarbakht-Nia (2018) identifies silo mentality in policy formulation and monitoring process as the main bottleneck in breaking this policy-data vicious cycle and proposes a structured, principle-based and participatory user-producer engagement to integrate policy and data production processes. Three major issues that can be addressed by such policy data integration are (i) identifying (and creating) clear demand for (disaggregated) statistics for policy monitoring and evaluation, (ii) establishing interlink ages between four development pillars (Economic, Environmental, Institutional, and Social) at the planning as well as data production and dissemination levels, and (iii) enhancing inclusiveness of development plans.

Addressing above three issues requires regular and active engagement between producers and users of official statistics. However, it is not common that data producers actively participate in policy discussions to understand where the evidence for policymaking is missing, and policymakers often fail to specify what data and at what level of disaggregation are needed for monitoring sectorial and national policies. Advocating for user-producer dialogue and evidence-based decision is not a new topic (Heine K & Oltmanns E, 2016; Vardon M et al, 2016). However, the efforts are focused on making use of available data. There has been very little or no effort in identifying issues or target groups neglected by the policy that, in principle, must be addressed/targeted (Heine K & Mause K, 2004), and data currently being produced but neither demanded nor useful for any policy formulation/monitoring (Jules M, 2017).

Therefore, two characteristics are necessary for any effective framework for facilitation of user producer dialogue: a set of principles on which all parties can agree up on, and identification of issues to be addressed by and all target groups to be affected by (benefit from) policies. The EPIC (Every Policy Is Connected) is developed based on these two building blocks to facilitate a principle-based and participatory engagement of policy makers and data producers for effective "monitoring' of "inclusive" policies.

Structured engagement between data producers and users at the national level is critical to address the above problem. National statistical offices need to engage with national planning agencies, line ministries and other relevant national agencies to understand data and information needs, so that the official statistics that are ultimately produced are adequately responsive to policy needs and demands.

While the need for coordination and engagement between data producers and users is not a new discussion and has been registered time and again by the international statistical and policy communities, what has been lacking, or at least partially lacking, is specific guidance on how to promote and operationalize such engagement and collaboration in a practical sense. EPIC (Every Policy Is Connected), is a tool that attempts to bring practical guidance on systematic, purpose-driven engagement between users and producers of data. The participatory process that it promotes not only helps in identifying and streamlining data needs, identifying data gaps and even data waste, but also in reviewing and reformulating national comprehensive and sectorial policies and plans by engaging all relevant stakeholders at the national level.

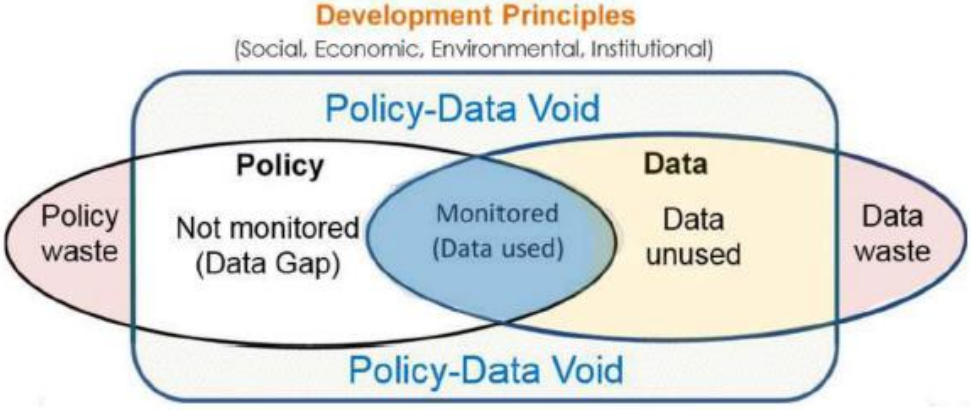
The tool guides the identification of priority population groups, issues and needs as stated in existing national development policies or plans; uses this information to identify data needs, including disaggregation requirements; and thereby works towards the development and/or strengthening of

monitoring/indicator frameworks for the specific national development policy or plan.

2. Conceptual framework

Inclusiveness and sustainability are symbiotic dimensions of development. One cannot be achieved without another. In 2015, the world leaders committed to "reach the furthest behind first", when signed the 2030 development agenda[1] at the general assembly of the United Nations. This means that all national and sectoral policies endeavour to be inclusive of needs and priorities of vulnerable groups that are most likely to miss the development train. This ambition cannot be achieved without a paradigm shift in planning, monitoring and evaluation process. Vulnerable groups of population are not easily identifiable by single-dimensionally developed policies. The same way that development pillars (economic, environmental, institutional, and social) are interlinked, vulnerability is often a result of various deprivations that cut across the four dimensions. Therefore, demand for and use of evidence (produced from disaggregated statistics) on the situation if vulnerable groups arises from deep understanding of the issues pertaining those groups and reflect them in the policy documents. This is the first step in defining what has to be measured and where. In other words, identifying issues that require Issues for Action and beneficiary target groups are primary to indicators and disaggregation requirements in the policy planning and evaluation process. To understand policy-data dynamic better, Bidarbakht-Nia (2018) proposes a framework within which policy and data interact and defines a knowledge space that can be expanded by linking policy with data (figure 1).

Figure 1- Policy-Data Space diagram



In all four dimensions of development, there are agreed principles embedded in international conventions ratified by the UN Member States.

[1] http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

Those principles are bedrocks of national policies that are reflected in the international conventions. At the same time, national statistical systems are responsible to inform such policies and naturally should be guided by the same underlying principles. When policy-data (or user-producer) dialogues are taking place on the basis of only endorsed policies (rather than agreed principles), it is likely that dialogues focus only on monitored policies (data used) and "data gap" for monitoring existing policies. In a less realistic situation, such dialogue may also focus on data and statistics that are produced by not being used (or useful) for policy monitoring (data unused). In all three cases, the focus of dialogue is on what we "know" about needs and used of data. While there are obvious cases of lack of policy that is consistent with agreed guiding principles; such as policies that explicitly assure equal opportunities for members of society (non-discrimination) as recommended by Universal Declaration of Human Rights[2]; or policies that are accounting for the impact of economic activities on ecosystem and people's healthy lives as recommended by Rio Declaration on Environment and Development[3]. It is often seen that policy-data dialogue is taking place around what is already reflected in the policy and hardly discuss what is NOT IN the policy. Ironically, most of the vulnerable groups and social, economic, environmental and institutional issues that relate to such groups are not acknowledged by the policy documents and therefore are left out of both policy and data (policy-data void). At the same time, statistical systems and planning organizations spend Millions of dollars annually on collecting data that are never being used or formulating policies that never been implemented (policy and data waste).

EPIC has been developed to help users and producers of data to expand their knowledge space from "data used" to void and waste in both policy and data by maximizing effective and structured policy-data interaction. In other words, to cut the waste, close the gap and fill the void in an integrated and participatory manner. The fundamental principle that is cornerstone in development of EPIC is that every tool that aims to successfully facilitate policy-data integration has to (a) focus on common interest of policy and data, and (b) benchmark against neither policy nor data, but a set of principles that are agreed up on by both data producers and decision makers. To achieve this objective, EPIC is designed to focus on Issues for Action and Target Groups as common interest of all stakeholders. Moreover, EPIC benchmarks the needs against a set of core concepts that cut across four development domains (Economic, Environment, Institutional, and Social), taken from internationally agreed frameworks, and naturally the expected outputs are both policy and data recommendations. Utilizing EPIC allows for a participatory process for

[2] http://www.un.org/en/universal-declaration-human-rights/index.html
[3] https://www.un.org/documents/ga/conf151/aconf15126-1annex1.htm

stakeholders to map policy onto data availability and enables articulation of new data requirements as well as an opportunity to strengthen content of policies. The tool is developed to identify existing unmet demand for data, lack of demand and potential demands for the future, and mismatch between data demand and supply.

3. Features of the tool

EPIC consists of three major components: Issues for Action and Target Groups; core concepts; and linking core concepts with Issues for Action and Target Groups to develop a national sustainable development indicator set. The Issues for Action, Target Groups and core concepts serve as inputs in the process, while the indicator set is in effect one of the key outputs that emerges from the process of policy-data integration and systematic user-producer engagement.

Issues for Action and Target Groups

While *Issues for Action* signify specific national or local concerns on which the policy or plan intends to act or make an intervention, the Issues for Action would make more sense when the policy connects it with a Target Group pointing out specifically for whom/what the action is being taken or who/what is likely to benefit from the action. It should be noted Target Groups are not just population groups e.g. women, children, unemployed, poor, families, households, etc. (covering the social dimension), but also enterprises, establishments, sectors etc. (economic dimension); oceans, mountains, freshwater, cities, forests, species, etc. (environmental dimension), as well as service providers, agencies, organizations, etc. (institutional dimension).

Core concepts

The second component, core concepts, are derived from existing international conventions and declarations ratified or adopted by UN member States such as such as the Universal Declaration of Human Rights (underlying principles for core concepts covering the social and institutional dimensions);[4] principles for inclusive economic growth (underlying principles for core concepts covering the economic core concepts);[5] and Rio 92 Declaration on Environment and Development (underlying principles for the environmental core concepts).[6] The tool has identified 29 core concepts and countries applying the tool are free to identify additional core concepts if considered relevant. The initial draft of core concepts was inspired by the list of core concepts in a tool called Equiframe which was developed to assess public

[4] https://www.un.org/en/udhrbook/pdf/udhr_booklet_en_web.pdf
[5] https://sustainabledevelopment.un.org/index.php?page=view&type=400&nr=893&menu=1561
[6] https://www.un.org/documents/ga/conf151/aconf15126-1annex1.htm.

health policies against core concepts of human rights and identifying vulnerable groups that, in principle, must be targeted by public health policies (Amin M et al, 2011).

The core concepts are a unique feature of the tool as they allow each issue for action and Target Group to be assessed for its potential to address the social, economic, environmental and institutional pillars of development, as appropriate and relevant. In effect, each issue for action and Target Group could link to more than one core concept. In the real world, it is a known fact that often sectoral policies are applied in isolation. Also, adequate linkages are often not established between SDG targets. This stove-pipe approach to policy formulation and implementation often leads to inadequate understanding and appreciation of the multiple and simultaneous disadvantages and vulnerability faced by population/Target Groups. Thus, policies often do not reach the most marginalized groups, which in turn also results in fragmented data collection without looking at simultaneous or nested disaggregation characteristics. Thus, the core concepts could serve as an aid to bring about better integration of Issues for Action and Target Groups across the various pillars of development, to address the principle of "leave-no-one-behind" in the policy-making and data collection processes.

In addition to establishing linkages across the social, economic, environment and institutional dimensions of development, the core concepts also broaden perspectives and provide insight on Issues for Action and Target Groups, which could potentially have been relevant for coverage in the policy document, but were missed out. For instance, while enhancing women's labour force participation could primarily be an issue for action addressing the economic pillar of development (as it could concern the issue of creating decent work for women; their protection for working across borders etc.), the policy may fail to realize that unless women's unpaid work burden is not reduced or redistributed at the intra-household level (social dimension), their physical working environment is not improved in certain occupations or work places (environmental dimension), or the legal infrastructure in the country is not conducive (institutional dimension), related policies may not reach those facing multiple forms of deprivation. Thus, the core concepts provide insight on what may be potentially important and relevant in the national context, but has been missed out of the policy. They provide useful inputs and guidance for policy review or reformulation in the future.

National sustainable development indicator set

Assessing each issue for action and Target Group from the angle of the various pillars for development allows for the development of a more comprehensive supporting indicator framework, as the indicators required measuring these various dimensions (economic, social, environment,

institutional) are likely to differ. Understanding Issues for Action and Target Groups helps define measurable factors (or parameters), their characteristics (i.e. variables), leading to the identification of standard measures to explain the state of change for policy monitoring (i.e. indicators). If the EPIC tool is applied consistently across all sectors in a country, this can help put together a "one-stop-shop" of sustainable development indicators that are directly responsive to national policy demands and which address data gaps and prevent data waste. This comprehensive indicator set is compiled by aligning Issues for Action and Target Groups with existing national, regional and global indicators, allowing for changes in indicator description or formulation of new indicators as needed, but with careful consideration of international standards on indicator development.

4. Implementation

Prior to implementation of the tool, it is necessary that a team of experts from all relevant stakeholders being established and made two fundamental decisions:

1- What are the major policy documents that need to be assessed to identify issues for policy actions and beneficiary target groups? At the minimum, national statistical office, planning organization, relevant sector/ministry, and national focal point for implementation of the Sustainable Development Goals have to involve.
2- Which components of the policy document have to assess against core concepts.

After the team made two above decisions, it is important to familiarize themselves basic concepts in the tool, review and practice application of the tool. Major steps in implementation of the tool involve the following:

A. In all selected components of the policy documents, identify issues that require policy action.
B. Map identified issues for action to all relevant "core concepts". Each issue may be mapped on more than one core concepts.
C. Identify target groups under each core concept. This may include target groups already specified in the document or groups that are not specified but the team agrees that are important to be targeted.
D. For each issue under each of the core concepts identify relevant indicator from existing indicator frameworks. It could be any of the national, regional, or global frameworks.

References

1. Amin, M., & MacLachlan, M., Mannan, H., El Tayeb, S., El Khatim, A., Swartz, L., Munthali, A., Van Rooy, G., McVeigh, J., Eide, A., Schneider, M., 2011. EquiFrame: A framework for analysis of the inclusion of human rights and vulnerable groups in health policies. Health & Human Rights: An International Journal; 13 (2).
2. Bidarbakht-Nia, A., 2018; *Policy-Data Integration: key to achieving the SDGs for all, UNESCAP, Working Paper Series (SD/WP/07/April 2017).*
3. Heine, K., Mause, K., 2004. Policy Advice as an Investment Problem. Kyklos, Vol 57 (2004), 403–428.
4. Heine, K., Oltmanns, E., 2016. Towards a political economy of statistics. Statistical Journal of the IAOS 32 (2016) 201–209.
5. Jules, M., 2017. The most underutilised source of data for smart cities. CitiesToday, 10th February 2017 (available at: https://citiestoday.com/industry/underutilised-source-data-smart-cities/)
6. Vardon, M., Burnett, P., Dovers, S., 2016. The accounting push and the policy pull: balancing environment and economic decisions. Ecological Economics 124 (2016) 145–152.

Prediction of daily respiratory tract infection episodes from prescription data

Atikur R. Khan[1], M. Towhidul Islam[2], Tabin Hasan[3], Saleheen Khan[4]
[1]Gulf University for Science and Technology, Kuwai
[2]Independent University of Bangladesh, Dhaka
[3]American International University - Bangladesh, Dhaka
[4]Minnesota State University, USA

Abstract

Changing weather pattern may directly or indirectly affect the incidence and severity of respiratory tract infections causing huge economic burden for healthcare services. Early warning for severity and extent of this infection may help healthcare service providers to prepare for any epidemic well before in time. Our aim in this paper is to explore the relationship between respiratory tract infection episodes and climatic factors and to predict the number of daily episodes in different weather zones defined by the coverage areas of active weather stations in Bangladesh. Prescription data collected from clinics are integrated with climatic factors of the nearest weather stations, and this integrated dataset is used to predict the daily respiratory tract infection episodes in response to climatic factors. We apply panel generalized linear models and show that the number of episodes increases in a greater extent for increasing magnitude of rolling standard deviation of relative humidity and rolling mean of wind speed. A 7-day ahead forecast of number of episodes based on rolling window models of regression tree and random forest are produced to know the severity of epidemic for healthcare planning, and a further 1-day ahead confirmation forecast is produced to assess the necessity of healthcare service plan adopted based on a 7-day ahead forecast. Root mean squared forecast error computed both for 7-day ahead and 1-day ahead forecasts both from regression tree and random forest provide qualitatively similar results, except for three weather stations where unusually high number of episodes are observed because of climatic extremes and high level of air pollution.

Keywords

Panel generalized linear model; prescription data; respiratory tract infection; regression tree; random forest

1. Introduction

Respiratory tract infections (RTI) are the most common infections worldwide causing a considerable economic burden to healthcare services. There are substantial evidences on the seasonal variation of respiratory

morbidity and mortality, which results in increased use of health services and hospital admissions. Respiratory disease incidence are correlated with climatic factors including temperature, relative humidity and other air pollution related consequences. According to WHO (2014), there were almost 7 million premature death in 2012 as consequences of air pollution and 88% of those deaths were in developing countries. Being a developing country, Bangladesh is achieving a very fast growth in economic and infrastructure development with the cost of increasing air pollution and climatic changes. Increasing air pollution and fluctuation in climatic patterns are likely to increase the RTI incidences with increasing burden in healthcare services. Thus it is important to know the severity of RTI episodes well before in time for healthcare planning and management. In this paper, we explore the relationship between RTI episodes extracted from millions of prescriptions and climatic factors obtained from 35 weather stations in Bangladesh, and predict the number of daily RTI episodes by using panel generalized linear models, regression tree and random forest.

Our study design based on the integration of prescription data and climatic factors from 35 weather zones defined by 35 active weather stations in Bangladesh is the first of its kind in a developing country context. Though several studies have been done in other countries with hospital records and climatic factors, most of those research have used weekly or monthly data without considering time lag relationship between daily RTI episodes and climatic factors. In a recent study in Shenmu County of China, Liu et al. (2016) have analyzed meteorological data and medical data from hospitalized patients less than 16 years of age and have found that the meteorological factors (air temperature, atmospheric pressure, rainfall, hours of sunlight, wind speed and relative humidity) are significantly associated with the lower respiratory tract infection (LRTI). Pearson correlation and multiple linear regression models have been used to explore the relationship between LRTI episodes and climatic factors. The LRTI does not happen instantly, there is a burn-in period since microorganism or viruses get into the body, and this burn-in period depends on the immune system and level of resistance of body. Thus a time lag relationship between climatic factors and disease incidence needs to be considered to explore the effect of climatic factors on LRTI episodes.

It is not only the level of temperature, oscillation of temperature is more sensitive to at-risk group (elderly, children, infants, and patients with medical conditions). Older adults and children are more vulnerable to daily temperature oscillations (Xu et al., 2012). An Australian study showed a correlation between sharp temperature change between two neighboring days and increased emergency visits for childhood pneumonia, and this effect can last up to 3 weeks (Xu, Hu and Tong, 2014). This findings also support to

apply a model that incorporates a time lag relationship between response and predictor variables.

Diurnal temperature change (difference between the maximum and minimum daily temperature) is a measure for variation in temperature change and has a greater impact on respiratory tract infection (RTI) episodes. Changes in temperature, precipitation, relative humidity, and air pollution influence viral activity and transmission and may contribute to the size and severity of the epidemics (Eccles, 2002; Mirsaeidi et al., 2016; Zhang et al., 2000). A good forecast on the severity of the epidemic is likely to provide good information for healthcare planning at least for at-risk group. To forecast weekly number of hospitalization in São Paulo city, Alencar (2018) applied generalized additive model with autoregressive terms (GAMAR) and Poisson distribution and demonstrated that inclusion of seasonal parameter in the model provides better forecasting performance than GAMAR with binomial distribution. However, these models do not consider time lag relationship between the response and predictor variables constructed from climatic factors. Furthermore, our dataset is in panel form with daily number of RTI episodes for 35 weather stations. Thus, we consider panel generalized linear models and machine learning methods for prediction of daily RTI episodes.

To organize the rest of our paper, we discuss the data integration procedure and computation for rolling time series statistics of climatic factors in Section 2. In Section 3, we provide our analytic results to demonstrate the effect of rolling time series statistics on RTI episodes by fitting panel generalized linear model and provide 7-day ahead forecast along with 1-day ahead confirmation forecast from regression tree and random forest models. In Section 4, we discuss on our overall findings and provide concluding remarks.

2. Methodology

Clinical data and climatic data are integrated to form data tables. Prescriptions collected from different clinics by 4P Ltd., a prescription audit and marketing company, over two years have been used in this study. Number of daily RTI episodes counted from these prescriptions is then integrated with climatic data obtained from 35 weather stations across the whole country. Daily episodes of diseases of a particular area are linked with the daily climatic time series from the weather station closest to that area. Our data integration procedure has been presented in Figure 1.

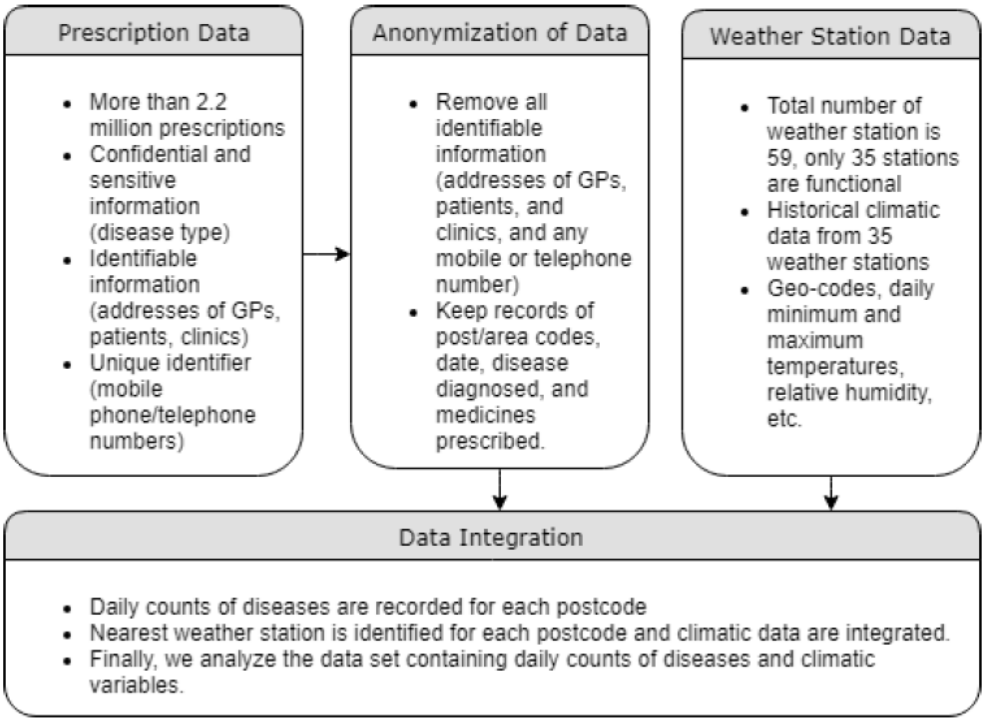


Figure 1. Data integration procedure

In this paper, we only consider daily episodes of respiratory tract infection (RTI) that includes both lower and upper respiratory tract infection, daily minimum and maximum temperatures, wind speed, sea level pressure, and relative humidity. Both MySQL quires and R routines have been used to process, analyse and visualize our data. We construct rolling time series statistics from these climatic variables to predict RTI episodes.

Let us compute rolling statistics for some climatic variables. If $x_i(t)$ is an instance of a climatic time series of the ith weather station at time (day) t, then $m-$lagged rolling mean and standard deviation can be computed as

$$\bar{x}_i(t|m) = \frac{1}{m}\sum_{j=1}^{m} x_i\,(t - m + j) \tag{1}$$

$$s_{x_i} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(x_i(t - m + j) - \bar{x}_i(t|m))^2} \tag{2}$$

Let $T_{max}(t)$ and $T_{min}(t)$ are maximum and minimum temperatures at time t. The difference between maximum and minimum temperatures, $d(t) = T_{max}(t) - T_{min}(t)$, indicates the amount of fluctuation in a day. Assuming that $d_i(t)$ is the difference between the maximum and minimum temperatures of weather station i on day t, we compute $m-$day rolling mean $\bar{d}_i(t|m)$ and standard deviation $s_{d_i}(t|m)$ by using Eq.(1) and Eq.(2). Similarly, for the ith

weather station, we compute rolling mean $\bar{h}_i(t|m)$ and rolling standard deviation $s_{h_i}(t|m)$ of relative humidity. In the next section, we explore underlying characteristics of $s_{d_i}(t|m)$ and $s_{h_i}(t|m)$ for $t = 1, \dots, T_i$ and $i = 1, \dots, M$ on RTI episodes where T_i is the length of the time series considered for the ith weather station and M is the number of weather stations.

3. Result

We use heatmap to explore the impact of rolling standard deviations of temperature difference and relative humidity on RTI episodes. Results shown in Figure 1 are test statistics computed for the null hypothesis $H_o: P = P_0$ where P_0 is the proportion of days in the dataset with positive counts for RTI episodes across all weather stations. Figure 1 clearly shows a region with rolling standard deviation of relative humidity greater than or equal to 6 with an indication that for this level of rolling standard deviation of relative humidity is likely to increase the RTI episodes.

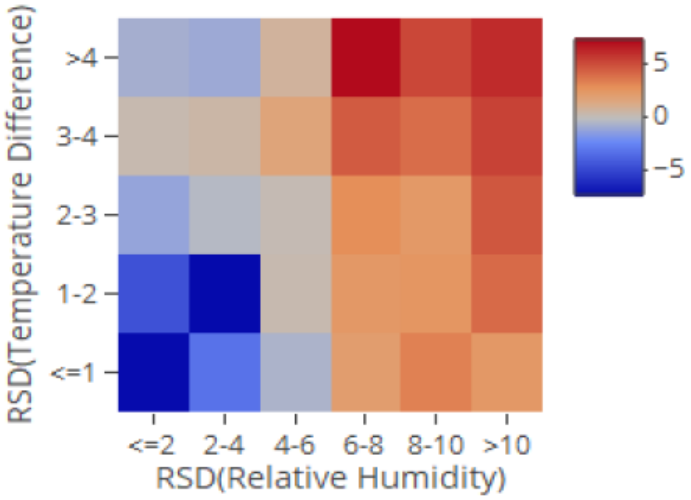


Figure 2. Heatmap for test statistics constructed from 8-day rolling standard deviation of temperature differences and relative humidity

Three different regions can be identified in Figure 2 as:

Category 1: This category emblems for higher degree of RTI episodes and the region under this category can be defined as $RSD(RH) > 6$, that is, $s_{h_i}(t|m) > 6$ with the notation used in Section 2.

Category 2: This can be classified as a category having medium level of RTI episodes and the region under this category is $RSD(RH) \leq 4 \cap RSD(TD) \leq 2$, that is, $s_{h_i}(t|m) \leq 4 \cap s_{d_i}(t|m) \leq 2$.

Category 3: This category is supposed to has a lower degree of RTI episodes and can be defined as $(4 < RSD(RH) \leq 6) \cup (RSD(RH) \leq 4 \cap RSD(TD) > 2)$, that is, this category can be defined based on defined rolling statistics as $(4 < s_{h_i}(t|m) \leq 6) \cup (s_{h_i}(t|m) \leq 4 \cap s_{d_i}(t|m) > 2)$.

Since RTI episode is a count response variable, we fit panel generalized linear models (PGLM) by using these categories as predictor variables along with rolling mean deviation of sea level pressure from normal level $RM(SP-1013.25)$ and rolling mean for wind speed $RM(WS)$. Results shown in Table 1 divulge that a negative binomial model for PGLM over a Poisson model is preferred based on the likelihood ratio test (LRT). The negative binomial model in PGLM reveals that the Category 2 and Category 3 of climatic condition is likely to exhibit almost 28% and 20% less RTI episodes compared to the climatic condition under Category 1.

We also note that for one unit increase in $(SP - 1013.25)$, RTI episodes are likely to increase by 1.6%. When $(SP - 1013.25) < 0$, low pressure in the sea causes rainfall that results in less dust particle in the air. Thus when $(SP - 1013.25) > 0$ there are less rainfall events and are likely to have more dust in the air. This little change may be due to regulation of dust and other particles in the air by rainfall. Further, one unit increase in $RM(WS)$, eight days rolling mean of wind speed, results in almost 16.51% increase in RTI episodes. This may be due to blowing more dust with increased wind speed, which is likely to affect people with dust allergies and other respiratory diseases related problems.

Table 1. Panel generalized linear model for RTI episodes

Variables	Poisson			Negative Binomial		
	β	$\exp(\beta)$	z − value	β	$\exp(\beta)$	z − value
Intercept	1.8452	6.3294	242.5375	1.7619	5.8235	34.8221
Category 2	-0.3319	0.7176	-45.0847	-0.3251	0.7225	-7.0642
Category 3	-0.2468	0.7813	-36.3982	-0.2163	0.8055	-4.8122
$RM(SP-1013.25)$	0.0245	1.0248	46.1541	0.0161	1.0162	5.5535
$RM(WS)$	0.1542	1.1667	77.3868	0.1427	1.1534	11.4618
Random effect: σ_u^2	1.1902			1.1941		
Log Likelihood	-24227.5800			-23003.8200		
AIC	48467.1600			46021.6300		
MSE	153.7109			153.6856		
LRT	--			2447.52		

Category 1: reference category. Here, $RM(SP-1013.25)$ and $RM(WS)$ are rolling mean deviation of sea level pressure from its normal level and rolling mean for wind speed, respectively.

We have already explored that the rolling time series statistics of climatic variables have significant effect on RTI episodes. Thus a forecasting exercise

of RTI episodes based on these climatic drivers may provide some insight regarding healthcare planning or for generating warning for at-risk patients. We use one year data to fit regression tree and random forest models (Choi et al., 2005; Lahouar & Slama, 2015) with predictor variables: rolling standard deviation of temperature difference, rolling standard deviation of relative humidity, $RM(SP - 1013.25)$ and $RM(WS)$. We fit rolling window model with 52 weeks data ($n = 364$), make only one forecast from each of the fitted model, compare the forecast with the original count of RTI and calculate mean squared forecast error (MSFE). Thus for $h-$step ahead forecast, we fit $T_i - n - h + 1$ rolling window models for the ith weather station, compute $T_i - n - h + 1$ forecasts and obtain MSFE values. Computed root mean squared forecast error (RMSFE) displayed in Figure 3 are square root of average MSFE computed from $T_i - n - h + 1$ forecasts for the ith weather station.

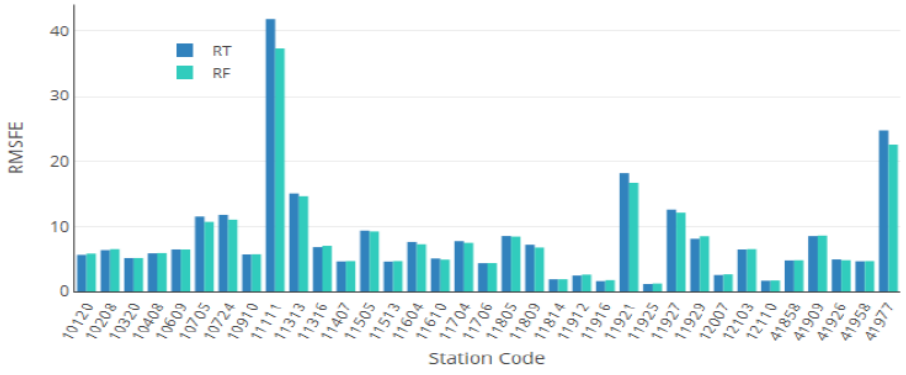


Figure 3. Root mean squared forecast error computed from regression tree (RT) and random forest (RF) models for 1-day ahead forecast

We find that both for 7-day ahead and 1-day ahead forecasts, RMSFE from both regression tree and random forest models are qualitatively similar. It is obvious that a 7-day ahead forecast will produce higher RMSFE than that of a 1-day ahead forecast. Though the RMSFE values are relatively low for most of the weather stations, weather stations with station codes 11111 (Dhaka City), 11921 (Chittagong), and 41977 (Ambagan, Chittagong) provide very high magnitude of RMSFE and the 11111 weather station produces unusually very high RMSFE both for 7-day ahead and 1-day ahead forecasts. Dhaka and Chittagong are two biggest and most air polluted cities in Bangladesh. Thus any changes in weather events affect these two cities much compared to other weather zones. Prediction of RTI episodes for these two cities requires further attention to explore underlying weather extremes and climatic factors.

4. Discussion and Conclusion

Prediction of daily number of RTI episodes provides an insight regarding the healthcare planning and early warning for an epidemic. Variations in climatic factors affect the residents directly or indirectly and may seriously affect the at-risk group. We have identified three different classes of variations in diurnal temperature change and relative humidity that are likely to increase the number of RTI episodes. Results obtained from panel generalized linear model demonstrate that these classes have significant impact on RTI episodes. Further, we apply flexible machine learning methods such as regression tree and random forest to obtain 7-day ahead forecast based on rolling statistics of climatic factors. This 7-day ahead forecast is suitable for planning healthcare services that may require in the following week, and a 1-day ahead forecast can be used to revise the planned healthcare services. We have also observed the higher magnitude of root mean squared forecast errors for highly air polluted cities which are likely to be a combined effect of air pollution and weather extremes, and a further research is required to investigate this phenomena.

References

1. Alencar, A. P. (2018). Seasonality of hospitalizations due to respiratory diseases: modelling serial correlation all we need is Poisson. Journal of Applied Statistics, 45 (10), 1813-1822.
2. Choi, Y., Ahn, H. & Chen, J. J. (2005). Regression trees for analysis of count data with extra Poisson variation. Computational Statistics & Data Analysis, 49(3), 893-915.
3. Eccles, R. (2002). An Explanation for the Seasonality of Acute Upper Respiratory Tract Viral Infections. Acta Otolaryngol, 122, 183-191.
4. Lahouar, A. & Slama, J. B. H. (2015). Day-ahead load forecast using random forest and expert input selection, Energy Conversion and Management, 103, 1040-1051.
5. Liu, Y., Liu, J., Chen, F., Shamsi, B. H., Wang, Q., Jiao, F., Qiao, Y. & Shi, Y. (2016). Impact of meteorological factors on lower respiratory tract infections in children. Journal of International Medical Research, 44(1): 30–41.
6. Mirsaeidi, M., Motahari, H., Khamesi, M. T., Sharif, A., Campos, M. & Schraufnagel, D. E. (2016). Climate change and respiratory infections. Annals of the American Thoracic Society, 13(8): 1223-1230.
7. WHO, 2014. Ambient (Outdoor) Air Quality and Health. World Health Organization. http://www.who.int/mediacentre/factsheets/fs313/en/.
8. Xu, Z., Etzel, R. A., Su, H., Huang, C., Guo, Y. & Tong, S. (2012). Impact of ambient temperature on children's health: a systematic review. Environmental Research, 117: 120-131.
9. Xu, Z., Hu, W., & Tong, S. (2014). Temperature variability and childhood pneumonia: an ecological study. Environmental Health, 13: 51.
10. Zhang, H., Triche, E. & Leaderer, B. (2000). Model for the analysis of binary time series of respiratory symptoms. American Journal of Epidemiology, 151(12): 1206-1215.

A new model selection criterion for finite mixture models

Jang Schiltz
University of Luxembourg, LSF, Luxembourg

Abstract

We present a generalization of Nagin's finite mixture model that allows non parallel trajectories for different values of covariates and illustrate its use by giving typical salary curves for the employees in the private sector in Luxembourg between 1981 and 2006, as a function of their gender, as well as of Luxembourg's gross domestic product (GDP). Afterwards, we propose a new model selection criterion for finite mixture models which is computationally easy and does not need a correction term for the number of parameters in the model.

Keywords

Statistical Models; Developmental trajectories; Trajectory Modeling; Model Selection

1. Introduction

Time series analysis is of the utmost importance for the research on various subjects in economics, finance, sociology, psychology, criminology and medicine and a host of statistical techniques have been developed to achieve it. In the 1990s, the modelization of the evolution of an age or time based phenomenon gave raise among other methods to latent growth curves modeling (Muthen 1989) and the nonparametric mixture model (Nagin 1999).

The nonparametric mixed model developed by Nagin (1985) is specifically designed to detect the presence of distinct subgroups among a set of trajectories. Compared to subjective classification methods, the nonparametric mixed model has the advantage of providing a formal framework for testing the existence of distinct groups of trajectories. This method does not assume a priori that there is necessarily more than one group in the population. Rather, an adjustment index is used to determine the number of sub-optimal groups. While the conceptual aim of the analysis is to identify clusters of individuals with similar trajectories, the model's estimated parameters are not the result of a cluster analysis but of maximum likelihood estimation (Nagin, 2005).

The remainder of this article is structured as follows. In section two, we present the basic version of Nagin's Finite mixture model, as well as one of his generalizations and we show two drawbacks of the model. In section three, we present a generalization of the model that overcomes these drawbacks and

we discuss model selection and group member probabilities for the new model. In section four, we highlight typical features of the new model by means of a data example from economics. In section five, finally, we introduce the classical criteria for determining the optimal number of trajectory groups in finite mixture models and propose a new criterion which does not need a lot of computer power to compute, even in case of large samples and does not depend on the number of parameters of the model.

2. Nagin's Finite Mixture Model

Starting from a collection of individual trajectories, the aim of Nagin's finite mixture model is to divide the population into a number of homogenous sub-populations and to estimate, at the same time, a typical trajectory for each sub-population (Nagin 2005). More, precisely, consider a population of size N and a variable of interest Y. Let $Y_i = y_{i1}, y_{i2}, \dots, y_{iT}$ be T measures of the variable Y, taken at times $t_1, \dots, t_T$ for subject number i. To estimate the parameters defining the shape of the trajectories, we need to fix the number r of desired subgroups. Denote the probability of a given subject to belong to group number j by π_j.

The objective is to estimate a set of parameters $\Omega = \{\pi_j, \beta_0^j, \beta_{01}^j, \dots; j = 1, \dots, r\}$ which allow to maximize the probability of the measured data. The particular form of Ω is distribution specific, but the β parameters always perform the basic function of defining the shapes of the trajectories. In Nagin's finite mixture model, the shapes of the trajectories are described by a polynomial function of age or time. In this paper, we suppose that the data follow a normal distribution. Assume that for a subject in group j

$$y_{it} = \sum_{k=1}^{s} \beta_k^j t_{it}^k + \varepsilon_{it} \tag{1}$$

where s denotes the order of the polynomial describing the trajectories in group j and ε_{it} is a disturbance assumed to be normally distributed with a zero mean and a constant standard deviation σ. If we denote the density of the standard centered normal law by ϕ and $\beta^j t_{it} = \sum_{k=1}^{s} \beta_k^j t_{it}^k$, the likelihood of the data is given by

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{t=1}^{T} \phi\left(\frac{y_{it} - \beta^j t_{it}}{\sigma}\right). \tag{2}$$

The disadvantage of the basic model is that the trajectories are static and

do not evolve in time. Thus, Nagin introduced several generalizations of his model in his book (Nagin 2005). Among others, he introduced a model allowing to add covariates to the trajectories. Let $z_1, \ldots, z_M$ be M covariates potentially influencing Y. We are then looking for trajectories

$$y_{it} = \sum_{k=0}^{s} \beta_k^j t_{it}^k + \alpha_1^j z_1 + \ldots + \alpha_M^j z_M + \varepsilon_{it} \quad (3)$$

where ε_{it} is normally distributed with zero mean and a constant standard deviation σ. The covariates z_m may depend or not upon time t. But even this generalized model still has two major drawbacks. First, the influence of the covariates in this model is unfortunately limited to the intercept of the trajectory. This implies that for different values of the covariates, the corresponding trajectories will always remain parallel by design, which does not necessarily correspond to reality.

Secondly, in Nagin's model, the standard deviation of the disturbance is the same for all the groups. That too is quite restrictive. One can easily imagine situations in which in some of the groups all individual are quite close to the mean trajectory of their group, whereas in other groups there is a much larger dispersion.

3. Our model

To address and overcome these two drawbacks, we propose the following generalization of Nagin's model. Let $x_1 \ldots x_M$ and $z_{i_1}, \ldots, z_{i_T}$ be covariates potentially influencing Y. Here the x variables are covariates not depending on time like gender or cohort membership in a multicohort longitudinal study and the z variable is a covariate depending on time like being employed or unemployed. They can of course also designate timedependent covariates not depending on the subjects of the data set which still influence the group trajectories, like GDP of a country in case of an analysis of salary trajectories. The trajectories in group j will then be written as

$$y_{it} = \sum_{k=0}^{s} \left(\beta_k^j + \sum_{m=1}^{M} \alpha_{km}^j x_m + \gamma_k^j z_{i_t} \right) + t_{it}^k + \varepsilon_{it}, \quad (4)$$

where the disturbance ε_{it} is normally distributed with mean zero and a standard deviation σ_j constant inside group j but different from one group to another. Since, for each group, this model is just a classical fixed effects model for panel data regression (see Woolridge 2002), it is well defined and we can get consistent estimates for the model parameters.

Our model allows obviously to overcome the drawbacks of Nagin's model. The standard deviation of the uncertainty can vary across groups and the trajectories depend in a nonlinear way on the covariates. Since our model is just a generalization of Nagin's finite mixture model, a lot of its main features and properties remain the same as in Nagin's model.

4. A data example

For the following example, we use Luxembourg administrative data originating from the General Inspectorate of Social Security, IGSS (Inspection gnrale de la scurit sociale). The data have previously been described and exploited with Nagin's basic model by Guigou, Lovat and Schiltz (2010, 2012). The file contains the salaries of all employees of the Luxembourg private sector who started their work in Luxembourg between 1980 and 1990 at an age of less than 30 years. This choice was made to eliminate people with a long career in another country before moving to Luxembourg. The main variables are the net annual taxable salary, measured in constant (2006 equivalent) euros, gender, age at first employment, residentship and nationality, sector of activity, marital status and the years of birth of the children. The file consists of 1303010 salary lines corresponding to 85049 employees, capped at 7577 EUR (2006 equivalent euros) per month.

We will not present here an exhaustive analysis of the whole dataset, but an illustration of the possibilities of our generalized mixture model and its differences from Nagin's model. We concentrate on the first 20 years of the careers of the employees who started working in Luxembourg in 1987, giving us a sample of 1716 employees. We compute typical salary trajectories for them, taking into account the gender of the employees, as well as their dependancy from the GDP of the country. Let us first highlight the differences with respect to Nagin's extended model.

Figure 1 shows a three group solution modeled by Nagin's generalized model representing the salary of employees in Luxembourg during the first 20 years of their professional career. We see that for the low salary group women and men are gaining exactly the same salary (with the consequence that there appears just one salary trajectory for the two lower salary groups on the graph instead of two) whereas in the middle and high salary groups, men earn more than women. Due to the limitations of the model, the evolution of the salaries seems to be exactly the same for men and women; their salary trajectories are strictly parallel.

Figure 1: Salary evolution by gender, modelized by Nagin's model

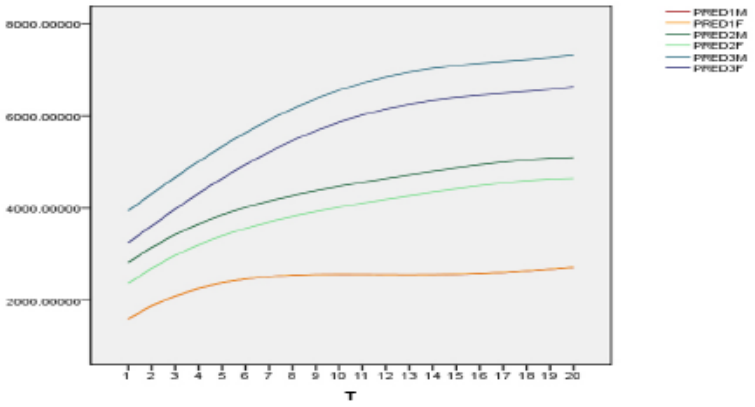
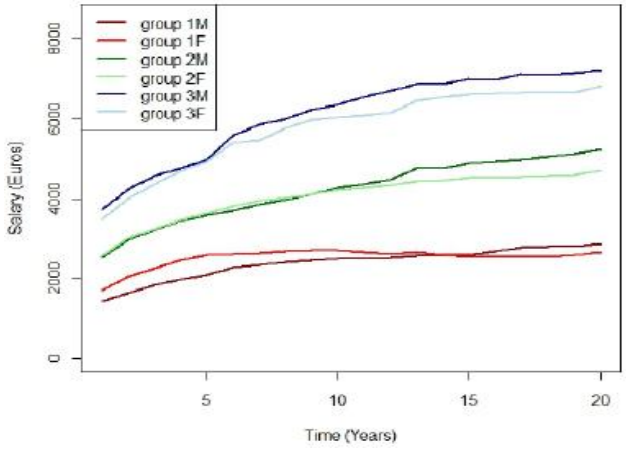


Figure 2 shows the three group solution for the 20 first year of Luxembourg employees calibrated with our model. We see a somewhat different and more realistic pattern emerging. For the high salary group the income of men and women remain more or less parallel, except for a short time interval around year five. This is however no longer the case for the middle and low salary groups. Here, we observe that the women in these groups have higher salaries then the men at the beginning of their career, but this is reversed somewhere in the middle and after 10 years for the middle salary group and 15 years for the low salary group the income of the men becomes higher than the one of the women.

Figure 2: Salary evolution by gender, modelized by our model



We obtained these results by calibrating the model

$$S_{it} = (\beta_0^j + \alpha_0^j x_i + \gamma_0 z_t) + (\beta_1^j + \alpha_1^j x_i + \gamma_1 z_t)t + (\beta_2^j + \alpha_2^j x_i + \gamma_2 z_t)t^2 \qquad (5)$$

where S denotes the salary, x the gender and z_t is Luxembourg's GDP in year $t - 1$ of the study. For figure 2, we replaced the variable z_t by the actual values of Luxembourg's GDP in the considered years. Table 1 shows the values of the parameters for a 3-group solution of model 8.

Table 1: Parameter estimates for model

Results for group 1

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	476.183	132.857	191.413	760.856
α_0	220.302	1.387	202.568	227.896
γ_0	0.582	0.071	0.407	0.710
β_1	206.446	27.850	146.632	266.084
α_1	123.219	4.909	121.582	126.895
γ_1	-0.077	0.007	-0.092	-0.062
β_2	-3.828	1.760	-7.602	-0.053
α_2	-8.922	0.1838	-9.089	-8.753
γ_2	0.002	0.001	0.002	0.003

Results for group 2

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	2243.017	236.843	1734.795	2750.771
α_0	-380.402	116.972	-636.957	-122.585
γ_0	0.180	0.011	-0.074	0.433
β_1	370.016	49.685	263.469	475.590
α_1	12.846	8.153	-41.197	66.703
γ_1	-0.049	0.012	-0.074	-0.023
β_2	-11.018	3.140	-17.741	-4.272
α_2	-1.491	0.755	-4.902	1.947
γ_2	0.002	0.001	0.001	0.003

Results for group 3

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	3293.714	335.402	2573.151	4011.944
α_0	-783.289	28.382	-892.997	-671.4
γ_0	0.189	0.025	-0.190	0.566
β_1	447.925	70.366	297.040	598.856
α_1	64.890	19.532	73.501	119.982
γ_1	0.036	0.017	-0.074	0.012
β_2	-13.873	4.447	-15.824	-9.174
α_2	-2.73	0.476	-4.196	-0.126
γ_2	0.001	0.001	0.000	0.002

The disturbance terms for the three groups are $\sigma_1 = 33.11$, $\sigma_2 = 54.18$ and $\sigma_3 = 78.85$ respectively. The dispersion is thus higher in the groups with higher salaries then in those with lower salaries. This makes sense, since in the low salary group a lot of employees just earn the minimal wage. Hence, a lot of them have the same salary.

Moreover this example illustrates the dependence of the trajectories on Luxembourg's GDP. We see that in the three groups, this influence is non linear, since γ_2 is always significantly different from 0. The trajectory equations from table 1 can now be used to predict the future evolution of the salaries for men and women as a function of GDP.

5. Model Selection in Finite Mixture Models

Till now, there has been no really satisfactory solution for a model selection procedure, in the sense of addressing the challenge to determine the optimal number of classes in a family of finite mixture models. Nagin (2005) emphasizes the need of an interplay of formal statistical criteria and subjective judgment and proposes to use the Bayesian Information Criterion (BIC), defined by

$$\mathrm{BIC} = \log(L) - 0,5k\log(N), \tag{6}$$

where k denotes the number of parameters in the model. The bigger the BIC, the better the model explains the data. The correction term $0.5k\log(N)$ is necessary, because the likelihood L is an increasing function of the number of groups and just taking the likelihood as criterion does hence not make any sense.

Nielsen et al. (2014) argue that the existing software does not always compute the BIC accurately and that furthermore BIC on its own does not always indicate a reasonable-seeming number of groups even when computed correctly. They propose the methodology of cross-validation error (CVE) instead, which consists in computing a CVE for each possible choice r of the number of latent classes, indicating the extent to which the model fails to perfectly model the data. The final choice of r is then the one that minimizes this CVE value. More precisely, CVE is defined by

$$CV\ E = \frac{1}{N}\sum_{i-1}^{N}\frac{1}{T}\sum_{t=1}^{T}\left|y_{i_t} - \hat{y}_{i_t}^{[-i]}\right|, \tag{7}$$

where $\hat{y}_{i_t}^{[-i]}$ it denotes the estimation of y_{i_t} obtained by running the model for the whole dataset, except line i. Apart from the fact that the numerical examples in Nielsen et al. (2014) do not really seem convincing, this method

is computationally extremely heavy, since it requires to execute the model $N(N-1)$ times, which can be quite problematic in the case of big data samples.

The criterion we propose also uses data to test the predictive power of the model. We actually propose to use the posterior probability $P(j/Y_i)$ of individual i's membership in group j that can be computed with the help of Bayes' theorem as

$$P(j/Y_i) = \frac{P(Y_i/j)\hat{\pi}_j}{\sum_{j=1}^{r} P(Y_i/j)\hat{\pi}_j}. \quad (8)$$

To determine the optimal number of groups r we maximize the posterior probability criterion (PPC) defined for $r > 1$ by

$$PPC = \sum_{i=1}^{N} (\max_{j=1,\ldots,r} P(j/Y_i)). \quad (9)$$

This actually means that we choose the number of groups that allows best to explain the predicted group membership of the people in the dataset. Having a PPC of N means that for all people in the dataset, it is almost sure to which group they belong. The big advantage of this criterion is that it gives a clear result and just requires a minimal number of computations. Moreover the criterion is always computed as a sum of N terms and does hence not depend on the number of parameters in the model. Therefore, there is no need for a correction term.

6. Conclusion

In this article, we presented Nagin's finite mixture model and some of its generalizations and showed some inherent shortcomings for possible applications. We addressed these by proposing a new generalized finite mixture model. A key characteristic is its ability to modelize nearly all kind of trajectories and to add covariates to the trajectories themselves in a nonlinear way.

We illustrated these possibilities through a data example about salary trajectories. We showed how to add a classical group membership predictor variable to the trajectories as well as a time series that does not depend on the subjects of the analysis but influences the shape of the trajectories in some of the groups.

Finally, we proposed a new methodology for determining the optimal number of groups in finite mixture models by introducing the posterior probability criterion.

References

1. Guigou, J.-D., Lovat, B., & Schiltz, J. (2010). The impact of ageing population on pay-as-you-go pension systems: The case of Luxembourg. Journal of International Finance and Economics, 10(1), 110{122.
2. Guigou, J.-D., Lovat, B., & Schiltz, J. (2012). Optimal mix of funded and unfunded pension systems: the case of Luxembourg. Pensions, 17(4), 208{222.
3. Jones, B.L., & Nagin, D.S. (2007). Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them. Sociological Methods & Research, 35(4), 542{571.
4. Muthen, B.O. (1989). Latent Variable Modeling in Heterogeneous Populations. Psychometrika, 54(4), 557{585.
5. Nagin, D.S. (1999). Analyzing Developmental Trajectories: Semi-parametric. Groupe-based Approach. Psychological Method, 4, 139{157.
6. Nagin, D.S. (2005). Group-Based Modeling of Development. Cambridge, MA: Harvard University Press.
7. Nielsen, J.D et al. (2014). Group-based criminal trajectory analysis and growth mixture modeling: A Monte Carlo simulation study. Communications in Statistics - Theory and Methods, 43(20), 4337{4356.
8. Schiltz, J (2015). A Generalization of Nagin's Finite Mixture Model. In: M. Stemmler, A. von Eye & W. Wiedermann. Dependent Data in Social Sciences Research. Heidelberg: Springer
9. Woolridge, J. (2002). Econometric Analysis of Cross-Section and Panel Data. Cambridge, MA: MIT press.

Player selection strategy: A quantitative perspective

Nandish Chattopadhyay[1], Prajamitra Bhuyan[2]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Department of Mathematics, Imperial College London, United Kingdom

Abstract
In most of the professional sporting leagues across the world, players are selected by team management and owners through various types of auctions. A ranked list of players is very useful for bidders to decide upon the preferences of possible buys. In this paper, we propose an efficient scoring system and ranking scheme of players based on performance and experience related attributes from the historical data. The proposed methodology maximally discriminates the players and incorporates both averages and the consistencies of the performance variables. The method is illustrated with real life data analysis and simulation study.

Keywords
Auctions; Consistency; Cricket; IPL; Rank

1. Introduction

Historically, professional sports have evolved into a diverse arena of social, economic and academic interest of research. It attracts huge fan following across geographical boundaries, and therefore allures massive investments. Professional sporting leagues take place throughout the year, all around the world and enjoy colossal viewership. Usually, these leagues are funded or run by franchises, which have different structures of ownership, and are in general conducted by the sports governing bodies, authorities and other related organizations.

A very fundamental aspect of these sports leagues in the process of building teams, which essentially makes the dynamics of these events quite different from the other mega-sports carnivals where teams are formed on the basis of national identity. The most prominent professional sports leagues include the football leagues of Europe namely the English Premier League in the UK, the La Liga in Spain, the basketball leagues like the NBA in the United States, and the cricket leagues like the Indian Premier League in India. In general, the teams participating in professional leagues are owned by clubs, companies or individuals on partnerships. These professional leagues are growing in popularity day by day. Therefore, building a good team is of prime interest for the club or franchise owners.

In most of these leagues, players are picked up by the franchises by auctions (open or closed bidding), where a player belongs to the highest bidder, and represents the team owned by that franchise for a stipulated amount of time. The rules and regulations of the auctions are set by the governing bodies of the respective leagues. A good understanding of the game and detailed knowledge of skill-sets, styles and strategies are essential for taking an informed decision during auction. However, any subjective perspective prove to be insufficient and deficient criteria in real-time decision making, especially with the increasing abundance of talented and hardworking sports-persons. In order to make an objective decision, one needs to rely on a ranking system, based on some statistical mechanism.

In the domain of professional leagues, varied research has been initiated for obtaining meaningful insight of its dynamics and functionalities in the recent past (Tingling, 2017; Schuckers, 2011). A lot of focus has been on the Indian Premier League, which is a cricketing tournament organized by the Board of Cricket Control in India, in a twenty-twenty format. Due to its novelty, dynamic nature with fresh auctions held every three years, and most importantly its massive viewership in India and abroad, the IPL entices colossal sponsorship and advertisements. Quite naturally, the investors are interested in maximizing their returns, which promotes team selection based on an objective decision making strategy.

It is worth noting that the existing research work related to the professional cricket leagues are based predominately on the hedonic pricing models (Karnik, 2010). For this purpose, Lenten et al. (2012) and Bhattacharya and Bhattacharya (2012) proposed to use regression model, considering the auction price of players as the response and performance related attributes as covariates. Note that, the aforementioned methodology makes an attempt to explain the valuation of the player purely on the performance related data, ignoring the intricacies of the auction procedure, which affects price. For example, the pricing model does not incorporate the sequential nature of the auction process and the valuation of the player is conditional on the events that have already unfolded. Therefore, it is not wise to model the valuation of a player based on performance attributes without considering the dynamic auction mechanism. However, the same performance related data can be effectively used to score and rank individual players, which can be utilized for the development of a sequential pricing model. Croucher (2000) proposed a simple performance index defined as a product of performance related variables of an individual player. Also Saikia et al. (2012), proposed a player rating scheme to compare the performances of the players in international cricket and the IPL where they essentially developed a rating scheme of the players using a weighted average of various performance variables, with the weights varying inversely as the variation in the respective variables. However,

these methodologies fail to ensure that the players are maximally distinguished and provide an easy selection strategy among close competitor.

In this paper, we propose a methodology that maximally discriminates the individual players, overcoming the shortcoming of the existing methods. The players can be easily ranked on the basis of their averages, ignoring the consistency (variability), of the performance variables under consideration. This mechanism is justifiable when all the players are equally consistent with respect to all the performance variables. Likewise, the players could be rated according to their consistencies when they are indistinguishable with respect to their average performance. However in reality, there is dissimilarity among the players, with respect to both averages and consistencies of the performance variable. Moreover, multiple performance variables may be considered to develop an efficient rating mechanism. In such cases, maintaining the trade-off that exists between the averages of the performances of the players and their respective variations is important but difficult. Furthermore, there could be significant correlations between some of the performance variables. This issue is not appropriately dealt with in any of the subjective or objective methods available in the literature. Our proposed methodology attempts to address all these aforementioned issues. We propose a new methodology for the player rating scheme in Section 2. In Section 3, we present the real-life data analysis on the performance of players in the Indian Premier League, considering both batsmen and bowlers. In Section 4, we present a simulation study of our proposed methodology. We end with some concluding remarks in Section 5.

2. Modeling Methodology

In order to develop an effective player rating scheme, let us consider that we have a total of n players, each player has played k_i matches for $i = 1, \ldots, n$. Let us define $X_{ij}^{(p)}$ to be the value corresponding to the performance of the i-th player, in the j-th match, for the p-th performance variable, for $p = 1, \ldots, m$. A typical data structure is represented in Table 1. In order to rank the individual players by associating a score with each of them, we will take a weighted average of the performance variables. Let us denote Yij as the score of the i-th player in the j-th match, which is formally written as:

$$Y_{ij} = \sum_{p=1}^{m} l_p Z_{ij}^{(p)}$$

where l_p is the weight corresponding to the p-th performance variable, and $Z_{ij}^{(p)} = \frac{X_{ij}^{(p)} - min_{ij} X_{ij}^{(p)}}{max_{ij} X_{ij}^{(p)} - min_{ij} X_{ij}^{(p)}}$ is the normalized value of the p-th performance variable for the i-th player in the j-th match. Note that the aforementioned normalization of variables is only for the sake of easy interpretation.

There are two major aspects that one needs to consider. Firstly from a pool of available attributes, one has to choose variables which are logically important, and pertain to non-overlapping information. Secondly the individual weights of the attributes are to be estimated in such a way which would maximally differentiate the players. Therefore, it is necessary that the estimated weights maximize the ratio of the variation of performances among different players to the variation of performances of individual players, i.e., the ratio of between-group variation to the within group variation considering the performances of a particular player as a group. This can be formally written as the following optimization problem:

$$(\hat{l}_1,\ldots,\hat{l}_m) = \underset{(l_1,\ldots,l_m)\in(-\infty,\infty)^m}{\operatorname{argmax}} \frac{\sum_{i=1}^{n} k_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(y_{ij}-\bar{y}_i)^2}, \quad (1)$$

and it has a closed form solution $(\hat{l}_1,\ldots,\hat{l}_m) = \hat{e}_1$, where $\hat{e}_1$ is the eigen-vector corresponding to the greatest among the non-zero eigen-values of $W^{-1}B, B = \sum_{i=1}^{n} k_i(\bar{y}_i - \bar{y})^2$ and $W = \sum_{i=1}^{n}\sum_{j=1}^{k_i}(y_{ij} - \bar{y}_i)^2$. See Johnson and Wichern (2007, p-610) for more details.

Note that the solution to this unconstrained optimization problem may lead to some of the weights having a negative value which may not be the appropriate sign in many cases depending upon the nature of the associated variable. Though that would maximally discriminate the players, it is not guarantee that the sign is appropriate for the highest score is associated with the best player. Since we are interested in obtaining ranks using the scores, our purpose would be defeated. Hence, it is necessary to add some constraints that ensure that all the weights are positive, for all those variables that positively impact the player's abilities. Similarly, for those variables that negatively impact the player's abilities, the corresponding weights should be negative. Without loss of generality, we consider that all the performance variables under consideration have positive impact on the player's abilities for further discussion. In order to fulfill our purpose, we consider the following constrained optimization problem:

$$(\hat{l}_1,\ldots,\hat{l}_m) = \underset{(l_1,\ldots,l_m)\in[0,\infty)^m}{\operatorname{argmax}} \frac{\sum_{i=1}^{n} k_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(y_{ij}-\bar{y}_i)^2}. \quad (2)$$

However, there is no closed form solution for this, and one needs to employ numerical optimisation methods for finding optimal solution (Nocedal and Wright, 1999).

For the ease of interpretation, we consider the normalized weights $w_p = \frac{l_p}{\sum_{p=1}^{m} l_p}$, so that the score $S_{ij} = \sum_{p=1}^{m} w_p Z_{ij}^{(p)}$, for $i = 1,\ldots,n$, and $j = 1,\ldots,k_i$, lies in the range [0, 1]. The rating is therefore obtained by sorting

the players, with respect to the average score $S_i = \frac{1}{k_i}\sum_{j=1}^{k_i} S_{ij}$, for $i = 1, \ldots, n$, in the descending order.

2.1 Variable Selection

While dealing with sports and related fields, which have a variety of quantitative aspects serving as performance attributes, selecting the most important ones is a crucial task. It may so happen that a certain subset of variables might be insignificant. The information contained in them, might already have been incorporated by some other variables. This might particularly happen in case of higher correlation among variables. In such an event, their corresponding weights might be close to zero. Therefore, using all the available variables will make the model unnecessarily cumbersome. In order to deal with such issue, we propose a variable selection technique to choose a suitable subset of variables for adequately explaining the final score. However, there is an involved cost of forfeiting the optimal solution and settling for a reasonable sub-optimal solution.

In order to choose an optimal subset of variables, from the complete set of available ones, we propose a backward subset selection technique. Since our methodology involves solving a constrained maximization problem, presented in equation (2), the maximum value of the objective function is attained when we use all the variables to obtain the full model $M^{(k)}$with k variables. Subsequently, dropping variables one by one leads us to sub-optimal solutions. In this method, we begin to generate the models $M^{(p)}$ s, for $p = k - 1, \ldots, 1$, each time by dropping the variable corresponding to the lowest associated weight. Having noted the corresponding values attained by (2) for all the models, we generate a scree plot of these values versus the number of variables rejected. In order to determine the appropriate number of variables, we plot the value of the objective function against the number of variables rejected, known as a Scree plot. We look for a knee (bend) in the Scree plot thereafter, to decide on a suitable number of variables.

Table 1: Data Structure Required for Analysis

Player ID	Match ID	Attribute 1	...	Attribute m
1	1	$X_{11}^{(1)}$	⋮	$X_{11}^{(m)}$
⋮	⋮	⋮	⋮	⋮
1	k_1	$X_{1k_1}^{(1)}$	⋮	$X_{1k_1}^{(m)}$
2	1	$X_{21}^{(1)}$	⋮	$X_{21}^{(m)}$
⋮	⋮	⋮	⋮	⋮
2	k_2	$X_{2k_2}^{(1)}$	⋮	$X_{2k_2}^{(m)}$
⋮	⋮	⋮	⋮	⋮
n	1	$X_{n1}^{(1)}$	⋮	$X_{n1}^{(m)}$
⋮	⋮	⋮	⋮	⋮
n	k_n	$X_{nk_n}^{(1)}$	⋮	$X_{nk_n}^{(m)}$

3. Data Analysis:

In this section, we applied our proposed methodology on the data obtained for the Indian Premier League. The data has been collected from www.cricsheets.org, www.espncricinfo.com and www.cricbuzz.com. We have considered the performance attributes of the players from the latest two seasons, IPL9 and IPL10. We have ignored cricketers who would not be available for the auctions scheduled next year due to retirement. However, the experience related attributes have been measured with respect to the inaugural edition of the IPL in 2008.

The data available at the aforementioned sources, was ball-wise and in the form of comma-separated-values tables. In order to summarize the raw data match-wise in the structure provided in Table 1, we used a Structured Query Language. Primarily, we considered two aspects of a player, the amount of experience and the way the player has performed in the previous matches. We have divided our analysis into the two segments in the game of cricket, batting and bowling. One can also extend this exercise to more granular levels of disciplines in cricket, like all-rounders, wicket keepers, etc.

For the batsmen, we have two experience related attributes, the number of matches played and the number of innings played, and three performance related attributes, namely runs scored, balls faced, and the strike rate. For the analysis, we have considered the number of innings that a batsman has played and ignored the number of matches which also includes the games where the batsman did not bat. As a performance variable, we considered the strike rate which is defined as the ratio of the total runs scored to the number of balls faced. Naturally, there exists very high correlation between the runs scored and the strike rate. Also, it is not fair to compare batsmen playing at different

positions with respect to number of run scored. In order to avoid such difficulties, we have used a filter for the minimum number of balls faced by a batsman (a minimum of 10 balls faced at least), and ensured coherence of the values of the strike rate (e.g., avoiding scenarios where a batsman faced just a single ball and scored a four, but making very little contribution to the team in terms of runs) among the batsmen. In this case, the strike rate provides combined information about the runs scored by the batsman and the number of balls used to score the same. We carried out the analysis for batsmen considering strike rate and the number of innings. The estimated value of the weights w_1 and w_2 are 0.746 and 0.254, respectively. The observed scores of the batsmen and the corresponding ranks are provided in Table 2.

Table 2: Ranking of batsmen

Rank	Player	Score	Rank	Player	Score	Rank	Player	Score
1	SK Raina	0.981	13	Yuvraj Singh	0.464	25	RR Pant	0.121
2	V Kohli	0.918	14	DJ Hooda	0.462	26	AJ Finch	0.120
3	MS Dhoni	0.872	15	S Dhawan	0.466	27	BB McCullum	0.118
4	RG Sharma	0.786	16	RV Uthappa	0.377	28	KD Karthik	0.117
5	G Gambhir	0.718	17	MK Pandey	0.337	29	Q de Kock	0.117
6	DA Warner	0.685	18	WP Saha	0.315	30	KK Nair	0.105
7	SR Watson	0.588	19	MC Henriques	0.221	31	RA Jadeja	0.104
8	AB de Villiers	0.558	20	AM Rahane	0.212	32	DR Smith	0.089
9	CH Gayle	0.551	21	SV Samson	0.207	33	KL Rahul	0.041
10	YK Pathan	0.538	22	JC Buttler	0.167	34	HM Amla	0.038
11	DA Miller	0.487	23	M Tiwary	0.158	35	KS Williamson	0.012
12	M Vijay	0.483	24	NV Ojha	0.154			

Next, we have analysed the match-wise data available for the bowlers. Similar to the case of batsmen, we have only considered the number of innings in which the bowler bowled, as the experience variable. Among the performance attributes, economy (runs conceded per over) indicates miserliness of a bowler. The performance of the bowlers is also monitored using average (runs conceded/number of wickets taken) and strike rate (number of balls bowled/number of wickets taken). Interestingly, both the metrics carry important information on the number of wickets taken by a bowler and rest of the information are redundant as those are already incorporated in economy rate. Therefore, we derived a new performance metric wickets-economy ratio (WER), defined as the ratio of the number of wickets picked up by the bowler in a particular match to the corresponding economy rate. The WER is directly proportional to the utility of a bowler, and a higher WER indicates that the bowler takes large number of wickets with low economy rate. We set a filter for the bowlers as well, and considered only those innings where they bowled at least two or more overs out of their stipulated quota of four overs. Based on our analysis considering WER and the number of innings, the estimated weights w_1 and w_2 are 0.791 and 0.209, respectively.

The observed scores of the individual bowlers and corresponding ranks are provided in Table 3.

Table 3: Ranking of bowlers

Rank	Player	Score	Rank	Player	Score	Rank	Player	Score
1	S Narine	0.941	8	A Mishra	0.575	15	I Tahir	0.175
2	B Kumar	0.660	9	U Yadav	0.510	16	R Khan	0.079
3	A Patel	0.658	10	J Bumrah	0.465	17	B Stokes	0.069
4	M Sharma	0.638	11	M McClenaghan	0.420	18	P Cummins	0.066
5	S Sharma	0.626	12	S Aravind	0.400	19	C Woakes	0.059
6	C Morris	0.614	13	Z Khan	0.304	20	B Thampi	0.058
7	Y Chahal	0.594	14	K Yadav	0.227			

4. Simulation Study

In this section, we have simulated the dataset as represented in Table 1, and tested our proposed methodology for validation purposes. As mentioned in Section 1, our method takes care of any hypothetical simplistic situation where just a single performance aspect is considered. In the more realistic scenarios with multiple aspects, we tested our method by first simulating two variables, similar to the analysis in Section 3 and then with four variables. When the data of the performance variables are simulated in order with an initial mean of 20, with a 10% incremental, keeping their consistencies same, the estimated scores are obviously in accordance irrespective of the choice of the weights. However, when the performance means are same, with varying consistencies, we tested our proposed methodology. We simulated the data by keeping the means of the two performance variables at 20 and 30, and the variances were set with a 10% incremental for each player, with the lowest player having a variance of 10 and 30 respectively for the two variables. We found out that it validates the fact that the scores are increasing in order of the increasing consistencies, or is in inverse order of the correspondingly increasing variances. So, in these two simplistic scenarios, our method generates results that are in line with the subjective decisions. However, in the real-life scenario, where both the means and variances are variant in an intertwined fashion, subjective decisions are not possible, thus necessitating our proposed methodology. We validated our claim of ensuring maximally differentiating players by subjecting a same dataset to three methods, the methodology proposed in this paper, and methods proposed by Saikia et al. (2012) and Croucher (2000). We have observed that the scores generated by our methodology have a higher variance, in fact it is more than twice the variance of the other, while the Spearman rank correlation coefficient between them is around 0.8

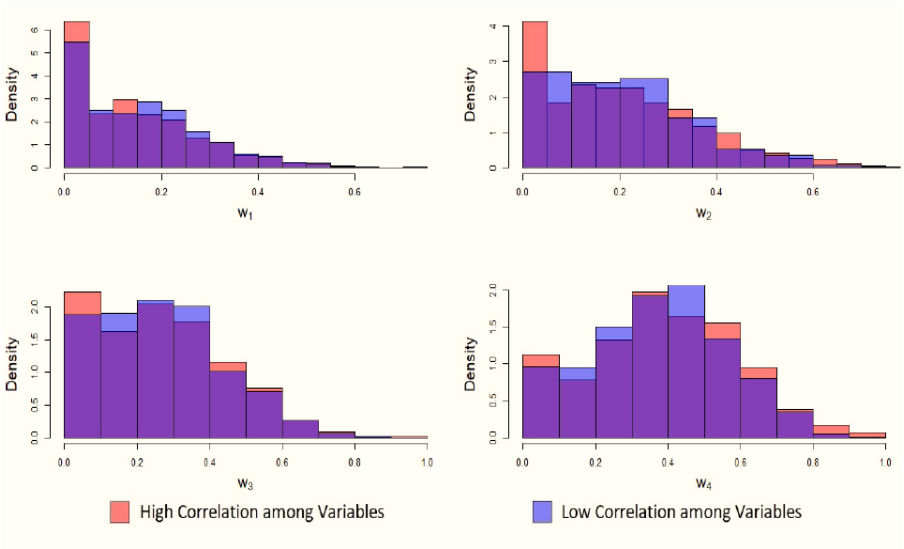


Figure 1: Behaviour of weights with varying correlation

4.1 Effect of Correlated Variables

We studied the effect of varying correlation among variables on our proposed methodology with four variables. The data for the performance of the players was simulated with low and high correlation, considering the pairwise correlation coefficients {0.1,0.2,0.2,0.3,0.3,0.4} and {0.7,0.8,0.7,0.8,0.9,0.9}, respectively. Our objective here is to identify the effect of correlation on the weights. The means of the variables were set to be {15,30,45,55} and their corresponding variances being f10,15,25,30g. The data was simulated with 1000 replications, and for every instance of simulated data, the weights were calculated using the methodology discussed in Section 2. In Figure 1, the superimposed histograms are presented. The histograms of the four weights with higher correlation are plotted in red and the ones with lower correlation are plotted in blue. While there is significant overlap as expected, there is a tendency of the weights shifting towards zero when the correlation is higher. This is more apparent for the weights w_1 and w_2, and comparatively less for the rest.

4.2 Subset Selection

In this section, we illustrate the proposed backward subset selection technique (see Section 2.1) through a simulation study. For this purpose, we simulated the performance data of the players by considering 8 variables with the mean vector (20,30,40,50,45,55,35,25) and the covariance matrix

$$\Sigma = \begin{bmatrix} 10 & 0.9 & 0.4 & 0.7 & 0.5 & 0.5 & 0.4 & 0.6 \\ 0.9 & 15 & 0.8 & 0.5 & 0.6 & 0.8 & 0.4 & 0.3 \\ 0.4 & 0.8 & 20 & 0.6 & 0.9 & 0.3 & 0.4 & 0.9 \\ 0.7 & 0.5 & 0.6 & 21 & 0.3 & 0.4 & 0.8 & 0.9 \\ 0.5 & 0.6 & 0.9 & 0.3 & 17 & 0.4 & 0.3 & 0.8 \\ 0.5 & 0.8 & 0.3 & 0.4 & 0,4 & 23 & 0.4 & 0.9 \\ 0.4 & 0.4 & 0.4 & 0.8 & 0,3 & 0.4 & 14 & 0.9 \\ 0,6 & 0.3 & 0.9 & 0.9 & 0.8 & 0.9 & 0.9 & 22 \end{bmatrix}$$

At first, we obtained the full model $M^{(8)}$ by considering all the variables and noted the corresponding value of the objective function. We obtained the subsequent sub-optimal solutions thereafter, by dropping the variable which had the least associated weight, as described in Section 2.1 and the corresponding the scree plot is shown in Figure 2. We clearly observe that the objective function attains the highest value when all the variables are considered, and there is a drop once we start to eliminate variables. However, upon the removal of a certain number of variables, there is a steep decline in the values of the objective function, in the subsequent sub-optimal solutions. In Figure 2, we observe that the knee occurs at about the point where 2 variables are dropped. Therefore, $M^{(6)}$ appears to be the suitable model. One can see that compromising two variables is sensible as the value of the objective function does not drop significantly, and the model is less cumbersome.

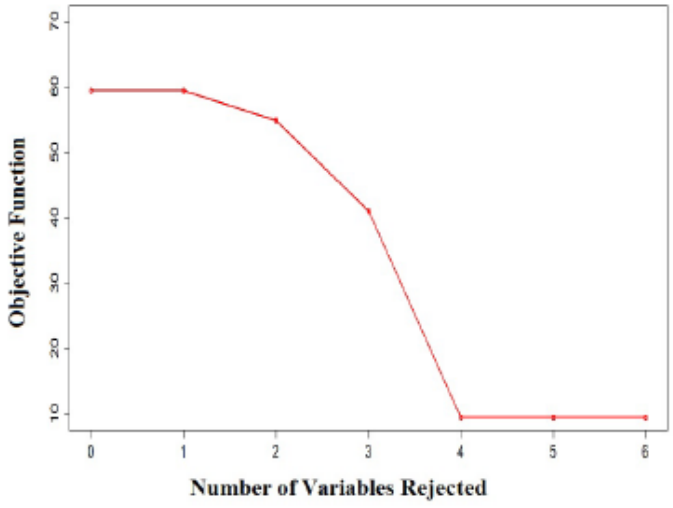


Figure 2: Selecting the subset using Scree Plot

5. Concluding Remarks

In this paper, we have put forward a proposition for an effective strategy for player selection by developing a scoring methodology based on historical data, which in principle ensures maximal discrimination amongst the players. Our proposed methodology is scalable, making it more useful for prospective

users. For example, one can use the proposed methodology to rank individuals with a specific skill type like opening batsmen, middle order batsmen, fast bowler, spinner etc. This will be particularly helpful as auctions are usually held skill-set wise, where each subset of a certain department is considered at a time. The proposed scoring mechanism can also be used to get an appropriate estimate of the worth of individual players. The evaluation of optimal valuation of a player needs further research considering the complexity and sequential nature of the auctions.

This methodology can also be generalized to a wide spectrum of fields, wherever there is data available in the form of repeated class-wise observations in different coherent aspects and an efficient ranking methodology is important. One interesting example could be the education sector, where the prospective candidates for graduate admission are ranked based on their performances in the past assessments. Also, one can use the proposed methodology to obtain an ranked list of individuals as per their intelligence based on the results from repeated IQ tests.

Acknowledgement
The authors are thankful to Dr. Arnab Chakraborty for many helpful comments and suggestions.

References

1. Bhattacharya, S. and Bhattacharya, S. (2012). Auction of players in indian premier league: The strategic perspective. *International Journal of Multidisciplinary Research*, 2.
2. Croucher, J. S. (2000). Player ratings in one-day cricket. In *Proceedings of the fifth Australian conference on mathematics and computers in sport*, pages 95–106. Sydney University of Technology Sydney, NSW.
3. Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Education.
4. Karnik, A. (2010). Valuing cricketers using hedonic price models. *Journal of Sports Economics*, 11(4):456–469.
5. Lenten, L. J., Geerling, W., and Konya, L. (2012). A hedonic model of player wage determination from the indian premier league auction: Further evidence. *Sport Management Review*, 15:6071.
6. Nocedal, J. and Wright, S. (1999). *Numerical Optimization*. Springer.
7. Saikia, H., Bhattacharjee, D., and Bhattacharjee, A. (2012). Is IPL responsible for cricketers performance in twenty20 world cup? *International Journal of Sports Science and Engineering*, 06:096–110.
8. Schuckers, M. (2011). An alternative to the nfl draft pick value chart based upon player performance. *Journal of Quantitative Analysis in Sports*, 7(2).
9. Tingling, P. M. (2017). Educated guesswork: Drafting in the national hockey league. In *Handbook of Statistical Methods and Analyses in Sports*, pages 343–356. Chapman and Hall/CRC.

Time-lagged variables and incidence of pneumonia in wet-dry tropical North Australia

Oyelola Adegboye[1*], Emma McBryde[1], Damon Eisen[1, 2]
[1]Australian Institute of Tropical Health and Medicine, James Cook University, Townsville, QLD
[2]Townsville Hospital and Health Service, QLD

Abstract

Few studies have focused on incidence of pneumonia in relation to time-lagged variables. We investigated the attributable risk of pneumonia associated with time-lagged weather variables in wet-dry tropics of Australia. We used distributed lag nonlinear models to estimate the relative risk associated with prolong exposure to weather conditions based on data from large cohort of patients hospitalized due to pneumonia between 2006 and 2016 North Queensland. The disease was identified using the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification (ICD10-AM) code J10.0* - J18. The risks associated with temperature were immediate and higher at moderate low (20 ^{o}C) temperature (RR=2.28, 95% eCI: 1.94 – 2.68). The cumulative effect over lag range 0 – 15 lag weeks revealed higher risk at moderate low rainfall (40 mm) with RR=5.49 (95% eCI: 2.27 – 13.24). About one-third, 28.4% (95% eCI: 19.7 – 36.2) of pneumonia cases were attributable to temperature (mostly due to moderate temperatures) while the overall proportion of cases attributable to rainfall (50.8%, 95% eCI: -6.4 – 76.5) was higher than attributable to temperature however, not significant. The findings in this study can inform a better understanding of the health implications and burden associated with pneumonia to support discussion-making in healthcare and establish a strategy for prevention and control of the disease among vulnerable groups.

Keywords

Wet-dry tropics; Pneumonia; Lagged variables; Temperature; Rainfall

1. Introduction

Lower respiratory infections accounted for about 40% of the total infectious disease burden in 2011 in Australia [1]. Australian hospital statistics revealed that about 4.3% of total primary diagnosis during the period 2012-2013 were classified as disease of respiratory system [2]. Potentially preventable hospitalization (PPH) accounts for 8.2% of all hospital admissions during the same period, of which 0.3% were vaccine preventable [2]. Queensland has second largest burden of PPH preceded by Northern Territory (17.2% vs. 23.2%) in Australia, one-third of which were vaccine preventable influenza and pneumonia [2].

The weather-pneumonia associations varied across the regions of the world. For example, positive association of temperature-pneumonia was observed in Mediterranean *climate of* California, United States of America [3], subtropical regions of China [4] and Australia [5, 6]. While in tropical regions of South Asia and Sub-Sahara Africa, the disease is associated with wet-rainy season (with less sunshine) [7-9].

The aspects of weather effects and seasonal variation of pneumonia has been largely unexplored in Australia [10] with majority of the studies conducted in the subtropical region [5, 6]. Wet-dry tropics North-East coastal region of Australia is characterized by distinct wet and dry seasons with high temperature throughout the year. Most of the rainfall in this region occurred during the summer season with high temperature.

In this study, we investigated the influence of temperature and rainfall on pneumonia in wet-dry tropics of North Queensland using a time series analysis via distributed lag nonlinear model (DLNM) analysis of data-linkage data between 2006 and 2016. The DLNM is a novel and flexible modelling structure for dealing with lagged nonlinear relations between or among time series structures. It will efficiently capture and control the behaviour of study variables in the exposure range and time dimension. The results of the time series analysis was used to identify vulnerable groups and estimates disease burden attributable to varying exposure-lag-response relationships. Also, given that pneumonia incidence is recorded throughout the year, adequate and reliable quantification of exposure-response is of utmost importance.

2. Methodology

Data sources

The data used in this study was part of a data linkage project from a large retrospective cohort of Townsville Hospital patients discharged with an ICD10-AM code for an infectious disease from 1 January 2006 to 31 December 2016. The use of ICD10-AM codes for infectious diseases have been shown to be closely correlated with clinical diagnoses in Australian research [11, 12].

In this study, every patient hospitalized at Townsville hospital assigned ICD10-AM codes J10.0* - J18* (a diagnosis of pneumonia including cases due to influenza) were included in this study. Other variable extracted were age, sex, indigenous status, admission source and presence of comorbidities.

Furthermore, individual pneumonia cases were aggregated to weekly data to investigate seasonality of pneumonia and the role climatic variables. Data on climate variables, daily mean temperature and daily mean rainfall were obtain from Australian Bureau of Meteorological. Daily mean temperature was averaged to weekly mean temperatures while daily mean rainfall was aggregated to total weekly rainfall.

Ethical approval was obtained from the THHS Human Research Ethics Committee (HREC/16/QTHS/221) and the Queensland Public Health Act (RD007802) for the data linkage project.

Statistical analysis

Estimation of the climate-pneumonia association

Weekly cases of pneumonia and climatic variables (rainfall, temperature) were analysed using distributed lag non-linear model (DLNM) [13-16] to investigate the association between pneumonia cases and rainfall or temperature in THHS from 2006 to 2016.

The weekly counts of pneumonia cases was fitted via quasi-Poisson generalized linear regression models adjusting for season, long-term trend, weekly mean temperature (ºC) and total weekly rainfall (mm). We used distributed lag non-linear models (DLNMs) [13-16] to model the potential non-linear and delayed (lagged) effects of temperature and rainfall.

$$Y_t \sim quasiPoisson(\mu_t)$$
$$log(\mu_t) = \propto + \sum_{j=1}^{J} s_j(x_{ij}\beta_j) + \sum_{k=1}^{K} \gamma_k u_{tk}$$

Where Y_t represents the weekly observed pneumonia cases on week t with mean μ_t, $\propto$ is the model intercept. The function, s_j is used to specify the functional relationship between variables x_j and the nonlinear exposure-response curve, defined by the parameter vectors β_j. The variables u_k include other predictors with linear effects specified by the related coefficients γ_k.

Previous studies have suggested that the effect of a specific exposure event is not limited to the period when it is observed, association may spread over a few time periods [15, 17]. Therefore, we modelled the non-linear and delayed effects of a rainfall and temperature through functions s_j which define the relationship along the two dimensions of predictor and lags. That is, the exposure-lag-response was modelled by applying a bi-dimensional cross-basis spline function describing simultaneously the dependency of the relationship along the temperature range and its distributed lag effects.

The relaxed cross-basis parameterization for exposure-lag-response is given by:

$$s_j(x,t) = \int_{l_0}^{L} f \cdot w(x_{t-l}, l)dl \approx \sum_{l_0}^{L} f \cdot w(x_{t-l}, l) = w_{x,t}^{T}\eta$$

Where the bi-dimensional function $f \cdot w(x_{t-l}, l)$ define the *exposure–lag–response function*, and models simultaneously the exposure–response $f(x)$ curve along temperature/rainfall range and lag–response curve, $\omega(l)$ [14].

Attributable risk measure

The attributable measures; attributable fraction (AF) and attributable number (AN) are the most useful indicator of exposure-related health burdens [18, 19]. We estimated the fraction of pneumonia cases attributed (AF) to weekly mean temperature and total weekly rainfall, separately using the optimum weather values as references.

Attributed fraction (AF) measure was derived from prediction of the overall cumulative exposure-response relationship in the DLNM model. Using the minimum incidence percentile, x_0 across the entire exposure spectrum as the reference and cut-off for optimum temperature/rainfall value, we used a backward perspective [18, 19], assuming that the risk at week t was attributable to a series of exposure, x events in the past, $t - l_0, \ldots, t - L$.
The attributable fraction ($b - AF_{x,t}$) for a given exposure is derived as follows:

$$b - AF_{x,t} = 1 - e^{-\sum_{l=l_0}^{L} \beta_{x_{t-l},l}}$$

Where $\beta_{x_{t-l},l}$ represented the risk associated (logRR) with lagged exposure, x at time, $t - l$.

All statistical analyses were performed with R statistics software v3.4.0 [20], with the package "dlnm" to create the DLNM [13].

3. Result

The weekly time series distributions of cohort of pneumonia cases were plotted in Figure 1. The time series decomposition shows increase trend and seasonal patterns in cases of pneumonia over the years. Similarly, the pattern of seasonality (alternating highs and low) of pneumonia cases is inversely mirrored by mean weekly temperature and total weekly rainfall (not shown). The summaries of cases and climate variables were presented in Table 1. There were negative correlations between weekly pneumonia cases- and mean temperature (r = -0.224, P<0.001), minimum temperature (r = -0.217, P<0.001), maximum temperature (r = -0.218, P<0.001) and total rainfall (r = -0.099, P=0.017).

Correlation among temperature variables were higher than 0.7 (not shown), therefore to prevent issues with multi-collinearity, we based this study on mean weekly temperature and total weekly rainfall.

The best DLNM model (out 128 candidates models) described the weather-pneumonia association by lag up to 15 weeks and quadratic B-splines function for temperature/rainfall-pneumonia relationship and linear function for lag-pneumonia relationship with a total degrees of freedom of 6 (based on smallest QAIC=3355.8). The model also include a natural cubic spline

function for long term temporal trend and seasonality with 8 dfs generated per year of study. The pneumonia-lag and weather-pneumonia associations were presented in Figures 2 & 3, respectively with reference at optimum temperature value of 28.8°C (95% empirical confidence interval, eCI: 16.0 - 31.4) and optimum total weekly rainfall value of 332.2 mm (95% eCI: 0 - 516).

Higher pneumonia risk and immediate effect (lag 0) was observed at lower temperatures while the effect of rainfall is delayed (up to lag 15) at low total weekly rainfall. The risk associated with rainfall (of 30 mm) appeared significant after lag of 5 weeks and progressively increases for longer exposure. However, significant risk associated with temperature exposure of 20 °C was immediate and lasted for 1-12 weeks.

Table 1: Summary characteristics of pneumonia cases, weekly temperature and rainfall.

Variables	Mean	SD	Min	Percentile			Max	Cor. (p-value)*
				25	50	75		
Pneumonia cases	15.3	7.94	2.0	10.0	14.0	19.0	62.0	
Mean temperature (°C)	24.8	3.26	15.98	22.43	25.53	27.51	31.43	-0.224 (<0.001)
Minimum temperature (°C)	20.32	4.21	8.78	17.48	21.14	24.07	27.26	-0.217 (<0.001)
Maximum temperature (°C)	29.27	2.52	20.27	27.26	29.67	31.27	35.60	-0.218 (<0.001)
Rainfall (mm)								
Total weekly	23.5	60.32	0	0	1.4	13.5	516.0	-0.099 (0.017)
Weekly average	3.55	9.29	0	0	0.2	2.0	78.1	-0.118 (0.005)

*Correlation between pneumonia cases and climate variables

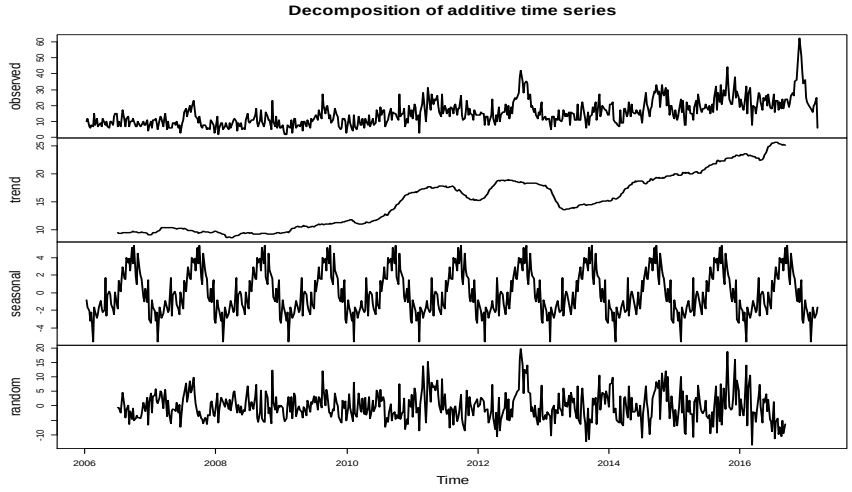


Fig. 1: Decomposition of weekly cases of pneumonia additive time series during the study period, 2006-2016.

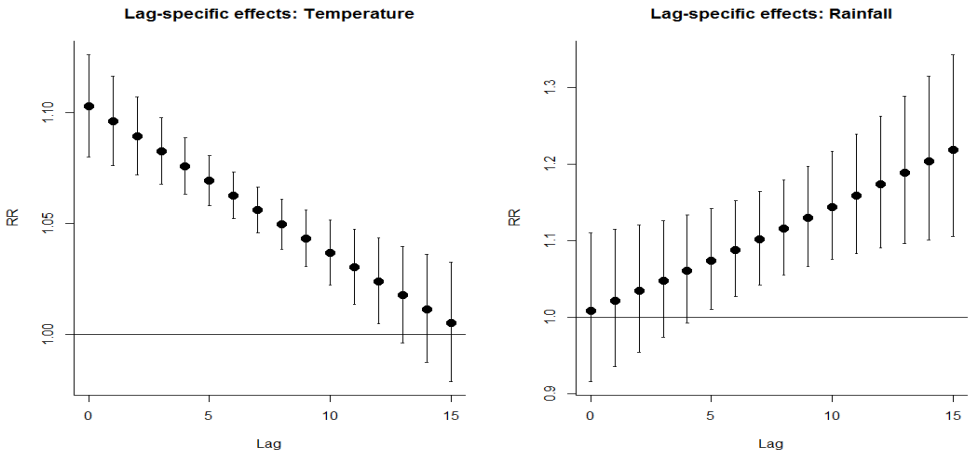


Fig 2: Lag-specific effects on pneumonia at different weather values. (Left) Temperature of 20 °C. (Right): Rainfall of 30 .

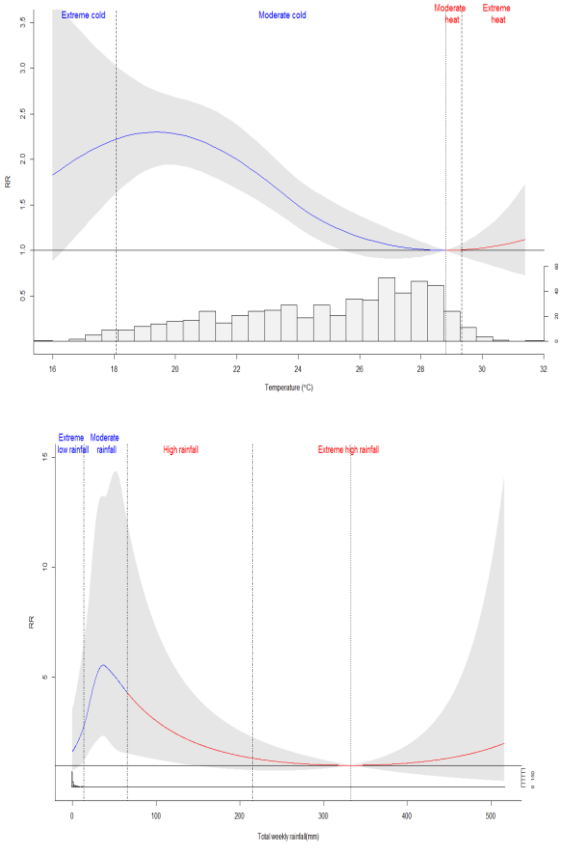


Fig 3: Overall effects of weather variables on CAP. Top: For a unit increase in temperature. Bottom: Total weekly rainfall.

The overall picture of the cumulative effects of temperature and rainfall over the range 0 – 15 weeks are presented in Figure 3. The figures are divided

into segments of extreme and moderate cold/heat as well as extreme and moderate low/high rainfall. The risks associated with temperature (with reference to optimum temperature of 28.8 ^{o}C) were significantly higher for temperatures lower than 25 ^{o}C. For example, the cumulative associated risk over the range 0-15 lags weeks at extreme cold (16 ^{o}C) and moderate cold (20 ^{o}C) were 1.83 (95% eCI: 0.88 – 3.78) and 2.28 (95% eCI: 1.94 – 2.68, respectively (Table 3). The risk were higher for shorter lags and decreases over longer. In the case of rainfall, the cumulative effect over lag range 0 – 15 lag weeks revealed higher risk at moderate low rainfall (40 mm) with RR=5.49 (95% eCI: 2.27 – 13.24).

Using backward perspective and the optimum value as references, the estimated proportions of incidence of pneumonia attributable to temperature and rainfall are presented in Table 4. The total attributable fraction due to temperature is 28.4% (95% eCI: 19.7 – 36.2). Taken separately, the total fraction of cases attributed to varying temperature is predominantly due to moderate temperatures (27.0%, 95% eCI: 17.8 – 34.8) against extreme heat (0.07%, 95% eCI: -0.4 – 0.6) (Table 4). The overall proportion of cases attributable to rainfall is higher than attributable to temperature however, not significant (28.4%, 95% eCI: 19.7 – 36.2 vs. 50.8%, 95% eCI: -6.4 – 76.5). Most pneumonia cases attributable to rainfall occurred during moderate rainfall (18.4%, 95% eCI: 8.5 – 26.4).

Differences were observed in the associated risk of pneumonia and varying temperatures and rainfall among different groups. For example, looking at specific estimated risk associated with 15 weeks cumulative exposure of 20^{o}C temperature, we observed; increased relative risk of pneumonia in females (RR=2.90, 95% CI: 2.31 – 3.64) compared to males (RR=1.87, 95% CI: 1.52 – 2.30), among indigenous (RR=2.55, 95% CI: 1.82 – 3.57) vs. non-indigenous (RR=2.22, 95% CI: 1.86 – 2.65); emergency admission (RR=2.44, 95% CI: 2.01 – 2.95) vs. others (RR=1.96, 95% CI: 1.45 – 2.66) and older patients aged >14 years (RR=2.29, 95% CI: 1.43 – 3.65) against RR=2.62 (95% CI:1.99 – 3.43) for 14 years and less. Similarly, the pneumonia-rainfall associated risk of exposure to a total weekly rainfall of 40 mm (at the end of 15 weeks) is higher for females (RR=6.27, 95% CI: 1.87 – 21.12) than males (RR=4.84, 95% CI: 1.54 – 15.07), higher among indigenous (RR=6.75, 95% CI: 1.08 – 42.36) vs. non-indigenous (RR=4.81, 95% CI: 1.99 – 13.64); lower for emergency admission (RR=4.82, 95% CI: 1.71 – 13.54) vs. others (RR=7.41, 95% CI: 1.36 – 40.28 and lower for older patients aged >14 years (RR=5.89, 95% CI: 1.40 – 24.78) against RR=11.36 (95% CI: 0.67 – 190.73) for 14 years and less.

4. Discussion and Conclusion

The 2015 global burden of disease revealed that lower respiratory tract infections is the fifth leading cause of death worldwide[21]. Among these,

mortality caused by pneumococcal pneumonia (55.4%) is the highest among LRIs [21]. In Australia, pneumonia accounted for an estimated 1.5% of all overnight hospital admission in 2012-2013 [2]. In this study, we quantified the lagged-pneumonia and exposure-pneumonia associations due to temperature and rainfall in wet-dry tropics of Australia. The burden of the disease attributable to the two weather variables was also estimated. Some countries have reported increases pattern of infectious disease among minorities [22], among children [23, 24] and older age group [25] which is consistent with our study.

This study has several strengths. To our knowledge this is the first study to estimate the burden of pneumonia attributable to weather in Australia. The use of DLNMs to explore attributable risk among vulnerable groups, give more insight into varying prolong exposure to weather in the community which will be useful in discussion-making. The two major limitations were noted in this study. First, the choice of lag-exposure use to investigate the delayed effect of weather variables was based on model selection criteria for better fit and not scientific justification. Second, few smooth functions were explored to capture the exposure-lag-response relationships. These smoothing methods are difficult to validate in DLNM [16]. Several lags up to 15 weeks (although, we did not extend the lags beyond 15 weeks) and functional relationships for exposure-lag-response were assessed by QAIC.

References

1. Hoy, W.E., *Australian burden of disease study: impact and causes of illness and death in Australia 2011.* 2016.
2. Australian Institute of Health and Welfare, *Australian hospital statistics 2012–13*, in *Health services series*. 2014, AIHW: Canberra.
3. Green, R.S., et al., *The effect of temperature on hospital admissions in nine California counties.* International journal of public health, 2010. **55**(2): p. 113-121.
4. Song, G., et al., *Diurnal temperature range as a novel risk factor for COPD death.* Respirology, 2008. **13**(7): p. 1066-1069.
5. Xu, Z., et al., *Impact of temperature on childhood pneumonia estimated from satellite remote sensing.* Environmental research, 2014. **132**: p. 334-341.
6. Xu, Z., W. Hu, and S. Tong, *Temperature variability and childhood pneumonia: an ecological study.* Environmental Health, 2014. **13**(1): p. 51.
7. Chan, P., et al., *Seasonal variation in respiratory syncytial virus chest infection in the tropics.* Pediatric pulmonology, 2002. **34**(1): p. 47-51.
8. Paynter, S., et al., *Sunshine, rainfall, humidity and child pneumonia in the tropics: time-series analyses.* Epidemiology & Infection, 2013. **141**(6): p. 1328-1336.
9. Enwere, G., et al., *Epidemiology and clinical features of pneumonia according to radiographic findings in Gambian children.* Tropical medicine & international health, 2007. **12**(11): p. 1377-1385.
10. Murdoch, K.M., et al., *What is the seasonal distribution of community acquired pneumonia over time? A systematic review.* Australasian Emergency Nursing Journal, 2014. **17**(1): p. 30-42.
11. Skull, S.A., et al., *ICD-10 codes are a valid tool for identification of pneumonia in hospitalized patients aged > or = 65 years.* Epidemiol Infect, 2008. **136**(2): p. 232-40.
12. Skull, S.A., et al., *Hospitalized community-acquired pneumonia in the elderly: an Australian case-cohort study.* Epidemiol Infect, 2009. **137**(2): p. 194-202.
13. Gasparrini, A., *Distributed Lag Linear and Non-Linear Models in R: The Package dlnm.* JOURNAL OF STATISTICAL SOFTWARE, 2011. **43**(8): p. 1-20.
14. Gasparrini, A., *Modeling exposure–lag–response associations with distributed lag non-linear models.* Statistics in Medicine, 2014. **33**(5): p. 881-899.
15. Gasparrini, A., B. Armstrong, and M.G. Kenward, *Distributed lag non-linear models.* Statistics in medicine, 2010. **29**(21): p. 2224-2234.
16. Gasparrini, A., et al., *A penalized framework for distributed lag non-linear models.* Biometrics, 2017. **73**(3): p. 938-948.

17. Adegboye, O. and M. Adegboye, *Spatially correlated time series and ecological niche analysis of cutaneous leishmaniasis in Afghanistan.* International Journal of Environmental Research and Public Health, 2017. **14**(3): p. 309.
18. Gasparrini, A. and M. Leone, *Attributable risk from distributed lag models.* BMC Medical Research Methodology, 2014. **14**(1): p. 55-55.
19. Gasparrini, A., et al., *Mortality risk attributable to high and low ambient temperature: a multicountry observational study.* The Lancet, 2015. **386**(9991): p. 369-375.
20. R Core Team, *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. 2017.
21. Troeger, C., et al., *Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015.* The Lancet Infectious Diseases, 2017. **17**(11): p. 1133-1161.
22. Torzillo, P., et al., *Etiology of acute lower respiratory tract infection in Central Australian Aboriginal children.* The Pediatric infectious disease journal, 1999. **18**(8): p. 714-721.
23. Paynter, S., et al., *Childhood pneumonia: a neglected, climate-sensitive disease?* The lancet, 2010. **376**(9755): p. 1804-1805.
24. Adegbola, R. and S. Obaro, *Diagnosis of childhood pneumonia in the tropics.* Annals of Tropical Medicine & Parasitology, 2000. **94**(3): p. 197-207.
25. Sharma, S., et al., *Community-acquired syndromes causing morbidity and mortality in Australia.* Commun Dis Intell Q Rep, 2017. **41**(1): p. E49-E57.

Assessment of condominium occupancy rate in Bangkok and its vicinity from electricity meter data analytics*

Jittima Dummee, Kuntip Trongthamakit
Bank of Thailand, Bangkok, Thailand

Abstract

The purpose of study is to apply the use of micro-level data as supplement indicators for assessing economic condition in couple with other traditional aggregate data. As urbanization has prevailed thoroughly in large cities, living in high-rises is often people's choices. It follows that demand for residential units, i.e., condominiums can be quantified. In the study, electricity meter data are adopted as proxy to estimate resident occupancy rate. The result reveals electricity data can be effectively used as a supplementary economic indicator. However, it is worth noting that on its own supplement indicator alone is not a substitute for conventional analysis yet it is synergistic with other variables in helping to enhance accuracy, profoundness as well as comprehensiveness of the findings.

Keywords

Electricity data; Micro data; Occupancy rate; Real estate

**Views expressed in this paper do not necessarily reflect views of the Bank of Thailand.

1. Introduction

Apart from the 4 requisites, electricity, is an essential necessity in our life. It has become the fifth requisite in modern civilization long before the age of smart phones. In economic context, electricity is considered as input or "factor of production" for goods and services, thus, economic growth usually coincides with higher energy consumption and *vice versa*. Regarding this, the study utilises this concept to investigate occupancy rate of residential buildings in Bangkok and surroundings.

Electricity data is public utility data which contains recondite information about human, at least, geographically and behaviourally. Sensibly to say it can

* This study would not have been possible without support from following people. Authors wish to express their gratitude to Mr. Permsuk Sutthinoon, Dr. Pichit Patrawimolpon, Dr. Somsajee Siksamat, Dr. Don Nakornthab and Mr. Chatchawan Intarak, for valuable comments, assistance and full support. Special thanks are also addressed to Mr. Suwatchai Chaikhor and colleagues from Monetary Policy Group, for initiating the study of this topic as well as staff from Information Technology Group for data preparation. Lastly, many thanks to Mr. Krit Chalermdumrichai as an editor for this paper. All remaining errors are the authors' own.

mirror broad base economic activities in general. In addition, such data can be utilized for economic research in various ways, for example, the U.S. Energy Information Administration (2014) has tracked electricity consumption data to indicate economic growth relative to the GDP.

In Thailand, main entities supplying electricity for households consisting of the Electricity Generating Authority of Thailand (EGAT), the Metropolitan Electricity Authority (MEA) and the Provincial Electricity Authority (PEA). Data used by those authorities are for strategic planning and power management, yet for the Bank of Thailand (BOT), the approach of data usage is an alternative. Household electricity consumption data, in tradition, is part of components in Private Consumption Index (PCI) as well as industrial electricity consumption, one kind of manufacturing indicator, however, data adoption for the above is in aggregate format.

The MEA and the BOT has entered into a Memorandum of Understanding (MOU[1]) for data collaboration. Electricity data obtained from the MEA is micro–level data which is a new and unconventional approach. The data acquired cover wide range of economic units including households, businesses and authorities. Major advantages of such data are it has short time lag (around 2 - 3 weeks) and a long time series which is well sufficient for study and analytics. In the trial, some key strategic objectives are achieved involving utilizing micro-data along with the aggregate data to assess state of economy in a comprehensive and timely manner.

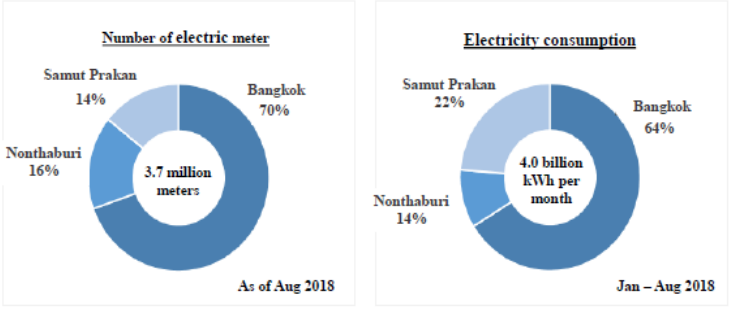
In the crucible, focus is placed on calibrating some indicators for property sector, particularly, for measuring demand for units in residential buildings – quite often this is referred to condominiums that share greater weights in real estate sector in the present. On methodology, an approach developed by Ecotagious (2016) is adopted. Basically, by applying daily electricity data from households in Vancouver, Canada to estimate Non-Occupancy Rate (NOR), also there is a study by the MEA and the Thai National Housing Authority (2009) aimed to project number of unoccupied dwellings in 3 provinces of Thailand by using monthly electricity data. In short, the purpose of the study is to investigate electricity usage data to measure occupancy rate (OR) for articulating real demand for units in condominiums located in 3 territories comprising Bangkok, Nonthaburi and Samut Prakan. Final attempt from findings will serve as downstream indicators for real estate market analysis. Prior to the use of micro-data, the BOT detected just only up- and midstream indicators, for instance, number of new unit launched and units sold. With regards to this, it enhances monitoring capability to be more comprehensive starting from up-, mid- and downstream level.

[1] The MOU was signed on 21 June 2016.

2. Methodology

Power meter data under the study is supported by the MEA, an entity supplying electricity for 3 provinces. Data frequency is on monthly basis dating from January 2014 onwards. On average, there are more than 3.7 million active meters in the observation, with 4.0 billion kilowatt - hours (kWh) capacity usage per month. Evidentially, Bangkok, the capital city shares the largest portion for meters and electricity consumption to carter energy demand from business districts, public transportation systems and residential areas. (Figure 1)

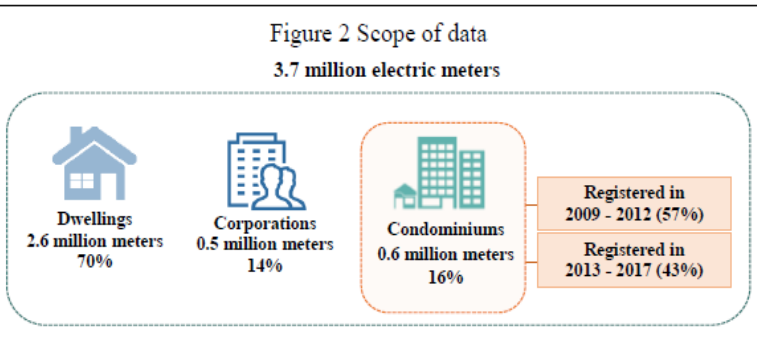
Figure 1 Number of meters and the amount of electricity consumption classified by Province



Source: MEA and calculated by BOT

Since the data from 3.7 million meters in 3 provinces is the population frame, yet the focal point of the analysis is to observe the OR of residential buildings, so some rules need to be applied to filter out the unrelated. To replicate the fact, relationship between master-meter and sub-meter[2] is purposefully defined. Regarding this, it turns out that there are 3,591 condominium buildings comprising 605,744 units representing the sampling data, as shown in Figure 2.

Figure 2 Scope of data



Source: MEA and calculated by BOT

To ensure the appropriateness of obtained sampling derived from counting electricity meters, number of sampling units is then compared with number of new residential units disclosed by the Real Estate Information

[2] Master-meter is defined as main meter of a building. Sub–meters, subsequently connected from Master-meter, are ones that attached to each individual residential units within the building.

Center (REIC), an agent taking charge of recording number of condominium units launched. Findings shown in table below.

Table 1. The number of new residential units and the number of electric meters

Data	2013	2014	2015	2016	2017	Total
New residential units	67,267	65,313	53,626	56,721	57,563	300,490
Electric meters	48,616	45,679	56,659	62,324	47,849	261,127
(%)	(72.3)	(69.9)	(105.7)	(109.9)	(83.1)	(86.9)

Source: MEA, REIC and calculated by BOT

Remarks: Two data sources for new residential units launched in 3 territories from 2013-17.

Results are concluded as follows:

1. The 5-year average ratio between registered electricity meter and REIC data stands around 87%. This exhibits meter data can be treated as one proxy for estimating new residential units launched in each period.
2. In some years, numbers of meter data exceed new residential units launched, for instance, in 2015-16. This was caused by cumulative effect, for new unsold units launched in 2013–14 were sold-out, also had meter registered in years after.

To model the Occupancy Rate (OR), one presumption of un-occupancy units is required. In the study, use of 20 kilowatt-hours (kWh)[3] or above energy consumption per month is set as a "Threshold" for occupied units. For other group, units with power utilization less than 20 kilowatt-hours shall be treated as unoccupied. It follows naturally that occupancy rate can be calculated by taking number of units with electricity usage equal or above the *Threshold*, dividing by total number of units (Equation 1). If OR is high, this means demand for residential units is elevated. The relationship also holds in the opposite direction.

$$\text{Occupancy Rate (OR)} = \frac{\text{Number of occupied units}}{\text{Total number of units}} \qquad \text{(Equation 1)}$$

[3] In the study, a Threshold of electricity usage (in units of kilowatt-hours per month) is imposed, but not non-occupancy duration criteria as defined by Ecotagious (2016). To improve the Threshold's validity and precision, non–occupancy period such as students' non-occupancy during school holidays should be factored-in to calculate average energy consumption. Following this concerns, the NOR and OR can reflect better reality. On balance, the computed OR without imposition of duration criteria can be regarded as minimum occupancy level, actual residency rate could be more.

Up to this point, justification for the state of property market is an interesting issue. Referring to business practice in property sector adopted by the Agency for Real Estate Affairs (AREA)[4], achieving 70% sold-out is regarded as breakeven or near-breakeven for estate projects, so it is logical to adopt 70% *Threshold* for the same purpose. In plain words, if the OR exceeds 70 percent, it signifies positive condition, also true contrariwise. Notwithstanding, one data constraint should be cautious, the OR it is not able to differentiate between two types of residency: "owner" and "renter".

Back to the OR, low occupancy can be resulted by different determinants. One interpretation is it may cause by multi-unit owners which result in leaving some units unoccupied. The other cause may by former tenants who had moved out for reasons. In such cases, this could contribute some absence to the demand for units. As a result, rendition of the occupancy rate has to be done with proper judgments. On the other end, supply factor could pose some effect to the OR, new supply can lead to a sudden increase in total units while the number of occupied rooms remain stable or slightly increasing. With reference to the scenario, low OR may be temporary, since it could take a few months for new tenants to move in. Technically speaking, by monitoring speed of move-in, some potential indicators can be constructed to measure real demand for units of the projects launched.

3. Empirical Findings

Results from adopting power meter data to approximate the OR for residential units in 3 territories can be summarized as follows:

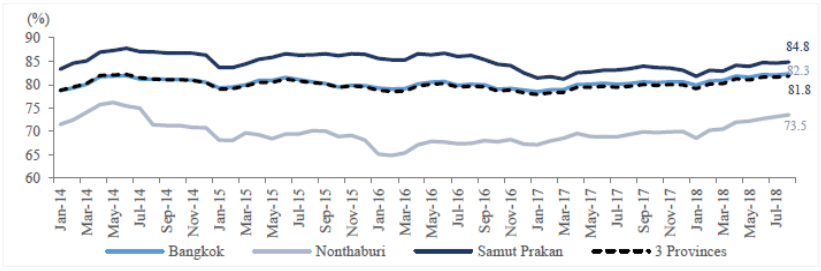
a) Average occupancy rate, from 2014-18, had soared above 70 percent. To be precise, the OR during the period stood at 80, 70 and 85 percent for Bangkok, Nonthaburi and Samut Prakan respectively. Following more recent data, it reveals overall OR had extended to 81.8 percent in August 2018. For Nonthaburi, improvements in OR are resulted by extensions of public transportation, allowing residents to have better connectivity by conveniently commuting from their location to others. As a consequence, the OR has constantly increased from 65.1 percent in January 2016 to 73.5 percent in August 2018. Even with improved public transport system in the province, OR in Nonthaburi is relatively lower than two peer territories, Bangkok (82.3 percent) and Samut Prakan (84.8 percent) (Figure 3).

   For units identified as unoccupied, such group can be defined by adopting *Threshold* or level of energy consumption per month as mentioned earlier to classify this type into 2 categories, empty rooms

[4] Information gathered from AREA staff. In business practice, achievement of a project is justified under sales of residential unit is above 70 percent within 1.5 years since its launch.

and unoccupied rooms[5]. It is observed that the first represents 12 percent while the latter shows 6 percent of the samples.

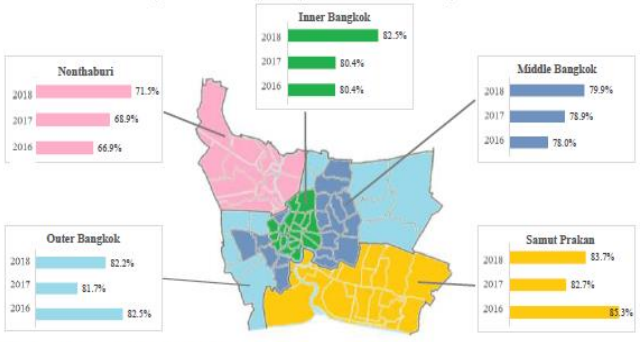
Figure 3 Occupancy rate in Bangkok, Nonthaburi and Samut Prakan



Source: MEA and calculated by BOT

b) In Figure 4, the study was conducted on provincial basis for in-depth analysis. Thus, one territory in the observation, "Bangkok" is divided into 3 areas as - inner, middle and outer Bangkok[6]. Findings show that for inner and middle Bangkok and Nonthaburi, from 2016-18, the OR had continually increased while outer Bangkok and Samut Prakan had declined in 2017. This can be explained individually in the following. For outer Bangkok, it was resulted by stagnant demand for residential units that lasted in 2017 but for Samut Prakan, caused by an increase in supply in late 2016. In short, tumble in 2 areas can be described as demand and supply factor respectively. In 2018, however, the occupancy rate of the latters had remarkably recovered.

Figure 4 Occupancy rate classified by area



Source: MEA and calculated by BOT
Remark: 2018 is the data from January to August 2018

c) The study is furthered by classifying aging of residential buildings into 2 clusters:
1) Old Meters – for those meters registered prior to 2012

[5] Empty rooms imply for meters with zero electricity consumption while unoccupied rooms imply for those which consume electricity less than 20 kWh per month.
[6] Referred to area division, Bangkok Metropolitan Administration, Office of Permanent Secretary

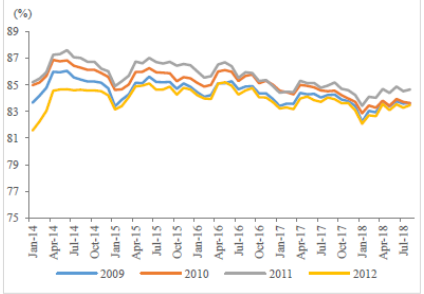
2) New Meters – for those meters registered from 2013 onwards

For Old Meters (Figure 5), findings discover that the OR of old meters had stood over 80 percent. Notwithstanding, there was a slightly decline in OR for this group after 2016. Of this group, 8 percent of residents purchased more units due to deteriorating conditions of existing buildings. Regarding this, this pattern is the so-called "relocation effect" of residents. Following the logic, "relocation effect" has exposed some drops to the old-meter cluster.

For New Meters (Figure 6), it was found that the OR has continued to rise yet the level is still below the old meters group. Some stylized facts classified by year of registration are shown below:

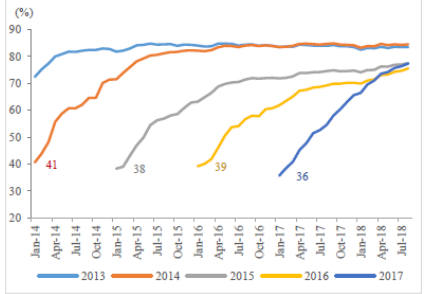
- For 2014, the initial OR stood at 41 percent. It took 11 months to take the OR to reach 70% *Threshold*. This was partly due to less supply launched and sluggish economic condition.
- For 2015-16, the initial OR was around 39 percent and required 18-22 months to achieve the *Threshold*. This was due to more supply launched
- For 2017, initial OR remained at 36 percent and it get 15 months to achieve the *Threshold*. This was supported by economic prosperity during the year compared to the two previous years.

Figure 5 Occupancy rate of old condominiums



Source: MEA and calculated by BOT

Figure 6 Occupancy rate of new condominiums



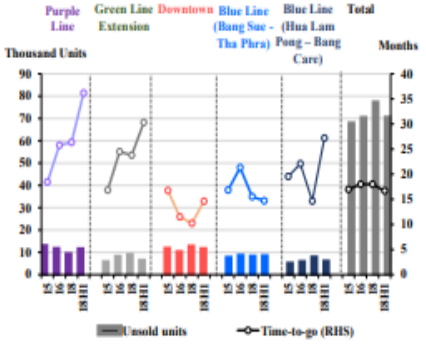
Source: MEA and calculated by BOT

d) In this section, other indicators apart from power meters was adopted to investigate some further aspects, for instance, "unsold units - stock" and "time-to-go" (length of time that would take to complete the sales of existing unsold units by assuming fixed supply for certain time). In simple mathematics, an assumption of sales per month is equal to average sales since the launch of projects is addressed. Relevant data are supported by the Agency for Real Estate Affairs (AREA).

From Figure 7 and 8, it is observed that there existed a condition of oversupply of residential units in Bangkok and its vicinity during the first half of 2018 yet

the situation slightly improved from 2017. Judging from "time-to-go" aspect, time interval was reduced from 18 to 17 months. The situation implies current demand could match with additional supply in that period, OR thus improved. However, there are certain locations that are not consistent with overall findings due to factor- or location-specific reasons.

Figure 7 Unsold condominium units in Bangkok and its vicinity



Source: AREA and calculated by BOT
Remark: as of June 2018

Figure 8 OR classified by the train line

Occupancy rate (%) 2018H1	Registered before 2015	Registered from 2015 onwards
Downtown	86.5	82.9
Green Line Extension	85.3	72.0
Purple Line	76.6	68.1
Blue Line (Bang Sue - Tha Phra)	79.1	66.0
Blue Line (Hua Lam Pong – Bang Care)	75.6	57.8

Source: MEA and calculated by BOT
Remark: as of June 2018

4. Conclusion

In the study, micro-data was derived to quantify demand for residential units in high-rise condominiums. To pursue the analysis, occupancy rate (OR) is constructed by using data from individual electricity meters. This source of data is regarded as new and unconventional for analytics. Findings discover, from 2014-18, the OR for Bangkok, Nonthaburi and Samut Prakarn had stood over 70 percent. This suggests that demand for residential units in the territories has remained robust. When taking aging aspect of buildings into account, the OR of residential units, which power meters registered prior to 2012, exhibits a mild declining trend, nonetheless, the ratio has been gyrating around 80-85 percent level. In relation to demand for new residence (proxy by using data from meters registered during 2013-17), the observed OR detects consistent uprise for more recently built units.

Besides, some property sector indicators such as "unsold units" and "time-to-go" variables are taken into consideration to enhance finding results more comprehensive. Analytics discover that 3 observed indicators have moved hand-in-hand with minor discrepancies in some locations. The OR also reflects similar direction. As illustrated above, the adoption of electricity meter data can be employed to assess occupancy rate (OR) of residential units which factually can be regarded as an additional useful economic indicator. It, however, should be detected along with other related barometers, namely number of new projects launched as well as units sold which will allow comprehending property market condition from up- to downstream. This

would, in turn, provide an embracive assessment of property market conditions in different period of time. Last but not least, in time to come, the study can be extended and improved the use of micro-level data to construct accurate and reliable economic indicators, this is not only confined to academic works but for policy attention and formulation.

References

1. Ecotagious Inc. (2016). Analysis of Housing Occupancy in the City of Vancouver Using Electricity Meter Data Analytics. *Stability in Vancouver's Housing Unit Occupancy*. February 2016.
2. Michael, A.,Turner, et.al. (2006). Give Credit Where Credit is Due: Increasing Access to Affordable Mainstream Credit Using Alternative Data. Information Policy Institute, Political & Economic Research Council.
3. Vipin, A., Jozef L. (2014). Electricity Use as an Indicator of U.S. Economic Activity. U.S. Energy Information Administration.

Semi-parametric single-index predictive regression

Hsein Kew[1], Weilun Zhou[1], Jiti Gao[1], David Harris[2]
[1]Monash University, Melbourne, Australia
[2]The University of Melbourne, Melbourne, Australia

Abstract
This paper proposes a semi-parametric single-index predictive model with multiple integrated predictors that exhibit cointegrating behaviour. We apply this predictive model to re-examine stock return predictability in the United States. It is well documented in the empirical finance literature that the most commonly used predictors (such as dividend-price ratio and earning-price ratio) can be characterised as integrated time series. We consider the case in which these integrated predictors can plausibly be modelled as cointegrated. We present some new evidence that the quarterly U.S. stock market returns are nonlinearly predictable when we account for cointegration among the predictors over the 1927-2017 periods and the post-1952 period.

Keywords
Stock return predictability; Single index model; Cointegration; Semi-parametric models

1. Introduction

Linear predictive models have been widely used in empirical economics and finance. For example, there is by now a large empirical literature that examines the predictability of stock returns using a variety of lagged financial and macroeconomic variables, including dividend-price ratio, earning-price ratio, dividend-payout ratio, book-to-market ratio, cay, interest rates, term spreads and default spreads; see for example Cochrane (2011), Lattau and Ludvigson (2001) and Rapach and Zhou (2013). They consider a multivariate predictive model of the form

$$y_t = \alpha + \beta^{\mathrm{T}} x_{t-1} + e_t$$

where y_t is the dependent variable, typically is the stock return at time t, x_{t-1} is a $p \times 1$ vector of predictors, typically is the lagged financial variables known at time $t - 1$, and e_t is an error term. They provide empirical evidence stock return are predictable because they reject the following null hypothesis of no predictability $H_0: \beta_1 = \cdots \beta_p = 0$.

Numerous studies, including Campbell and Yogo (2006) and Kostakis, Magdalinos and Stamatogiannis (2015) have found evidence that many of

these predictor variables are highly persistent and are often integrated of order one. However, they did not consider the case in which the predictor could potentially cointegrated; that is $\beta^{\mathrm{T}} x_{t-1} \sim I(0)$. We thus extend the linear multivariate predictive regression model, focusing on predictors that can plausibly be modelled as cointegrated and to allow the possibility that stock return depends in a nonlinear way on predictors.

We use Goyal and Welch updated quarterly data over the 1927-2017 sample period. The dataset were obtained from Amit Goyal's website at http://www.hec.unil.ch/agoyal. Their dataset is one of the most widely used datasets in research on stock return predictability. The dependent variable, y_t, is the US equity premium, which is defined as the log return on the S&P 500 index including dividends minus the log return on a risk-free bill.

2. Methodology

To allow for potential non-linearity and cointegration among the predictors, we consider a semiparametric single index model of the form

$$y_t = g_0\left(\theta_0^{\mathrm{T}} x_{t-1}\right) + e_t$$

where $x_t = \left(x_{1,t}, \ldots, x_{d,t}\right)^{\mathrm{T}}$ is a vector of d-dimensional potential integrated predictors, $g(.)$ is an unknown nonlinear integrable function and is often called the link-function in the literature, θ_0 is the single index parameter such that $\theta_0^{\mathrm{T}} x_{t-1}$ is stationary and e_t is a martingale difference sequence. Thus our model allows for the presence of cointegration among the integrated predictors. Also our model includes the linear parametric multivariate predictive model as a special case since function $g(.)$ can take a linear form. In the case of a univariate integrated predictor with $\theta_0 = 1$, Kasparis, Andreou and Phillips (2015) established the statistical theory for the estimation of the $g(.)$ function. Following the estimation procedure discussed in Dong, Gao and Tjostheim (2016), a profile approach is used to derive the estimators of the unknown link-function and the unknown single index parameters.

Among the 14 financial and macroeconomic variables that Goyal and Welch (2008) use to predict the equity premium, we consider the following four pairs of potentially cointegrated variables: (a) dp and ep; (b) 3-month T-bill rate (tbl) and long-term yield (lty); (c) baa and aaa rated corporate bond yields; and (d) dp and dividend yield (dy). Goyal and Welch (2008) provide the definitions and sources of these predictors.

3. Result

For initial illustration, Figure 1 plots those four pairs of variables using quarterly data in the subperiod 1952--2017 and demonstrates that the two series in each of the four pairs considered are positively correlated and they

appear to be cointegrated. As mentioned in the introduction, the use of the first pair to predict equity premium is motivated by a casual glance at Figure 1 of Campbell and Yogo (2006), who show that dp and ep could cointegrate. Fama and French (1989) use the term spread (defined as tbl minus lty) and the default spread (defined as baa minus aaa) to predict the equity premium and under the assumption that these spreads are stationary, their work implies cointegrating relationships between tbl and lty and between baa and aaa.

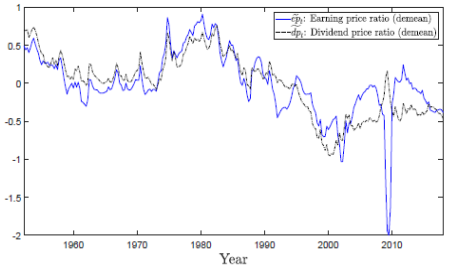
Preliminary Augmented Dickey-Fuller (ADF) test indicates that every variable has a unit root and the Engle-Granger ADF test suggests the existence of cointegration in each of the four pairs. These tests provide statistical evidence supporting the impressions of cointegrating relationships from visually inspecting Figure 1. We then proceed to test the hypothesis that the US equity premium is predictable using cointegrated predictors. By applying Hermite polynomial expansion, we can re-write single index model as

$$y_t = g(u_{t-1}) + e_t \approx \sum_{i=0}^{k-1} d_i \, u_{t-1}^i \; + \; e_t$$

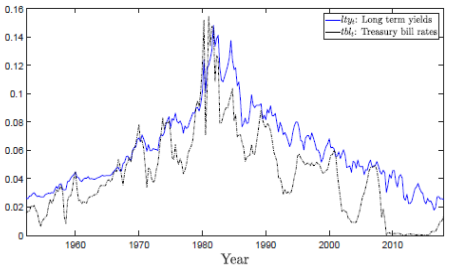
where $u_{t-1} = \theta_0^{\mathrm{T}} \, x_{t-1}$ and the truncation parameter k determined by the Generalised Cross Validation method (see Gao, Tong and Wolff (2002)). Under the null hypothesis of no predictability, $d_1 = d_2 = \cdots = d_k = 0$ and so the model reduces to the constant expected equity premium model. Given that $u_t \sim I(0)$, the no predictability null hypothsis can be tested using the heteroscedasticity robust F-statistic. The OLS coefficient estimates and their White standard errors can be obtained in the standard way from a multiple regression of y_t on the lagged of u_t .

For each pair of variables, we estimate the single index predictive model and report in Table 1 the least squares estimates of the coefficients, the results of the F-tests under the null hypothesis of no predictability and the adjusted R^2 statistic for each pair. Numbers in parentheses below the coefficients are t-ratios (based on White standard errors) and below the F-tests are p-values. Panel A reports the results for the whole sample period 1927-2017. Following Kostakis, Magdalinos and Stamatogiannis (2015), we also consider the post-1952 period because the interest rate variables are expected to be linked together after the Federal Reserve abandoned the interest rate pegging policy in 1951. Moreover, Kostakis, Magdalinos and Stamatogiannis (2015) and Campbell and Yogo (2006) report weak or no evidence of stock return predictability in the post-1952 period. Our results for this subperiod are reported in Panel B.

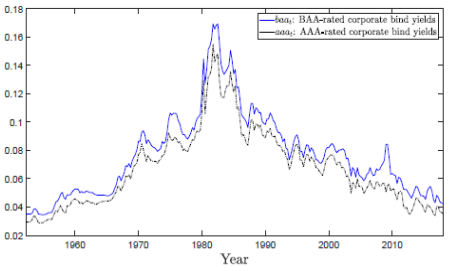
Figure 1: Time series plots of cointegrated predictors



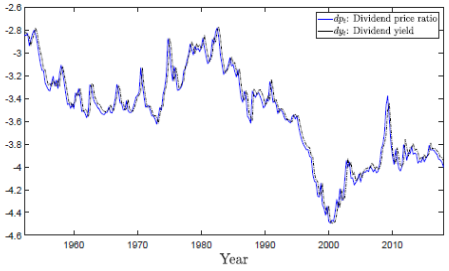
(a) ep_t and dp_t (demean)



(b) lty_t and tbl_t



(c) baa_t and aaa_t



(d) dp_t and dy_t

Notes: This figure plots the following four pairs of cointegrated predictors: (a) demean ep_t (earning price ratio) and dp_t (dividend-price ratio), (b) tbl_t (annualized 3-month T-bill rate) and lty_t (annualized long-term yield), (c) annualized BAA_t and AAA_t -rated corporate bond yields, and (d) dp_t and dy_t (dividend yield). The sample period is 1952:Q1 through 2017:Q4.

Table 1: Estimates of the single index model parameters and predictive test

Pair of predictors	$\widehat{d}_0$	$\widehat{d}_1$	$\widehat{d}_2$	$\widehat{d}_3$	$\widehat{d}_4$	F-test	R^2
			Panel A: 1927Q1 - 2017Q4				
ep and dp	0.0221***	0.0685**	−0.2288**	−0.1998***		4.1762	0.03
	(3.3005)	(2.3507)	(−2.2317)	(−2.9103)		(0.0063)	
lty_t and tbl_t	0.0167***	0.5449				2.1357	0.00
	(2.9890)	(1.4614)				(0.1448)	
BAA_t and AAA_t	0.0128*	9.0449**	−1602.0866***	52 371.3111***		11.2738	0.08
	(1.8883)	(2.5737)	(−4.0084)	(4.9546)		(0.0000)	
dp and dy	0.0223***	−0.4421***	0.7292**	7.0280***		10.2787	0.07
	(3.5254)	(−3.8853)	(2.2705)	(5.4615)		(0.0000)	
			Panel B: 1952Q1 - 2017Q4				
ep and dp	0.0192***	0.0585***	−0.0942**	−0.0848***		4.1318	0.03
	(3.3997)	(2.7383)	(−1.9821)	(−2.8597)		(0.0069)	
lty and tbl	0.0035	3.3153***	−30.5827	−3713.8087**		4.8013	0.04
	(0.5064)	(3.7572)	(−1.1124)	(−2.5512)		(0.0028)	
BAA and AAA	0.0234***	3.0394**				3.9085	0.01
	(3.6901)	(1.9770)				(0.0491)	
dp and dy	0.1483***	−1.2062***	2.4386***			6.7875	0.04
	(4.0082)	(−3.4235)	(3.0305)			(0.0013)	

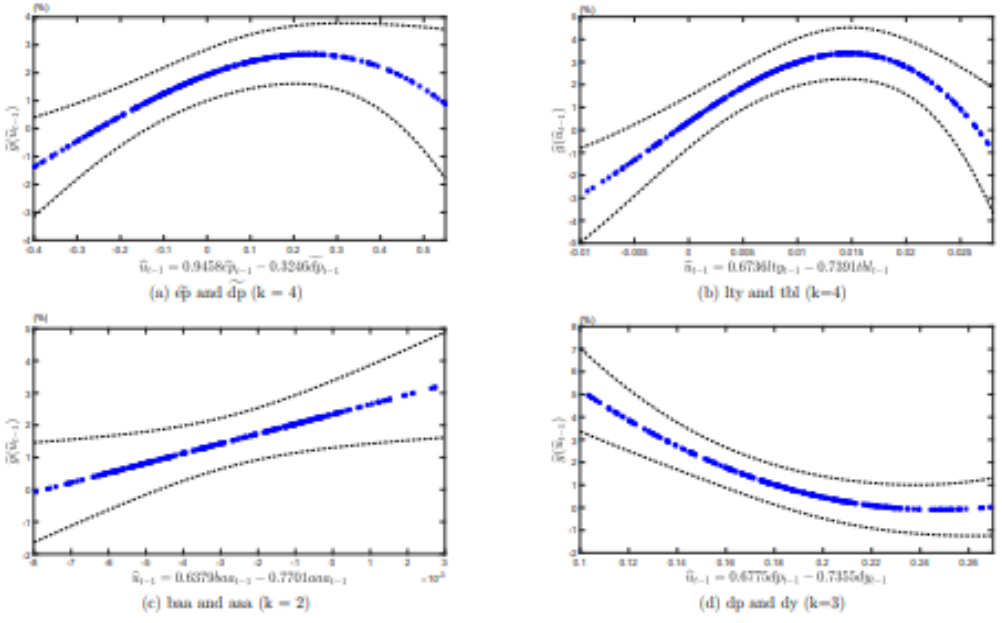
In Table 1, using the F-tests, we reject the null of no predictability at the 5% level in both the full sample and the post-1952 sample for all four pairs, with one exception. The pair of lty and tbl has no predictive ability for equity premium in the full sample but this pair is a strongly statistically significant predictor in the post-1952 sample with a p-value of 0.0028. Our result

supports the view that the term-structure variables are closely linked together after 1952 but not before.

We compare our results to studies that provide in-sample evidence regarding stock return predictability. Using the full sample, our finding of significant predictability is consistent with those of prior studies. Using the post-1952 period, however, while Campbell and Yogo (2006), Kostakis, Magdalinos and Stamatogiannis (2015) and Kasparis, Andreou and Phillips (2015) have found no or weak evidence of predictability in the post-1952 period using linear or non-linear predictive regressions with a single predictor or multiple predictors without allowing for the presence of cointegration, we do find strong evidence using bivariate cointegrated predictors.

Furthermore, the results in Table 1 provide ample evidence in favour of nonlinear predictability of stock returns using some pairs of cointegrated predictors since the coefficient on the highest power in the polynomial regression is statistically significant at conventional levels. To illustrate the approximate form of nonlinearity, Figure 2 plots the predicted value of equity premium, $\hat{g}(\hat{u}_{t-1})$ against $\hat{u}_{t-1}$ along with the 90 percent pointwise confidence intervals. This figure shows that the two pairs of lty-tbl and of ep-dp exhibit a hump-shaped relationship. This empirical finding of nonlinear predictability using these two pairs of cointegrated predictors highlights a useful feature of our semi-parametric single index predictive model.

Figure 2: Estimated link function $\hat{g}(\hat{u}_{t-1})$ at quarterly frequency



Notes: This figure plots the estimated link function of each pair of cointegrated predictors. The dashed line shows the approximate 90 percent pointwise confidence interval and the horizontal line depicts the average quarterly equity

premium (which is 1.52%). The confidence interval is constructed by the procedure described in Section 3. The sample period is 1952:Q1 through 2017:Q4.

As an illustration of the nonlinear predictability, consider the pair of quarterly lty and tbl with the cointegrating relation $\hat{u}_{t-1} = 0.674lty_{t-1} - 0.739tbl_{t-1}$ shown in Figure 2. The average quarterly equity risk premium in the post-1952 sample is 1.52 percent. The predicted value of quarterly equity premium, $\hat{g}(\hat{u}_{t-1})$, exceeds this 1.52 percent beginning at $\hat{u}_{t-1}$ = 0.0038 and then peaks at 3.39 percent at $\hat{u}_{t-1}$ = 0.0147. This peak occurs in the first quarter of 2016 when the annualized long-term yield is 2.43% and 3-month T-bill rate is 0.23%.

Panel A in Table 1 shows that the pair of baa and aaa gives the largest quarterly $\bar{R}^2$ in the full sample and this pair explains 8 percent of the variation in next quarter equity premium. As in previous empirical studies, we document in Table 1 small $\bar{R}^2$ statistics but they can signal economically significant predictability, as explained in Campbell and Thompson (2008) and Fama and French (1988).

4. Conclusion

This paper considers a semi-parametric single index predictive model with cointegrated predictors. We apply our model to study the predictability of U.S. stock returns. We provide new evidence that quarterly stock returns are predictable using the following pairs of cointegrated predictors: earning-price ratio and dividend-price ratio; 3-month T-bill rate and long-term yield; baa and aaa rated corporate bond yields; and dividend-price ratio and dividend yield using data over the 1927-2017 period and the post-1952 period.

References

1. Campbell, J. Y. and Yogo, M. (2006), `Efficient tests of stock return predictability', Journal of Financial Economics 81(1), 27-60.
2. Cochrane, J. H. (2011), `Presidential address: Discount rates', The Journal of Finance 66(4), 1047-1108.
3. Dong, C., Gao, J. and Tjostheim, D. (2016), `Estimation for single-index and partially linear singleindex integrated models', The Annals of Statistics 44(1), 425-453.
4. Fama, E. F. and French, K. R. (1989), `Business conditions and expected returns on stocks and bonds', Journal of Financial Economics 25(1), 23-49.
5. Gao, J., Tong, H. and Wol_, R. (2002), `Model specification tests in nonparametric stochastic regression models', Journal of Multivariate Analysis 83(2), 324-359.
6. Kasparis, I., Andreou, E. and Phillips, P. C. (2015), `Nonparametric predictive regression', Journal of Econometrics 185(2), 468-494.
7. Kostakis, A., Magdalinos, T. and Stamatogiannis, M. P. (2015), `Robust econometric inference for stock return predictability', Review of Financial Studies 28(5), 1506-1553.
8. Lettau, M. and Ludvigson, S. (2001), `Consumption, aggregate wealth, and expected stock returns', The Journal of Finance 56(3), 815-849.
9. Rapach, D. and Zhou, G. (2013), Forecasting stock returns, in `Handbook of Economic Forecasting', Vol. 2, Elsevier, pp. 328-383.

Comparative analysis of R&D statistical systems between China and major developed countries

Zhu Yingchun[1], Liu Huifeng[1], Sun Yunjie[1], Wu Da[2]
[1]Chinese Academy of Science and Technology for Development, Beijing, China
[2]Tianjin Research Centre for Statistics and Science and& Technology Development, Tianjin, China

Abstract
This article discusses the R&D statistical system of China, including its statistical structure, objects, scope and methodologies; in addition, it also analyses the main features of organizational systems of R&D statistics of major developed countries, such as, the U.S., Japan, Germany and the U.K. Compared to these countries, China is quite close to them in terms of levels of statistical training and data auditing, albeit only a short history of 30 years of its R&D statistics. Nevertheless, China should reference the experience of developed countries, using methods like, estimations and sampling to conduct R&D surveys in a flexible and diverse manner.

Keywords
R&D statistical systems; scope of statistics; statistical methodologies

1. Overviews and features of China's R&D statistical system

1.1 Evolution of China's R&D statistical system

In the 1970s, Chinese experts and scholars began to engage in relevant research on science and technology (S&T) statistical methodologies. In 1978, the National Bureau of Statistics (NBS), the State Commission for Science and Technology (SCST) and other two departments jointly organized what was called, "General Survey of Employees Working in Natural Science and Technology in China", through which to gasp the basic situation of personnel working in the field of natural science and technology. In 1985, SCST along with NBS and other related government agencies introduced for the first time ever relevant definitions on indicators and standards of classification from UNESCO's Manual for Statistics on Science and Technology Activities, and they co-implemented the "General Survey on Science and Technology in China". Since then, the S&T, education and statistics authorities in China have set up three independent S&T statistics systems respectively for government research institutes, higher education institutions, and medium and large industrial enterprises, with the establishment of level-by-level survey systems on the regional basis. And this survey became the cornerstone of S&T statistics in China. In 1988, SCST organized experts to revise the S&T statistics indicators and carried out a nation-wide R&D investment sample survey, through which

data on R&D activities of China's medium and large industrial enterprises, government research institutes and higher education institutions were garnered systematically for the first time. In 1991, the NBS established the national S&T consolidated report system and promulgated the national R&D congregate data and structured data for the first time. The system of S&T statistics set up in 1991 based on the executive arms of S&T activities is in use today. In 2000, seven government departments including the Ministry of Science and Technology (MOST) jointly undertook the first time National R&D Resources Inventory, with the survey scope covering organizations with R&D activities in various sectors and industries of the national economy. This inventory was considered an important milestone in the history of China's S&T statistics development. In 2009, China conducted its second National R&D Resources Inventory to further improve its R&D statistics system, giving more prominence to the R&D indicators in the S&T statistics and thus pushing forward the S&T statistics to extend to the R&D statistics.

1.2 China's R&D statistical work system

The R&D statistical work system in China was set up based on its executive department of S&T activities, with its MOST taking in charge of statistics on government research institutes and public institutions engaging in scientific research and technology service, with NBS being responsible for statistics on the S&T activities of the enterprise, with MOE for statistics on the S&T activities of higher education institutions, with the State Commission of Science and Technology for National Defence Industry for statistics on government research institutes in the sector of science and technology industry of national defence, and with NBS for consolidating statistics data around the country (see Fig. 1).

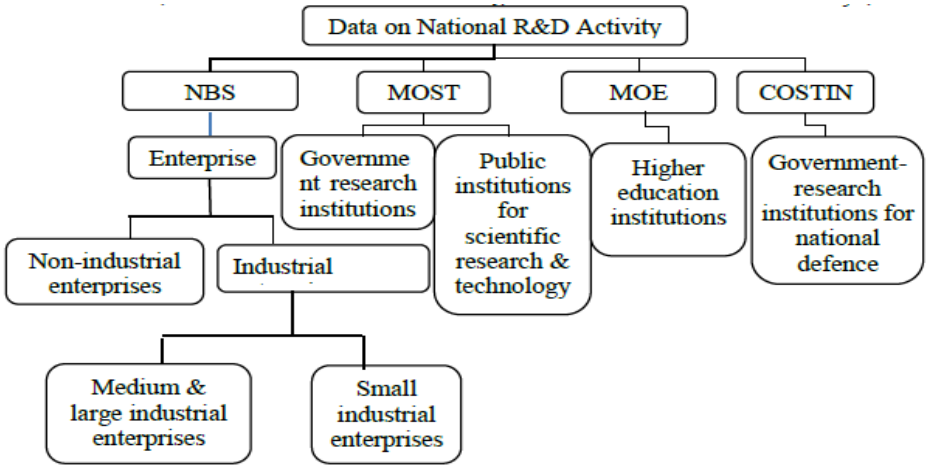


Fig.1 China's R&D Statistical System

1.3 China's R&D statistical scope and survey methodologies

Prior to 2009, China's R&D statistics only covered government research institutes, medium and large industrial enterprises and higher education institutions, with a lack of statistics on its agricultural enterprises, construction enterprises and service enterprises. In 2009, China undertook its second National R&D Resources Inventory, with the statistics objects being the legal entities in the R&D activity intensive industries of the national economy. The scope of survey has, henceforth, been constantly expanded, with the major medium and large enterprises in the service industry also incorporated into the annual statistical survey (see Table 1).

Table 1. Scope of China's R&D Statistics

Year	Survey	R&D Survey Scope
2009	2nd R&D Resources Inventory	Legal entities in agriculture, forestry and fishing; mining; manufacturing; production & supply of power, fuel gas and water; construction; transport, warehousing & postal; information transmission, computer service and software; financial; leasing & commercial service; scientific research & technology service; water conservancy, environmental & public utilities management; education; health & social activity; culture, sports & entertainment.
2013	3rd Economic Census	1. Higher education institutions (incl. affiliated hospitals); 2. Above- designated industrial enterprises; 3. Above designated size construction enterprises; 4. Enterprises and public institutions in the scientific research and technology service sector; 5. Above designated legal entities enterprises in transportation, warehousing & postal; information transmission, computer service and software; leasing and commercial service; water conservancy, environmental & public utilities management; health & social activity, cultural, sports and entertainment industries.
2014-2016	Annual Statistical Survey	The same as in 2013.
After 2017	Annual Statistical Survey	1. Higher education institutions (incl. affiliated hospitals); 2. Above designated industrial enterprises; 3. Enterprises & public institutions in scientific research and technology service; 4. Above designated size building enterprises; 5. Medium & large legal enterprise entities in transportation, warehousing & postal; information transmission, computer service and software; leasing and commercial service; water conservancy, environmental & public utilities management; health & social activity, cultural, sports and entertainment industries..

The survey method combines the complete enumeration with the census. In addition, higher education institutions, above designated industrial

enterprises, enterprises and public institutions in the scientific research and specialized technology service industries as well as major service industries use the complete enumeration on an annual basis, and data for other industries for non-survey years using the previous census data as the substitutes to ultimately got the consolidated national data.

2. Main features of R&D statistical systems of major developed countries

Developed countries began their research on relevant concepts and measurement methods for R&D activities in the early 1940s, and they split the R&D statistics from the socioeconomic statistics during 1950s-1960s according the conceptual framework in OECD's Frascati Manual, and established an independent statistics system.

2.1 Government agencies play an important part in R&D statistics

Generally speaking, R&D statistics of various countries are conducted by their government agencies, although the R&D statistical organization mechanism in some countries such as, Japan is characterized by coordination between government and non-government institutions with the latter taking the responsibility of statistics. And the R&D statistics organization mechanisms of various countries can be divided, based on the nature of department, into the following two categories: first is the government comprehensive statistics department (i.e. Bureau of Statistics) responsibility system, under which the Statistics and Survey Department of the Bureau of Statistics of the Ministry of Internal Affairs and Communications of Japan, for instance, is responsible for organization and implementation, and in the case of the U.K., the Department of Business, Energy & Industrial Strategy (BEIS) takes the responsibility; second is the government agency managing or financing S&T activities responsibility system , under which the National Science Foundation (NSF)of the U.S. takes in charge of it, for instance.

2.2 Multiple measures are taken to limit the statistical scope within a reasonable range

Statistics objects of various countries are basically divided into four categories, which are enterprises, government agencies (including government research institutes), higher education institutions, and private non-profit sectors. Multiple methods and measures are taken by various countries for these four types of statistical objects according to their actual situation, with the institutions engaging in no or little R&D activities excluded from the scope of survey and the volume of their statistical work reduced on the premises of not affecting the data accuracy. Typically, the below methods are taken: (1) for government agencies, the scope of statistics is limited to the

government-sponsored research institutions, laboratories and centers; (2) for the education sector, the scope of statistics is limited to those institutions engaging in S&T activities in an organized manner, such as, colleges and universities; (3) in terms of industry coverage, the industries conducting little R&D activities are not incorporated into the scope of statistics; (4) set up the starting point of surveys, set up the limiting conditions for the quantitative indicators, such as, staff and expenditure, and include the objects only meeting the survey starting point in the scope of statistics (see Table 2).

Table 2. R&D Statistics Scopes of Some Countries

	Government sector	Higher education sector	Enterprises
US	R&D centers sponsored by the federal government, various federal government agencies & their	All colleges and universities with annual R&D budget over US$150,000	All enterprises with the number of employees over 5.
	subordinated institutions; state government		
Japan	Central & local government research institutes in natural science, social humanities	Comprehensive universalities, junior colleges, research institutes affiliated to universities, joint research institutions of universities, laboratories	Enterprises with the registered capital over 10 million Yen
Germany	Government research institutes, government-sponsored independent non-profit institutions and independent research institutions affiliated to colleges and universities.	Relevant colleges & universities, teachers' universities, art colleges, management colleges, divinity schools, university hospitals, federal and state institutions belonging to university budget engaging in college management or college staff management.	All enterprises engaging in R&D activities (including those doing so internally or by entrusting external R&D institutions with doing so).
UK	Various departments in the governments, local governments	Estimations	Enterprises within the scope of registration

Notes: the private non-profit sector accounts for a very little proportion in the total amount, which is not included in the table, due to lack of data.

2.3 Most countries use the complete survey on research institutes and higher education institutions while the sample survey on enterprises

In terms of methodologies used by various countries for their R&D statistics, there are mainly complete surveys, sample surveys and estimations. And there are two scenarios for the complete survey. First is conducting the

complete survey of all the institutions within the scope of statistics, which is only suited to the relatively fewer institutions within the scope. Currently, this method is generally used for the survey of government research institutes and higher education institutions. When this method is used, it is typically necessary to set up the survey starting point within the given scope of statistics and further narrow the scope to an acceptable range, for instance, the US's R&D statistics on colleges and universities. Second is conducting the complete survey of institutions with R&D activities within the scope. This is a kind of method most frequently used at present. When this method is used for a survey, it is necessary to determine in advance whether the object engages in the R&D activity. Most countries use the sample survey (stratified sampling) for collecting data on enterprises R&D activities, due to huge numbers of respondents in the enterprises; countries, such as, the U.S., the U.K. and Japan, for instance, stratify their enterprises by the industry, R&D scale and staff numbers. For the enterprises survey, most developed countries use the method of "long questionnaire" combined with "short questionnaire" and they usually ask the key surveyed enterprises to fill out the "long questionnaire" and other enterprises to fill out only the "short questionnaire". This will greatly reduce the workload while ensuring the timeliness and convenience of the garnered data. As the supplement to the complete survey and the sample survey, estimations are made by some countries in their sectoral or industrial surveys, such as, the R&D statistics by the UK on its local government sectors and higher education sectors (see Table 3).

Table 3. R&D Statistical Methodologies of Some Countries

	Government Sector	Enterprises	Higher education sector
US	Complete survey	Sample survey	Complete survey
Japan	Complete survey	Sample survey	Complete survey
Germany	Complete survey	Sample survey	Complete survey
UK	Estimations	Sample survey	Estimations

Notes: the private nonprofit sector accounts for a very small proportion in the total amount, whose statistical method is not included, due to lack of data.

2.4 The survey of service industry is given much attention

The R&D statistics of the above four countries involves extensive industries with collecting the data on non-manufacturing industry, except for the manufacturing industry; especially in recent years, with the vigorous development of the service sector, various countries have given sufficient emphasis on the R&D statistics on the service industry. In Japan, for instance,

only 3 industries were covered in 2000, but the coverage was increased to 7 in 2005; Germany increased its coverage of 3 industries in 1993 to 8 in 2005.

3. Major conclusions and revelations

3.1 China's R&D statistics is currently up to the level of developed countries in certain respects

Through nearly three decades of its development, China's R&D statistics has grown out of nothing, then achieved excellence and gradually established a relatively perfect and internationally accepted R&D indicator system with a standardized statistical procedure, as well as a well-trained professional team of R&D statistics. Albeit a relatively short history of its R&D statistics, compared to developed countries, China has become relatively mature in R&D statistical practices, such as, training, auditing and other related aspects in order to ensure data quality and even surpassed part of develop countries in one way or another. Every year, China collects data of over 2,000 higher education institutions, over 3,000 government research institutes and over 400,000 enterprises. Statistical data are filled out by objects and checked and verified by statistics specialists at the city, provincial and national levels. The computer programs combined with manual checks are used for auditing and checking purposes. In order to ensure the quality of the reported data, China on an annual basis conduct training in statistical indicators and reporting methods for the objects of statistics and urban and provincial statistics staffs. In addition, China developed an improved statistical regulatory framework to guarantee the authenticity and effectiveness of its statistical data.

3.2 China should draw lessons and experiences in certain aspects from developed countries

In recent years, a series of new features have appeared in China's R&D activity. Firstly, with the deepening of implementation of the new policy of "Mass Entrepreneurship and Innovation", there has been a vigorous development of innovative activities of small and micro businesses at below designated size, with a constant flow of new industries, new industrial formats and new modalities. Secondly, with its talent and market advantages, China has attracted a large number of MNCs to establish R&D centres in China to undertaking R&D activities. Thirdly, there has been a vigorous development of non-state and non-enterprise research institutions providing the society with R&D outsourcing and professional technological services. Currently, R&D activities of small and micro businesses at below designated size, foreign funded R&D institutions and emerging R&D institutions have yet to be incorporated into the scope of statistics, and the R&D statistics system is therefore unable to quickly reflect the changes in the development of new industrial formats, new modality and new industries during mass

entrepreneurship and innovation. For this reason, it is necessary to accelerate the establishment of a modernized service-oriented R&D statistics system which is adaptive to new normal. In this regard, we may, in terms of statistics scope , use relevant experience of developed countries for reference; we may further strengthen the statistics of the service industry and incorporate various more sectors such as, small and micro businesses at below designated size, make space and new types of R&D institutions into the scope of statistics; and we may, in terms of specific statistical methodologies, use sample surveys and estimations on the respondents that are newly incorporated into the statistics system, adopting the method of filling out the "short questionnaire".

References

1. Department of Development & Planning, Ministry of Science and Technology, China Association of Science and Technology Indicators. User Manual of Science and Technology Statistics [M]. Beijing: Science and Technology Literatures Press, 2008:27.
2. National Science Foundation. NESES Surveys [EB/OL].[2016-12-20].https://www.nsf.gov/statistics/surveys.cfm.
3. Bureau of Statistics, Ministry of Internal Affairs and Communications: Science and Technology Research and Survey. "Survey Abstract" [EB/OL].[2016-11-24]. http://www.stat.go.jp/data/kagaku/gaiyou/index.htm#gaiyou12.
4. Bureau of Statistics, Ministry of Internal Affairs and Communications: Science and Technology Research and Survey. "Survey Results" [EB/OL]. [2017-1-27]. http://www.stat.go.jp/data/kagaku/kekka/index.htm.
5. National Bureau of Statistics. Transformation of the British Statistics [J] Data. 2010(9):30-31.
6. Department for Business Innovation & Skill. SET statistics: science, engineering and technology statistics 2013[EB/OL]. [2017-27] https://www.gov.uk/government/statistics?departments%5B%5D=department-for-business-innovation-skills&page=2&publication_filter_option=statistics.
7. National Bureau of Statistics, Office for the National R&D Resources Inventory. Document Assembly of the 2009 Second National R&D Resources Inventory (Synthesis) [M]. Beijing: China Statistics Press. 2011(6).83.

A statistical modelling framework for mapping malaria seasonality

Michele Nguyen[1], Jennifer Rozier[1], Suzanne Keddie[1], Rosalind E. Howes[1], Timothy C. D. Lucas[1], Daniel J. Weiss[1], Katherine E. Battle[1], Peter W. Gething[1], Ewan Cameron[1], Harry S. Gibson[1], Mauricette Andriamananjara Nambinisoa[2]
[1]Malaria Atlas Project, Big Data Institute, University of Oxford, Oxford, UK
[2]National Malaria Control Programme, Antananarivo, Madagascar

Abstract

Many malaria-endemic areas experience seasonal fluctuations in cases because the mosquito vector's life cycle is dependent on the environment. While most existing maps of malaria seasonality use fixed thresholds of rainfall, temperature and vegetation indices to find suitable transmission months, we develop a spatiotemporal statistical model for the seasonal patterns derived directly from case data.

A log-linear geostatistical model is used to estimate the monthly proportions of total annual cases and establish a consistent definition of a transmission season. Two-component von Mises distributions are also fitted to identify useful characteristics such as the transmission start and end months, the length of transmission and the associated levels of uncertainty. To provide a picture of "how seasonal" a location is compared to its neighbours, we develop a seasonality index which combines the monthly proportion estimates and existing estimates of annual case incidence. The methodology is illustrated using administrative level data from the Latin America and Caribbean region.

Keywords

Seasonality; Spatiotemporal Statistics; Geostatistics; Infectious diseases; Malaria

1. Introduction

Malaria is a disease caused by the Plasmodium parasite and remains a major cause of child mortality in sub-Saharan Africa (World Health Organisation 2018). Like that of many other infectious diseases, malaria transmission exhibits seasonality across endemic areas. Understanding location-specific seasonal characteristics is useful for maximising the impact of interventions, developing early warning systems as well as improving models relating indicators of transmission and disease (Stuckey et al. 2014).

To this end, maps of malaria seasonality have been developed. By using thresholds on environmental factors, one can determine the months suitable for transmission (Cairns et al. 2012, Gemperli et al 2006). Since seasonal

malaria chemoprevention has been shown to be most effective when delivered over three months, these maps can be useful for targeting such interventions (Cairns et al. 2012).

Despite their functionality, the threshold-based maps have several limitations. Although the environment is a key driver of seasonality, there are other contributors such as migration (Martinez 2018). The same environmental factors could also affect different areas differently: rain, for example, can both create and wash away mosquito breeding sites depending on the local topology and rainfall intensity (Martinez 2018). Using environmental thresholds does not allow for other potential drivers or account for the variation of responses.

Another class of seasonality maps relates to concentration indices. To quantify the distribution of malaria cases in each district over a year, Mabaso et al. (2005) used Markham's concentration index which was previously used to determine rainfall concentrations. Their concentration maps from the case numbers estimated using a Bayesian spatiotemporal regression model displayed clearer spatial patterns than those derived from raw case numbers. Spatiotemporal models smooth out idiosyncratic deviations to enable us to focus on the main seasonal trend. They are also useful for relating the seasonality to input covariates and account for unknown spatiotemporal effects.

In this paper, we present a modelling framework for a cohesive and evidence-based analysis of malaria seasonality. Using a spatiotemporal geostatistical model, we obtain maps of various seasonality measures including the number of transmission periods in a year, as well as start and end months of each transmission season. Unlike previous work, we also present the uncertainty associated with each map. A seasonality index from the rainfall literature is adapted to give a visual impression of both the distribution and magnitude of malaria cases over a year. The methodology is illustrated using administrative level data from the Latin America and Caribbean (LAC) region.

2. Methodology

2.1 Original seasonality index

As suggested by Feng et al. (2013), "how seasonal" a location j is can be expressed as the product of an entropy measure (D_j) and the relative amplitude ($\frac{R_j}{R_{max}}$):

$$S_j = D_j \times \frac{R_j}{R_{max}},$$

$$where\ D_j = \sum_{i=1}^{12} p_{i,j}\, log_2\left(\frac{\bar{p}_{i,j}}{q_i}\right).$$

Here, $p_{i,j}$ is the average proportion for month i and $q_i = 1/12$, for $i = 1, \ldots, 12$. So, D_j quantifies how different the monthly proportions are from a uniform distribution over the year. In the context of malaria, R_j can represent the annual case or parasite incidence (API; the number of cases per 1000 people in a year) at location j and R_{max} the maximum API over the region. Since S_j separates the timing and amplitude aspects of seasonality and there are existing maps of API in the literature, we can focus on modelling the monthly proportions at each location, $p_{i,j}$.

By modelling proportions instead of the number of cases as done by Mabaso et al. (2005), we bypass the need to estimate catchment populations. This is useful when working with health facility data, which is becoming increasingly available in multiple countries. In addition, we restrict our analysis to dynamic environmental covariates like temperature while omitting static factors such as elevation.

2.2 Spatiotemporal monthly proportion model

From incidence data, we compute monthly median case counts at each location over a set number of years and obtain the monthly proportions by dividing the medians by their annual sum. The following spatiotemporal model is used to estimate monthly proportions over our study region:

$$\log(p_{i,j}) = X_{ij}^T\boldsymbol{\beta} + \phi_{ij} + \epsilon_{ij},$$

where X_{ij} is a m-dimensional covariate vector including an intercept, $\boldsymbol{\beta} \in \mathbb{R}^m$ is the corresponding parameter vector and $\epsilon \sim N(0, \sigma_e^2)$ denotes independent, identically distributed noise. The spatiotemporal Gaussian field ϕ is constructed as follows:

$$\phi_{ij} = \begin{cases} \xi_{1j}\ for\ i=1 \\ a\phi_{i,j-1} + \xi_{i,j}\ for\ i=2,\ldots,12, \end{cases}$$

and $\xi_{i,j}$ correspond to zero-mean Gaussian innovations which are temporally independent but spatially coloured with a Matérn covariance. Note that rescaling of the estimated proportions is required to ensure that they sum to one at each location. This is consistent with the fact that some locations are more or less sensitive to the variation in the underlying factors.

Before fitting the model, we exclude outliers ($\log(p_{i,j}) \leq -11\ for\ LAC$) to model prototypical seasonal behaviour. We also compute monthly medians

for our covariates over the study years and standardise them. The covariates we consider are rainfall from the Climate Hazards Group Infrared Precipitation and Station data (CHIRPS), enhanced vegetation index (EVI), daytime land surface temperature (LST_Day), diurnal difference in land surface temperature (LST_delta), night-time land surface temperature (LST_night), tasselled cap brightness (TCB), tasselled cap wetness (TCW) as well as the temperature suitability indices for *Plasmodium falciparum and Plasmodium vivax* (TSI_Pf and TSI_Pv). The data sources are detailed elsewhere (Kang et al. 2018). To account for delayed and accumulated responses to these environmental variables, we also test them at 1-3 month lags.

We set aside 30% randomly selected sites for validation. Working with the rest of the data, we reduce the set of covariates and account for multicollinearity by iteratively computing the variance inflation factors (VIFs) and removing the covariates with the highest VIF value until all the remaining covariates have VIF values less than 10. Next, we fit the model in R using integrated nested Laplace approximation (INLA) (Lindgren et al. 2011, Kang et al. 2018). Backwards regression using the Deviance Information Criterion (DIC) is used to select the best parsimonious model.

2.3 Deriving seasonality statistics

After checking that the model performs well on both the training and test data in terms of coverage probabilities and root mean squared errors, we refit the chosen model using all of the data. Seasonality statistics are derived for each posterior sample of the location-specific monthly proportions.

We regard a location as potentially seasonal if its entropy $D_j > 0$. If this is satisfied, we fit a rescaled, two-component von Mises (R2vM) density to the monthly proportions. Rewriting the month in a year as a random variable on a circle, $\theta = \frac{2\pi i}{12}$ where $i = 1, \ldots, 12$, the R2vM function is defined as: 12

$$f(\theta, s\,\omega, \mu_1, k_1\,\mu_2, k_2) = s[\omega f_1(\theta, \mu_1, k_1) + (1 - \omega) f_2(\theta, \mu_2, k_2)]$$

$$\text{where } f_k(\theta, \mu_k, K_k) = \frac{1}{2\pi I_0(K_k)} \exp\{K_k \cos(\theta - \mu_k)\},$$

and μ_k and K_k are the mean and concentration parameters of the k^{th} component respectively (k = 1, 2). Here, I_0 is the modified Bessel function and ω is a probability weight. The scale parameter $s > 0$ modulates between the continuous density function and monthly proportions over discrete months.

Using a circular distribution provides a continuous curve between the months of January and December. Using a two-component von Mises density, in particular, is convenient for identifying the peaks of the bimodal distribution since these correspond to the mean parameters (Pewsey et al. 2013). Although

an arbitrary number of von Mises components can be used, we only use two because areas with seasonal malaria transmission typically have one or two main seasons (Stuckey et al. 2014).

Based on the fitted R2vM curve, we identify the transmission periods by marking the months where the curve is at or above 1/12. In this way, we can also obtain the start, end and length of each season. If there is only one transmission period, we fit a rescaled, one component von Mises density to the monthly proportions and estimate the seasonality statistics based on this instead.

To obtain the uncertainty associated with the derived statistics, we summarise the results from 100 posterior samples of the monthly proportions. By looking at the proportion of times a location is deemed bimodal or unimodal, we can obtain the majority decision as well as the degree of certainty. Based on this, we can analyse the uncertainty in the estimated seasonal characteristics. For the start, end and length of the transmission, we can obtain the means and standard deviations.

2.4. Adjusted seasonality index

The seasonality index introduced in Section 2.1 does not work well for bimodal distributions since the entropy does not account for the two peaks and only considers the overall distribution in a year which typically appears as more even than a unimodal one. To better reflect the degree of seasonality, we adapt the entropy for bimodal distribution at location j using the fitted R2vM function as follows:

$$\widetilde{D}_j = \omega D_j^{(1)} + (1 - \omega) D_j^{(1)},$$

$$where\ D_j^{(k)} = \sum_{i=1}^{12} f_k(\theta, \mu_k, K_k)\, log_2\left(\frac{f_k(\theta, \mu_k, K_k)}{q_i}\right)$$

for k=1,2 and the terms are as defined before. For consistency, we also base the adjusted seasonality index for unimodal distributions on the fitted one component von Mises density.

3. Results

We illustrate our method using case data from the LAC region and restrict our analysis to Plasmodium vivax (Pv), the dominant malaria species there. We study the smallest administrative units available over each area and use their centroid coordinates as their point locations.

For the years 2009-2016, we can compute monthly median case counts for 1 ADMIN1 (state) unit and 567 ADMIN2 (municipalities) units in Brazil, 458

ADMIN2 units in Colombia, 21 ADMIN1 units in Venezuela, 1 ADMIN1 units in Panama and the ADMIN0 (national) level for Ecuador, Suriname and Paraguay. The model chosen by backwards regression was refitted to all of the data since both training and test coverage probabilities (90.244% and 90.204%) are close to the target of 95% and the root mean squared errors (0.0211 and 0.0187) are comparably low.

Table 1 shows the parameter estimates for the chosen monthly proportion model. Under a 5% significance level, we observe a positive relation between monthly proportions and EVI but negative relations with CHIRPS, CHIRPS_lag1 and EVI_lag2. The negative relation with rainfall could be explained by the increase in mosquito breeding habits when river recede during the dry season (Valle & Lima 2014) Intense rain could also reduce treatment seeking rates and hence the number of recorded malaria cases.

Figure 1 shows the estimates of the seasonality index which were computed using 2016 Pv API estimates and 100 realisations of the estimated monthly proportions. The strongest seasonality based on the mean index is observed in Colombia, near its borders with Brazil and Peru. It was also found that most of the LAC region experiences only one transmission season per year. As an example, we present the map of the mean start month of the first season and its standard deviation in Figure 2. These are calculated via the circular definitions and transformed back into months via multiplication with $\frac{12}{2\pi}$ (Pewsey et al. 2013). The mean start months are reasonably contiguous across space as expected. Such maps of seasonality characteristics will be useful for policymakers for planning interventions.

Term	Median	95% CI	Term	Median	95% CI
Intercept	**-2.205**	(-2.276 , -2.133)	LST_delta_lag1	-0.017	(-0.116, 0.082)
CHIRPS_lag3	0.014	(-0.010, 0.038)	TCB_lag1	-0.026	(-0.085, 0.033)
CHIRPS	**-0.065**	(-0.093, -0.036)	EVI	**0.092**	(0.019, 0.165)
TCB_lag3	0.014	(-0.042, 0.071)	EVI_lag2	**-0.124**	(-0.203, -0.044)
TSI_Pv_lag3	-0.070	(-0.157, 0.017)	obs. var (σ_e^2)	**0.188**	(0.173, 0.203)
LST_night	0.035	(-0.080, 0.149)	spde.var.nom (σ_f^2)	**1.167**	(1.057, 1.288)
CHIRPS_lag1	**-0.041**	(-0.070, -0.012)	spde.range.nom (κ)	**5.350**	(4.954, 5.750)
LST_delta_lag3	0.094	(-0.005, 0.193)	AR.rho (a)	**0.908**	(0.892, 0.922)

Table 1: Posterior medians and 95% credible intervals (CIs) for the parameters of the refitted model.

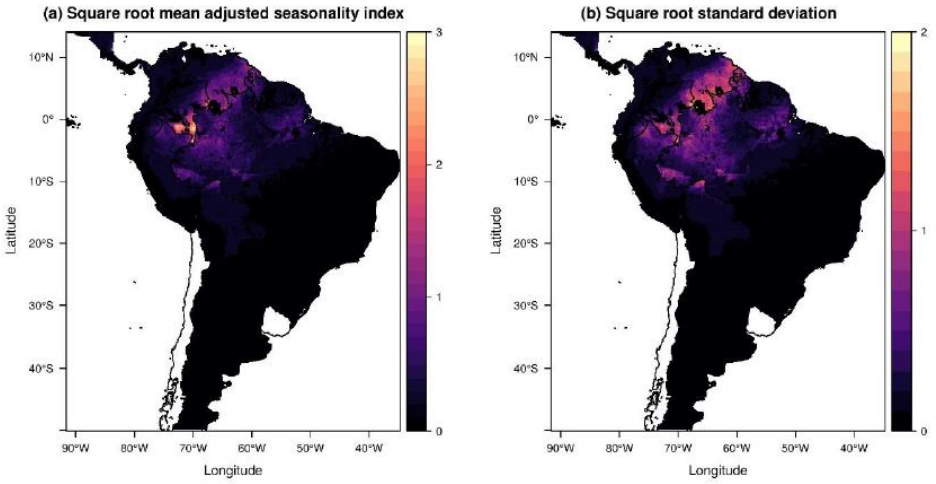


Figure 1: (a) The mean adjusted seasonality index estimates over the LAC region and (b) the standard deviation where square roots have been applied to highlight the spatial heterogeneity. The white regions denote areas with no Pv API data.

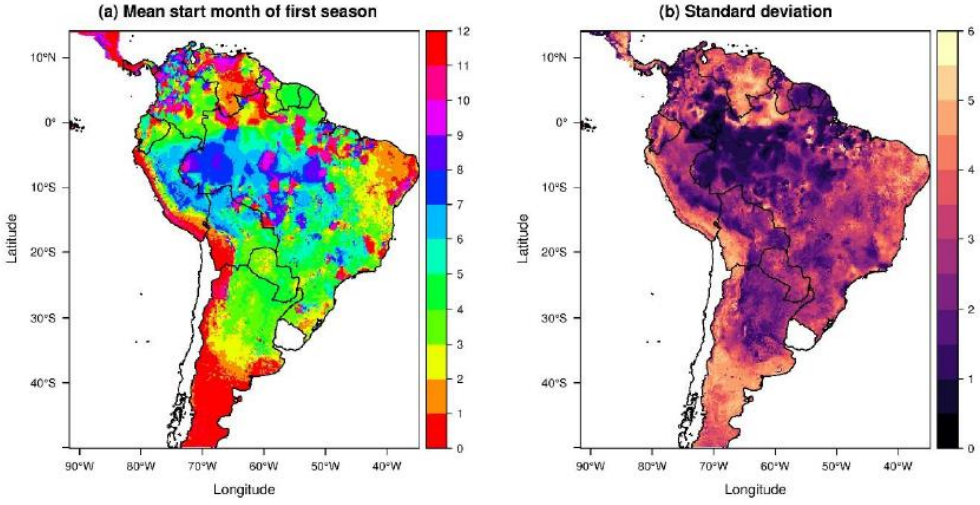


Figure 2: (a) Mean start month and (b) associated standard deviation for the first transmission season in the LAC region.

4. Discussion and Conclusion

Using a spatio-temporal statistical model, we related malaria seasonality to environmental drivers and derived useful information on transmission seasons. Although our focus has been on mapping these, our methodology has other use cases. For example, by multiplying the monthly proportion estimates with the API, we can compute the monthly case or parasite incidence (MPI) at each location without developing a separate model for MPI itself. In addition, the different stages of the methodology can be used separately: if

we have previously estimated monthly incidence, we can use the monthly proportion concept and algorithm to derive seasonal statistics.

Centroids of the smallest administrative units were used as point locations in the LAC data analysis. To avoid this approximation, continuous autoregressive (CAR) models can be used in place of the geospatial autoregressive process in our model. However, this has its own drawbacks (Valle and Lima 2014). The correlation between neighbouring units does not depend explicitly on the distance between them which is unnatural when we have units of varying sizes. Furthermore, the relation between malaria incidence and unit-representative covariates is likely to be weaker than at the point level.

The monthly proportion model identifies the dominant relationship between malaria cases and the environment in our study region. A key assumption is that this and the resultant seasonal pattern remain constant at least for the time period in our data. Since this may not be the case with climate change, there is a need to update models and investigate extensions to deal with varying relations.

As more countries adopt the District Health Information Software 2 (DHIS2) for instant recording of cases, using case data to establish seasonality patterns will be increasingly feasible and desirable. Currently, work is being done to apply this methodology to health facility case data from Madagascar.

References

1. Cairns, M., Roca-Feltrer, A., Garske, T., Wilson, A. L., Diallo, D. Milligan, P. J., Ghani, A. C. & Greenwood, B. M. (2012), 'Estimating the potential public impact of seasonal malaria chemoprevention in African children', *Nature communications* 3, 881.
2. Feng, X., Porporato, A. & Rodriguez-Iturbe, I. (2013), 'Changes in rainfall seasonality in the tropics', *Nature Climate Change* 3(9), 811-815.
3. Gemperli, A., Sogoba, N., Fondjo, E., Mabaso, M., Bagayoko, M., Briët, O. J. T., Anderegg, D., Liebe, J., Smith, T. & Vounatsou, P. (2006), 'Mapping malaria transmission in West and Central Africa', *Tropical Medicine & International Health* 11(7), 1032-1046.
4. Kang, S. Y., Battle, K. E., Gibson, H. S., Ratsimbasoa, A., Randrianarivelojosia, M., Ramboarina, S., Zimmerman, P. A., Weiss, D. J., Cameron, E., Gething, P. W. & Howes, R. E. (2018), 'Spatio-temporal mapping of Madagascar's Malaria Indicator Survey results to assess *Plasmodium falciparum* endemicity trends between 2011 and 2016', BMC Medicine 16(1), 71.
5. Lindgren, F., Rue, H. and Lindström, J. (2011), 'An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach', Journal of the *Royal Statistical Society: Series B (Statistical Methodology)* 73(4), 423-498.
6. Mabaso, M. L. H., Craig, M., Vounatsou, P. & Smith, T. (2005), 'Towards empirical description of malaria seasonality in southern Africa: the example of Zimbabwe', *Tropical Medicine & International Health* 10(9), 909-918.
7. Martinez, M. E. (2018), 'The calendar of epidemics: seasonal cycles of infectious diseases', PLoS Pathogens 14(11), e1007327.
8. Pewsey, A., Neuhäuser, M. & Ruxton, G.D. (2013), *Circular Statistics in R*, Oxford University Press.
9. Stuckey, E. M., Smith, T. & Chitnis, N. (2014), 'Seasonally dependent relationships between indicators of malaria transmission and disease provided by mathematical model simulations', *PLoS Computational Biology* 10(9), e1003812.
10. Valle, D. & Lima, J. M. T. (2014), 'Large-scale drivers of malaria and priority areas for prevention and control in the Brazilian Amazon region using a novel multi-pathogen geospatial model', *Malaria Journal* 13(1), 443.
11. World Health Organization (2018), *World Malaria Report* 2018, Geneva.

Multiclass classification of growth curves using Random Change Point Model with heterogeneity in the random effects

Vincent Chin *[11,2], Jarod Y. L. Lee[1,3], Louise M. Ryan[1,3,4,] Robert Kohn[1,5], Scott A. Sisson[1,2]

[1]Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers.
[2]School of Mathematics and Statistics, University of New South Wales, Sydney, Australia.
[3]School of Mathematical and Physical Sciences, University of Technology Sydney, Sydney, Australia.
[4]Harvard T.H. Chan School of Public Health, Harvard University, Cambridge, USA.
[5]School of Economics, University of New South Wales, Sydney, Australia.

Abstract

Faltering growth among young children is a nutritional problem prevalent in low to medium income countries and is in general defined as slower rate of growth compared to a reference healthy population of the same age and gender. As faltering is closely associated with reduced physical, intellectual and productive potentials, it is important to identify faltered children and be able to characterise different growth patterns so that target-specific treatments can be designed and administered. Our proposed multiclass classification model in this paper is built upon the broken stick model which is a piecewise linear model with breaks at the knots. Heterogeneity in the growth behaviour between children is captured by extending the broken stick model to mixture distributed random effects whereby the mixture components determines the classification of children into subgroups. We model the mixture distribution by a Dirichlet process prior. With this prior, we avoid having to choose the "true" number of components. Considering that children have different timings of growth stages, we propose replacing the fixed knots in the broken stick model by child-specific random change points. We illustrate our classification model on a longitudinal birth cohort from the Healthy Birth, Growth and Development knowledge integration (HBGDki) project funded by the Bill and Melinda Gates Foundation. Analysis on the dataset reveals 8 subgroups of children within the population. The largest subgroup consists of children with linear faltering trend while the others show varying degrees of growth catch-up at different stages between birth and age one.

Keywords

Bayesian non-parametric model; Child growth modelling; Dirichlet process prior; Longitudinal data; Mixture modelling

* Corresponding author: vincent.chin@student.unsw.edu.au

1. Introduction

According to the latest joint malnutrition estimates by United Nations Children's Fund, World Health Organization (WHO), and World Bank Group (2018), it is estimated that stunted growth is prevalent in 22.2% of the children population under the age of 5 in 2017 or over 150 million children worldwide. This is particularly serious in low to medium income countries where the rate of stunting is 35.0%. A major contributor to stunted growth is prolonged faltering, which comes with adverse consequences such as increased susceptibility to diarrhoea and respiratory infections (Kossmann et al., 2000), abnormal neurointegrative development (Benitez-Bribiesca et al., 1999) and capital loss to the labour market (Hoddinott et al., 2013). Therefore, it is imperative to take early preventive measures so that these impacts can be minimised. In order to implement the preventive measures, we need to first identify faltered children in the population of interest. In addition, it is also important to distinguish between the different types of growth patterns as each type represents particular growth behaviour. For example, children who caught up after faltering may have taken nutritional supplements. The strategy can then be replicated to other children in the cohort to improve their growth.

2. Methodology

A popular method for modelling longitudinal growth data in the epidemiology literature is the broken stick model defined as follows:

$$z_{ij} = \alpha_i + \beta_{0i}(t_{ij} - (t_{ij} - \xi_1)_+) + \beta_{Ki}(t_{ij} - \xi_K)_+ + \sum_{k=1}^{K-1} \beta_{ki}((t_{ij} - \xi_k)_+ - (t_{ij} - \xi_{k+1})_+) + \varepsilon_{ij}, \quad (2.1)$$

$$\alpha_i \sim N(\mu_\alpha, \sigma_\alpha^2), \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad (2.2)$$

for $i = 1, \ldots, N, j = 1, \ldots, Ji$ where $z_{ij} \in \mathbb{R}$ denotes the height-for-age z-score (HAZ) for child i on the j-th measurement occasion at age t_{ij}, x_+ gives the positive part of x and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)$ is an ordered vector of K predetermined internal knots or change points such that $\xi_1 < \cdots <, \xi_K$. The random intercept α_i and error ε_{ij} are both assumed to be normally distributed with parameter vectors given by $(\mu_\alpha, \sigma_\alpha^2)$ and such that $\xi_1 < \cdots <, \xi_K$. The random intercept α_i and error ε_{ij} are both assumed to be normally distributed with parameter vectors given by $(0, \sigma_\varepsilon^2)$ respectively. The child specific and time invariant α_i controls for the heterogeneity in the HAZ at birth around the population mean μ_α, and it is assumed to be uncorrelated with the error term ε_{ij}. The broken stick model fits $K + 1$ piecewise linear segments with breaks at $\boldsymbol{\xi}$ to model the growth trajectory calibrated in terms of the HAZ. The formulation in (2.1) enables an individual child's growth velocity to be

obtained directly from the regression coefficients because β_{ki} represents the rate of change in the HAZ between yearsξ_K and ξ_{k+1}.

So far, we have not mentioned any distributional assumption on the growth velocity vector $\beta_i = (\beta_{0i}, \ldots \beta_{Ki})^{\mathrm{T}}$. Anderson et al. (2018) and Lee et al. (2018) model β_i as realisations from a multivariate $N(\mu_\beta, \Sigma_\beta)$ distribution with mean vector μ_β and covariance matrix Σ_β. This signifies a homogeneous population model where individual growth profiles largely follow the trend of a global trajectory and the variability of deviation from this mean curve is determined by Σ_β. On average, the rate of growth is the same for all children in the population. However, this is rarely the case in practice. For example, Goode et al. (2014) find that higher socio-economic status has a positive impact on the HAZ through greater health consciousness and better household sanitation system. Therefore, we consider a normal mixture distribution where

$$\boldsymbol{\beta}_i \sim \sum_{g=1}^{G} \pi_g N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2.3}$$

for positive weights π_g summing to 1 in order to accommodate for a more complex composition in the population. Each mixture component in (2.3) corresponds to a particular type of growth pattern and each child belongs to one of these G subgroups. By clustering the children into different subgroups, further analysis can then be done to identify risk factors which cause the manifestation of certain growth behaviour.

Equation (2.3) requires the specification of the number of subgroups G, which is often not known *a priori* in practice. Therefore, we employ a Bayesian non-parametric approach to circumvent the model selection procedure in modelling the distribution of the growth velocity β_i. Conceptually, the number of parameters in a Bayesian non-parametric model is set to infinity and a prior distribution is posited on the infinite dimensional parameter space Θ. The complexity of the model (referring to the value of G in our setting) is then adapted to the amount of information available in the dataset. One such prior which has been widely used in various applications (da Silva, 2007; Blunsom et al., 2008) is the Dirichlet process (DP) prior established in Ferguson (1973).

In order to illustrate the usefulness of the DP prior, we formulate our problem of modelling the distribution of β_i by a mixture distribution in terms of a DP mixture model (Antoniak, 1974) in which

$$\boldsymbol{\beta}_i | (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) | H \sim H, \quad H \sim DP(\lambda, H_0), \tag{2.4}$$

for $i = 1, \ldots, N$ where $\phi_i = (\mu_i, \Sigma_i)$ is the parameter of a normal distribution specifying the mixture component associated with child i and $DP(\lambda, H_0)$

denotes a DP with concentration parameter $\lambda > 0$ and base distribution H_0. Here, one possible choice of H_0 is the normal-inverse-Wishart distribution.

To obtain meaningful results in any classification problems, we often require that $\phi_i = \phi_j$ for some $i \neq j$ so that each observation in the dataset does not belong to a cluster of its own, i.e. $G = N$. The DP prior exhibits such clustering property. Integrating out H from (2.4), Blackwell and MacQueen (1973) show that the conditional prior distribution induced on ϕ_i follows a Pólya urn scheme

$$\phi_i|\phi_1,\ldots,\phi_{i-1} \sim \frac{1}{\lambda+i-1}\sum_{j=1}^{i-1}\delta_{\phi_j} + \frac{\lambda}{\lambda+i-1}H_0, \tag{2.5}$$

where δ_ϕ is the point measure at ϕ. The parameter ϕ_i is generated by first drawing a sample from the base distribution H_0. Subsequent samples are then obtained by setting ϕ_i to be either a random draw from the current pool of parameters $\{\phi_1,\ldots,\phi_{i-1}\}$ with probability proportional to $i-1$ or a new sample from H_0 with probability proportional to λ. Since random draws from a continuous H_0 have zero probability of being identical, a large value of λ gives rise to a larger set of unique parameters $\{\theta_1,\ldots,\theta_G\}$ in $\{\phi_i,\ldots,\phi_N\}$. Teh (2011) show that for $N,\lambda \gg 0$,

$$\mathrm{E}[G] \simeq \lambda \log\left(1+\frac{N}{\lambda}\right),$$

indicating that the mean of G is data driven for a fixed concentration parameter λ and it scales logarithmically with the size of data N. Clustering effect of the DP as a result of the Pólya urn scheme in (2.5) makes it a popular option to model multimodal distributions without having to specify the number of components explicitly.

We have thus far treated the knot location vector $\boldsymbol{\xi}$ as predetermined and fixed across all children in the population. However, this is unrealistic in our context as different children react differently to treatment interventions such as the administration of vitamins or to negative experiences such as infections which will likely occur at different time points. The heterogeneity in the timing of treatment interventions or the occurrence of insults will likely cause individual trajectories to change course at different time points. Furthermore, fixing $\boldsymbol{\xi}$ results in a biased estimate of the growth velocity β_i in the broken stick model as regression lines between two neighbouring segments are connected at the internal knot. This would then affect the classification model because we summarise the growth pattern of children by β_i. Therefore, a sensible approach is to model the knot location within the interval of $[0,T]$ as child specific random effects $\boldsymbol{\xi}_i = (\xi_{ik},\ldots,\xi_{iK})$ whose distribution is expressed by

$$p(\boldsymbol{\xi}_i) \propto \prod_{k=1}^{K+1} (\xi_{ik} - \xi_{i,k-1}) \times \prod_{k=1}^{K} \mathbb{1}\left(\xi_{ik} \in \left(\frac{(k-1)T}{K}, \frac{kT}{K}\right)\right), \tag{2.6}$$

for $i = 1, \ldots, N$ where $\xi_{i0} = 0$ and $\xi_{i,K+1} = T$ for convenience and $\mathbb{1}(E)$ is an indicator function which takes value 1 if the event E occurs and 0 otherwise. The first term in (2.6) is the even-numbered order statistics from $2K + 1$ points uniformly distributed on $[0, T]$ used in Green (1995) in order to ensure that adequate spacing between internal knots is achieved probabilistically. Although it penalises short subintervals, it might still be possible that $\boldsymbol{\xi}_i$ are concentrated on regions where there is an abundance of data. We thus impose a hard constraint via the second term in (2.6) so that there is an internal knot within each subinterval of equal length on $[0, T]$.

We update $\boldsymbol{\xi}_i$ one component at a time using the Metropolis-Hastings algorithm with independent proposal distribution within the Markov chain Monte Carlo (MCMC) sampling scheme. The acceptance probability is

$$\min\left\{1, \exp(\ell_i(\tilde{\boldsymbol{\xi}}_i^{(k)}) - \ell_i(\boldsymbol{\xi}_i)) \times \frac{(\xi_{i,k+1} - \tilde{\xi}_{ik})(\tilde{\xi}_{ik} - \xi_{i,k-1})}{(\xi_{i,k+1} - \xi_{ik})(\xi_{ik} - \xi_{i,k-1})}\right\},$$

where $\tilde{\boldsymbol{\xi}}_i^{(k)} = (\xi_{i1}, \ldots, \xi_{i,k-1}, \tilde{\xi}_{ik}, \xi_{i,k+1}, , \ldots, \xi_{i,K})$ is a proposal vector of knot location with $\tilde{\xi}_{ik}$ uniformly sampled from the subinterval $((k-1)T/K, kT/K)$ and ℓ_i is the log-likelihood for child i.

3. Result

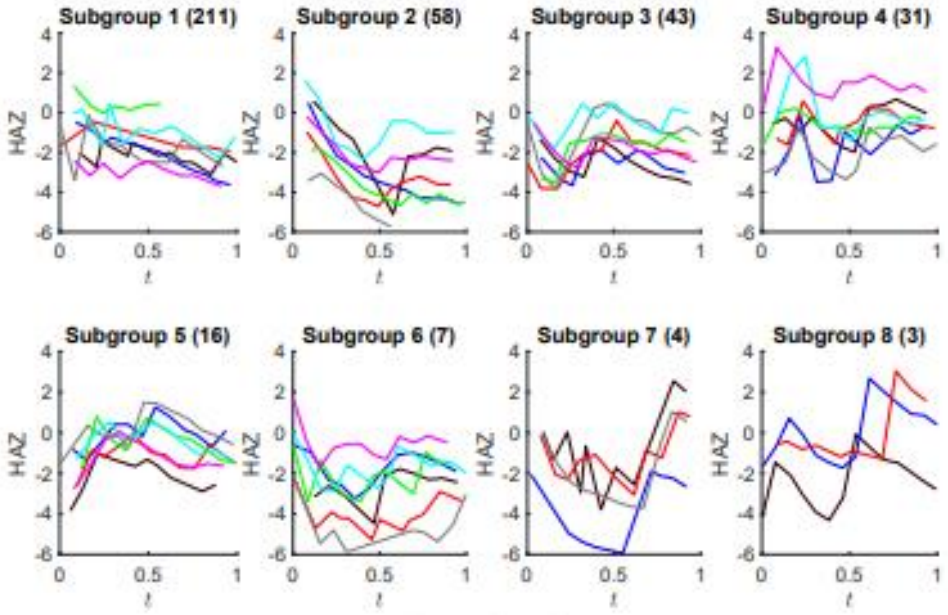
Our application example of classifying growth curves is based on a longitudinal study from the HBGDki project which analyses the prevalence of rotavirus infections in a birth cohort in Vellore, India (Paul et al., 2014). The sample population of 373 children are followed up for three years since birth and have their anthropometric measurements recorded. For the purpose of our analysis, we only focus on the HAZ up to one year old after removing outliers (HAZ < -6 or HAZ > 6) based on WHO recommendations. There are 5 to 15 observations for each child and the first measurement is taken between day 1 and 225. We convert the time scale to age in years and set the number of random change points $K = 3$. More sophisticated models can be formulated by allowing K to vary, for example by using the reversible jump algorithm introduced in Green (1995). However, we fix the value of K here as the number of measurements taken for each child is relatively small.

Figure 3.1a shows a random sample of raw trajectories in each subgroup obtained from the classification model, while their respective posterior mean curves are given in Figure 3.1b. Eight different subgroups of children are identified in the dataset. The largest subgroup which accounts for more than half of the child population shows a constant faltering pattern throughout the first year of the observational period. Subgroups 2 and 6 experience severe

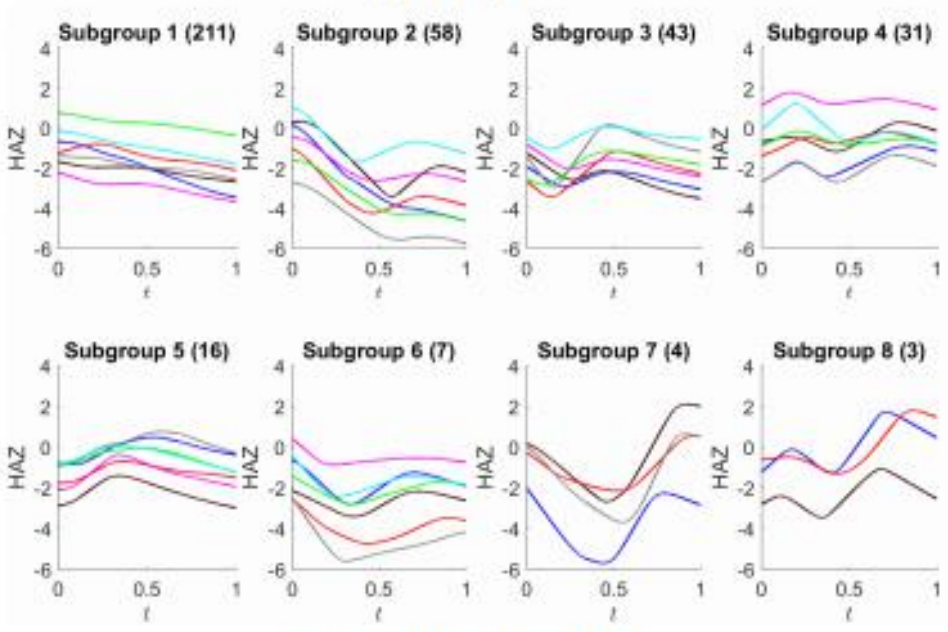
faltering and then the HAZ remains relatively stable after that. Generally, subgroup 2 has a longer duration of faltered growth. Children in subgroups 3 and 7 have a steep decline in the HAZ score before a short interval of recovery is observed, and then followed by another onset of faltered growth. The differences between these two subgroups lie in the time point at which growth catch up takes place and the rate of improvement. The growth catch up phase for children in subgroup 3 is milder and happens between $t = 0.25$ and $t = 0.50$ whereas huge jumps in the HAZ happen between $t = 0.50$ and $t = 0.80$ in subgroup 7. Another pair of clusters which are similar is subgroups 4 and 8 whereby the growth looks like a sinusoidal wave, but with different amplitudes. The HAZ for children in subgroup 5 reaches a peak and then starts to decline.

4. Conclusion

In this article, we use the broken stick model as the basis to propose an approach which incorporates a classifier within a regression model. This allows the classification of growth curves into different patterns based on the vectors of regression parameters to be achieved within a single model. In order to capture the heterogeneity in the growth velocity between children, we extend the broken stick model to allow for mixture distributed random slopes. The classification of an individual child's growth profile is then determined by the component of mixture distribution from which the vector of velocities derived from the regression model is generated. We model the distribution of growth velocity non-parametrically in a Bayesian framework using a DP prior. The DP prior adapts complexity of the model to the amount of data available without having to choose the number of mixture components, which is often unknown in practical applications. Another contribution of this paper is to introduce the idea of random change points into the broken stick model in order for the knots to adjust their locations instead of having them fixed. These change points are modelled as random effects so that the difference in the timing of growth phase between children is taken into consideration in the model.



(a) Raw trajectories.



(b) Posterior mean trajectories.

Figure 3.1: The raw trajectories and their respective posterior mean curves for a random sample of children from each subgroup. The number of children in the subgroup is given in bracket.

References

1. Anderson, C., R. Hafen, O. Sofrygin, and L. Ryan (2018). Comparing predictive abilities of longitudinal child growth models. *Statistics in Medicine.* To appear.
2. Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* 2(6), 1152–1174.
3. Ben´ıtez-Bribiesca, L., I. De la Rosa-Alvarez, and A. Mansilla-Olivares (1999). Dendritic spine pathology in infants with severe protein-calorie malnutrition. *Pediatrics* 104(2), 1–6.
4. Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via P´olya urn schemes. *The Annals of Statistics* 1(2), 353–355.
5. Blunsom, P., T. Cohn, and M. Osborne (2008). Bayesian synchronous grammar induction. *In Advances in Neural Information Processing Systems,* Volume 21, pp. 161–168.
6. da Silva, A. R. F. (2007). A Dirichlet process mixture model for brain MRI tissue classification. *Medical Image Analysis* 11(2), 169–182.
7. Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1(2), 209–230.
8. Goode, A., K. Mavromaras, and R. Zhu (2014). Family income and child health in China. *China Economic Review* 29, 152–165.
9. Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82(4), 711–732.
10. Hoddinott, J., H. Alderman, J. R. Behrman, L. Haddad, and S. Horton (2013). The economic rationale for investing in stunting reduction. *Maternal & Child Nutrition* 9(2), 69–82.
11. Kossmann, J., P. Nestel, M. Herrera, A. Amin, and W. Fawzi (2000). Undernutrition in relation to childhood infections: A prospective study in the Sudan. *European Journal of Clinical Nutrition* 54(6), 463–472.
12. Lee, J. Y. L., C. Anderson, W. T. Hung, H. Hwang, and L. M. Ryan (2018). Detecting faltering growth in children via minimum random slopes. *arXiv preprint arXiv:1812.05903*.
13. Paul, A., B. P. Gladstone, I. Mukhopadhya, and G. Kang (2014). Rotavirus infections in a community based cohort in Vellore, India. *Vaccine* 32(11), A49–A54.
14. Teh, Y. W. (2011). Dirichlet Process. In C. Sammut and G. I. Webb (Eds.), Encyclopedia of Machine Learning, pp. 280–287. Springer.
15. United Nations Children's Fund, World Health Organization (WHO), and World Bank Group (2018). Levels and trends in child malnutrition: Key findings of the 2018 edition of the joint child malnutrition estimates.

Master's programme in Data Analytics for Government: The UK experience

Solange Correa Onel, David Johnson
UK Office for National Statistics

Abstract

This paper presents the phases of the process of planning, launching and implementing an innovative Master's programme in Data Analytics for Government by the UK Office for National Statistics (ONS). This Master's programme is a flagship capacity building project created and managed by ONS in collaboration with multiple UK universities to support the "Better Statistics, Better Decisions" strategy for UK statistics. Successes and challenges in implementing the Master's programme are outlined and priorities for the next phase of the process are described.

Keywords

Capacity building; Data Science; academic partnership; training; skilled workforce

1. Introduction

The UK Office for National Statistics (ONS) is undergoing a data transformational journey to ensure that government can make the best possible use of traditional and alternative data sources currently available to improve people's lives. The 'Better Statistics, Better Decisions' strategy (Government Statistical Service, 2015) for UK statistics highlights the critical importance of readily accessible data, speedy data sharing across government departments, up-to-date data architecture and data processing mechanisms and a workforce with cutting-edge analytic skills to extract value from data. Growing data analytics skills across the UK government plays, therefore, a key role in this transformation.

This paper describes the implementation process of one of the flagship capacity building initiatives of the ONS: the UK Master's programme in Data Analytics for Government (MDataGov). Section 2 introduces the MDataGov Framework, its scope, structure, timeline and implementation plan. Section 3 presents the impact of the programme since its launch in 2017 and how the growth of a data analytics community in government is supported by former, current and prospective MDataGov students and academic partners. It also shows the MDataGov expansion plan currently in progress, aimed at broadening the geographic coverage of the Master's programme and increasing support to civil servants in other UK regions and abroad.

2. The UK Master's in Data Analytics for Government Framework

Overview

The Master's in Data Analytics for Government is a collaborative project between the UK Office for National Statistics and UK academic partners. The programme was launched in October 2017 and aims to build data science capability across government by equipping civil servants with a key set of skills required from a modern government data analyst.

Open to analysts from across the UK public sector, the MDataGov is a unique and flexible modular part-time Master's programme designed with and for government, combining skills in Statistics and Data Science that are essential for UK government data transformation.

As well analysts from ONS, students undertaking the full MDataGov programme are drawn from over 10 government departments, including key central departments like Her Majesty Revenue & Customs, Department of Health and Social Care, Home Office, Ministry of Justice, Department for Environment, Food and Rural Affairs, Department for Business Energy and Industrial Strategy and the Northern Ireland Statistics and Research Agency. This number rises to over 20 departments and agencies when CPD participants are factored in, including international partners like the National Institute of Statistics of Rwanda and Statistics Indonesia.

Programme Structure

In its standard format, the MDataGov programme consists of four core modules (Data Science Foundations, Statistics in Government, Survey Fundamentals and Statistical Programming), eight optional modules to be chosen from a range of courses in Statistics and Data Science (e.g. Time Series Analysis, Introduction to Machine Learning, Advanced Machine Learning, Introduction to Distributed Systems, Data Visualisation, Introduction to Survey Research, Advanced Statistical Modelling, Spatial Analysis, among others) and a Master's dissertation in data analytics. To be awarded the Master's degree students are required to complete 120 credits in taught components and a dissertation, yielding to 180 credits in total. The credit allocation per module in Master's programme varies across UK universities. Typically, a module is worth 10 credits and the dissertation module is worth 60 credits.

The MDataGov programme offers three possible exit points: Postgraduate Certificate (60 credits), Postgraduate Diploma (120 credits) and Master's degree (120 credits plus the dissertation) in Data Analytics for Government. The four core modules are required in any of these exit points.

The MDataGov is currently provided in a classroom-based format by three UK universities: University of Southampton, Oxford Brookes University and University College London. The minimum entry requirement is an upper second-class Bachelor's degree in a quantitative discipline from a UK

university. Modules combine theory and computer lab-based sessions and are assessed by written exams and/or coursework and oral presentations.

Modules are offered in a week-long format or throughout the standard academic terms and semesters. In the latter, lectures can be concentrated on 1 or 2 days a week, depending on the university, for ease of travel and minimum impact on civil servants' work.

Tables 1 and 2 show timetables for the 2018-2019 academic from Oxford Brookes University (Oxford Brookes University, 2017) and University of Southampton (University of Southampton, 2017), respectively, illustrating the delivery format adopted by each university. Table 3 shows the programme structure at University College London (University College London, 2017), which currently offers the traditional format of lectures distributed throughout the standard academic terms.

Table 1. Timetable for modules delivered at Oxford Brookes University. UK Master's in Data Analytics for Government. 2018-2019 academic year. Each module is worth 10 credits

Semester 1	**Monday**	**Friday**
10:00 – 12:00	Regression modelling	Intro to Machine Learning
13:00 – 15:00	Statistical Programming	Intro to Distributed Systems
16:00 – 18:00	Data Science Foundations	Intro to Survey Research
Semester 2	**Monday**	**Friday**
10:00 – 12:00	Advanced Statistical Modelling	Time Series Analysis
13:00 – 15:00	Statistics in Government	Data Visualisation
16:00 – 18:00	Survey Fundamentals	Advanced Machine Learning

Table 2. Timetable for modules delivery at University of Southampton. UK Master's in Data Analytics for Government. 2018-2019 academic year. Each module is worth 10 credits

Semester 1 Dates	**Module Title**
1st – 5th October 2018	Introduction to Survey Research
15th – 19th October 2018	Regression Modelling
29th October – 2nd November 2018	Statical Programming
12th – 16th November 2018	Demographic Methods
19th – 23rd November 2018	Data Science Foundations
26th – 30th November 2018	Survey Fundamentals

Semester 2 Dates	Module Title
28th January – 01st February 2019	Ecaluation and Monitoring
11th – 15th February 2019	Further Survey Estimation Methods
25th February – 01st March 2019	Time Series Analysis
18th – 22nd March 2019	Survey Data Collection
01st – 05th April 2019	Compensating for Non-response
29th April – 03rd May	Statistics in Government

Table 3. Programme structure at University College London. UK Master's in Data Analytics for Government. 2018-2019 academic year

Module Title	Credits	Academic Term
Core modules		
Statistics in Government	15	3
Data Science Foundations	15	1
Survey Fundamentals	15	1
Statistical Programming	15	1
Data Analytics for Government Dissertation	60	

Optional Modules (students choose 60 credits from the below)

Course Title	Credits	Term
Introduction to Survey Research	15	
Regression Modelling	15	1
Digital Visualisation	30	2
Survey Data Collection	15	2
Further Survey Estimation Methods	15	1
Advanced Statistical Modelling	15	2
Time Series Analysis	15	2
Spatial Analysis	15	1

MDataGov modules can be taken as part of the full MSc programme or as standalone (assessed or unassessed) module for Continuing Professional Development (CPD). Academic credits gained through assessed CPDs are counted towards the Master's degree, if students decide to join the full Master's programme at a later stage. The programme duration varies from 1 to 5 years, depending on the student.

Week-long MDataGov modules are also offered onsite at ONS offices. Onsite module delivery has proved to be extremely popular across the UK public sector, attracting over 80 students per academic year. As a result, ONS

has been offering an onsite annual CPD programme in collaboration with Oxford Brookes University. Table 4 shows ONS CPD programme scheduled in the 2018-2019 academic year.

Table 4. Timetable for modules delivered at ONS. UK Master's in Data Analytics for Government. 2018-2019 academic year. ONS and Oxford Brookes University

Module	**Date**
Introduction to Machine Learning	21st – 25th January 2019
	24th – 28th June 2019
Data Visualisation	21st – 25th January 2019
Survey Fundamentals	20th – 24th May 2019
Introduction to Distributed Systems	20th – 24th May 2019
Statistics in Government	10th – 14th June 2019

The MDataGov Framework Implementation

Implementation of the MDataGov Framework started 13 months before its formal launch in October 2017. From September to November 2016, ONS organised market events and consultations with a range of government departments and representatives of UK analytical professions to agree the core set of analytic skills required across government. In December 2016, ONS released the MDataGov Framework and opened a public Tender Process inviting universities to submit their proposals for appreciation. The Framework included guidelines on programme structure, module titles and contents and delivery mode. One of the key requirements of the Framework was the offering by each university of at least four optional modules in Statistics and at least four optional modules in data science related topics (e.g. machine learning, data visualisation, etc.).

From March to September 2017, ONS worked closely with the three successful universities, organising quarterly review meetings to ensure a smooth implementation of the programme. At this stage, ONS provided the three universities with complete material for four government-specific modules (Statistics in Government, Introduction to Survey Research, Survey Data Collection and National Accounts) free of charge.

In October 2017, the first intake of over 80 civil servants started the programme as either full Master's or CPD students. In October 2018, this number rose to over 100 civil servants. For some university partners, provision and delivery of the full programme of agreed modules took longer than anticipated. An extended deadline of August 2019 was agreed with affected partners, to enable them to adapt their existing curricula to meet government

requirements with support from ONS. This was Phase 1 of the Framework implementation process.

Phase 2 started in February 2018, when ONS identified the need to expand the list of MDataGov providers to broaden the MDataGov geographic coverage and incorporate a distance learning component into the Master's programme from the 2019-20 academic year. Provisions that started in Phase 1 are also being reviewed in Phase 2. ONS has adopted a more flexible approach in Phase 2 by targeting key academic partners in other UK regions to join the Framework. ONS is currently progressing discussions with other three UK universities, and Phase 2 is planned to be completed by October 2020.

3. Successes, Challenges and Next Steps:

The launch of the MDataGov programme has facilitated the provision of advanced data analytics skills across the UK public sector, attracting a considerable number of students every year and enabling them to bring new insights to analytical and policy challenges faced by the UK government. As a result, an increasing number of network initiatives has been organised by this community and supported by ONS, including an Annual MDataGov Symposium organised by the students on the programme, a dedicated online collaboration platform for students, and sponsorship programme for public sector analysts undertaking the MDataGov programme. ONS has also welcomed analysts from other National Statistical Institutes to undertake week-long modules from the Master's programme and exchange experiences regarding the implementation and launch of the MDataGov Framework.

The co-ordination and management of a smooth delivery of a Master's programme across multiple providers can be challenging. For instance, ONS needs to ensure attractive value-for-money for the civil service, relevance of modules for government, a balance between demand and providers, regular marketing exercise, appropriate in-house training infrastructure, and a flexible timeline to accommodate universities' extremely busy periods, among other factors.

By August 2020, the end of the current Phase 2 of the Framework, ONS and partners expect to offer an expanded MDataGov geographic coverage, a distance learning component with at least one academic partner and to have a third cohort of over 100 civil servants attending the programme.

The MDataGov programme thus plays a central role in enabling the delivery of the ambitious 'Better Statistics, Better Decisions' strategy, making an impact on the way UK government extracts value from data and enabling it to make better decisions for the benefit of the public good.

References

1. Government Statistical Service (2015) Better Statistics, Better Decisions: Strategy for UK Statistics, 2015-2020 https://gss.civilservice.gov.uk/policy-store/better-statistics-better-decisions-2/ Accessed 28 January 2019
2. Office for National Statistics (2017) Master's in Data Analytics for Government https://datasciencecampus.ons.gov.uk/capability/msc-in-data-analytics-for-government/ Accessed 28 January 2019
3. Oxford Brookes University (2017) Mater's in Data Analytics for Government https://www.brookes.ac.uk/courses/postgraduate/data-analytics-for-government/ Accessed 28 January 2019
4. University College London (2017) Master's in Data Analytics for Government https://www.ucl.ac.uk/big-data/Training_Education/MSc_Data_Analytics/ Accessed 28 January 2019
5. University of Southampton (2017) Master's in Data Analytics for Government https://www.southampton.ac.uk/demography/postgraduate/taught_courses/msc-data-analytics-for-government.page#modules Accessed 28 January 2019

Detecting life expectancy anomalies in England using a Bayesian hierarchical model

Areti Boulieri, Marta Blangiardo
MRC-PHE Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, UK

Abstract

In England, life expectancy has shown a steady increase over many years, however these improvements have recently started to slow down considerably. This work aims to investigate the changes in life expectancy in England over time and across its local authorities, and to identify local authorities with unusual time trends that might help with hypothesis generation and point to emerging risk factors. We analyse mortality count data in England for females at the local authority level (324 areas), from 2001 to 2016 (17 years), and by age group, assuming 19 age groups of 5 year bands. We develop a statistical model within the Bayesian hierarchical framework that accounts for spatial, temporal, and age effects, as well as for pairwise interactions. The space-time interaction parameter is used to detect areas whose time trends deviate from the national one. The detection rule that we specify focuses on areas that are detected as unusual over the last 5 years of the time period (2013 – 2017). The model is implemented in Integrated Nested Laplace Approximations (INLA). We found roughly 40 areas to be highlighted as unusual, following a different time trend in the mortality rates compared to the national trend.

Keywords

Bayesian statistics; spatio-temporal modelling; anomaly detection; life expectancy surveillance

1. Introduction

The study of life expectancy is of primary interest in public health practice, where it is needed to plan for health and social services. Recently, it has attracted a lot of attention (Hill 2018; Therrien 2018), due to the stalling effect that has been observed in several countries, including USA and UK (Olshansky et al. 2005; Hiam et al. 2018). That is, while life expectancy has been improving steadily since the early 80s when records began, mainly due to better lifestyles and healthcare, this phenomenon has started to slow down since 2012 onwards. According to Office for National Statistcs (ONS), between 2014 and 2015, life expectancy fell by 0.2 years for both sexes. A larger decrease was noticed for females, and for larger age groups. In addition, there have been fluctuations in life expectancy among local authorities in England (Office for

National Statistics, 2016). Even though life expectancy is known to vary substantially across space, very few studies have incorporated both temporal and spatial information to investigate how the spatial patterns of mortality evolve in time, and therefore to better understand the alarming behaviour recently seen in England. In addition, most studies have traditionally used standard statistical techniques, e.g. age-standardised mortality rates, thus not accounting for the noise in the data. It is therefore hard to determine whether spikes in deaths are true or data artefacts, in particular, when studying low populated areas. In this paper, we analysed mortality counts in England at the local authority level from 2001 to 2016 using Bayesian statistical methods, which borrow strength from spatial and temporal neighbours to reduce the high variability inherent to classical risk estimators, such as the crude mortality rate. The main objective of this work was to investigate whether life expectancy time trends are stable across England and highlight areas whose trends differ to the national one over the last 5 years (2012 to 2016).

2. Methodology

We used mortality counts from 2001 to 2017 at the local authority level in England. In total 324 local authorities were considered, after excluding the Isles of Scilly and City of London. Information on age and sex was available for each record. We used 19 5-year band age groups (0-4, 5-9, …, 90 plus) and we analysed males and females separately as these are expected to have very different mortality levels and trends. In this paper we present results for females only. Population data by age and sex for each local authority for the same time period were also used for the analysis. All data were provided by Public Health England (PHE), originally held by Office for National Statistics (ONS). Life expectancy tables were used to convert mortality rates to life expectancy rates.

We developed a statistical model to analyse mortality counts by age group, local authority, and year. The model was formulated within a Bayesian hierarchical framework, which allowed to assign prior specification to the unknown parameters through which we incorporate assumptions regarding the structure of the data. The model is as follows:

$$Y_{i,j,t} \sim Poisson(\lambda_{i,j,t}\, Pop_{i,j,t})$$

where $Y_{i,j,t}$ and $Pop_{i,j,t}$ are the mortality counts and population counts respectively in area $i = 1,\ldots,324$, age group $j = 1,\ldots,19$ and year $t = 2001,\ldots,2017$. Similarly, the parameter $\lambda_{i,j,t}$ represents the mortality rate which we model on the log scale as $\lambda_{i,j,t} = \alpha + \eta_i + \delta_j + \gamma_t + \nu_{j,t} + \xi_{i,j} + k_{i,t}$. The overall intercept α follows a flat prior. The spatial component η_i is assigned a convolution prior, widely known as the Besag York Mollie (BYM) model (Besag et al. 1991). This is a Gaussian prior, $\eta_i \sim N(u_i, \sigma_\eta^2)$ where u_i is

a spatially structured term following an intrinsic conditional autoregressive prior $u_i \sim ICAR(\boldsymbol{W}, \sigma_u^2)$ where $\boldsymbol{W}$ is the adjacency matrix specifying the spatial neighbourhood. The temporal component γ_t and the age component δ_j both follow a Gaussian random walk prior of order 1 (RW1) which is implemented as an ICAR prior. The interacton terms $\xi_{i,j}$ and $k_{i,t}$ follow normal iid priors, while the interaction term $\nu_{j,t}$ follows an age specific random walk prior of order 1 (RW1). All model hyperparameters follow a weakly informative half Normal prior. The analysis was carried out using integrated nested Laplace approximations (INLA), which is an alternative to the traditional Markov chain Monte carlo (McMC) methods for Bayesian inference specifically designed for latent Gaussian models (Rue and Held 2005; Blangiardo et al. 2013). In the context of spatio-temporal modelling, several Bayesian detection rules exist, and a trade off between true and false positives is usually desired. A common practice is to use Bayesian exceedance probabilities (Richardson et al. 2004; Lawson 2013). These base the detection rule on a cut off value on the posterior probability that the space-time interaction is above a reference threshold. The choice of the reference threshold value and the cut off value depends on the data, i.e. number of areas, time points and magnitude of expected cases (Ugarte et al. 2009). In our work, we were interested in detecting unusual behaviour both in terms of increasing or decreasing trend compared to the national one and we chose the reference threshold to be 1, and cut off values 0.05 and 0.95. Unusual areas were detected in terms of space-time interactions over the last 5 years of the time period, i.e. 2013-2017, ($Prob(\exp(k_{i,t}>1))<0.05$ or $Prob(\exp(k_{i,t}>1))>0.95$).

National mortality rates were calculated as the population-weighted average of age-specific mortality rates. Finally, mortality rates were converted to life expectancy using life table methods (Preston, Heuveline and Guillot 2001)

3. Result

Under the detection rule that we specified, we found 21 areas to be unusual over the last 5 years of the study period with an exceedance probability below 0.05, suggesting a decreased slope compared to the national one, and 32 areas with an exceedance probability above 0.95, suggesting an increased slope respectively (Figure 1).

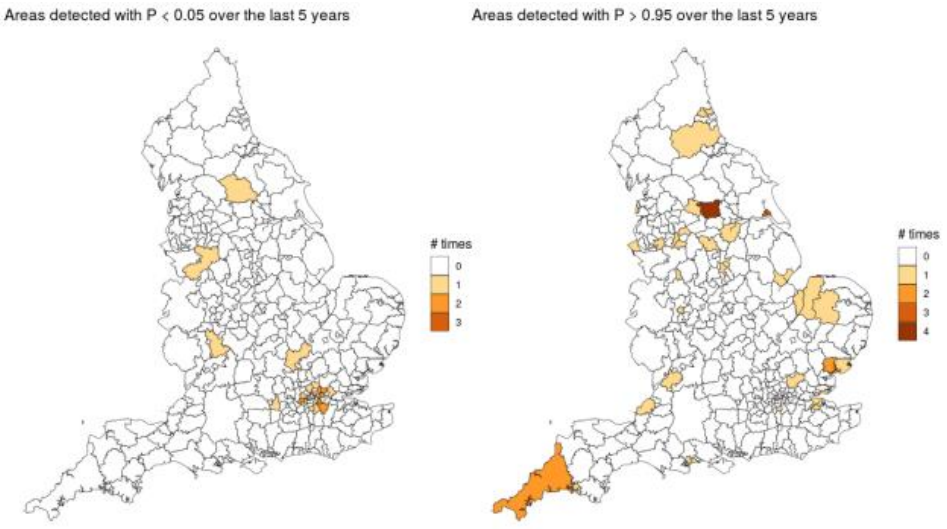


Figure 1: Local authorities in England that are detected as unusal

Figures 2 and 3 show the time trends of mortality rate and relative risk for two of the detected areas, Southwark (prob < 0.05) and Leeds (prob > 0.95), plotted against the national time trends. We can see that the life expectancy trend for the local authority of Southwark differs in year 2017 to the national trend, showing a lower rate than expected, while the opposite is observed for the local authority of Leeds in years 2013, 2015, 2016 and 2017.

4. Discussion and Conclusion

In this paper we analysed mortality data on females to estimate life expectancy trends in England from 2001 to 2017. We were interested in detecting areas whose time trends deviate from the overall national trend, suggesting unusual behaviour. For the analysis we used a Bayesian hierarchical model, allowing for spatial, temporal, and age effects, and we based our detection rule on exceedance probabilities of the space-time interaction. We found 21 areas to be unusual with a decreased slope compared to the national one, and 32 with an increased slope respectively. These findings call for further investigation as to the reasons behind these patterns. Potential extension of this work is to allow for deprivation, the major determinant of life expectancy (Woods et al. 2005; Smith, Olatunde, and White 2010), in order to assess potential associations with the unusual time trends.

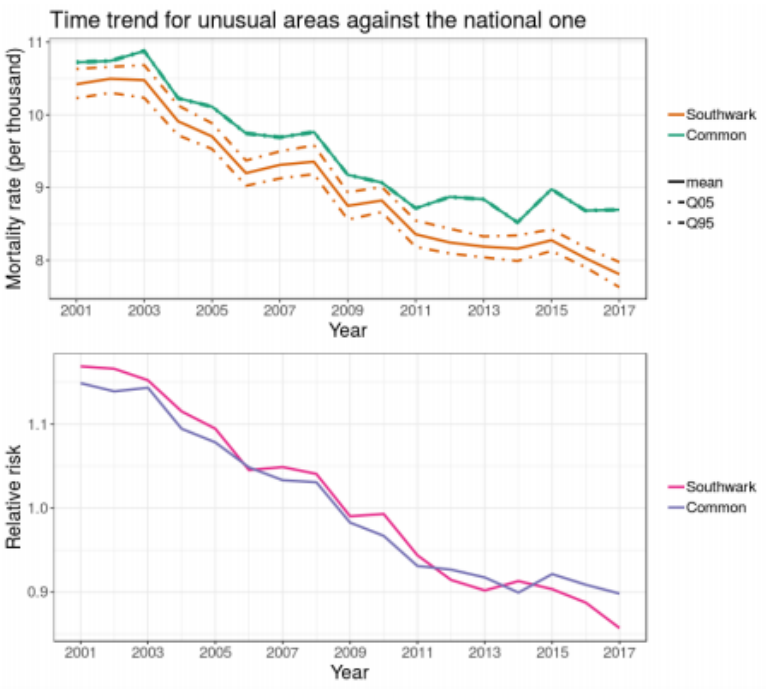


Figure 2: National time trend of mortality rate and relative risk plotted against the corresponding time trends for 'Southwark', detected as unusual with prob < 0.05.

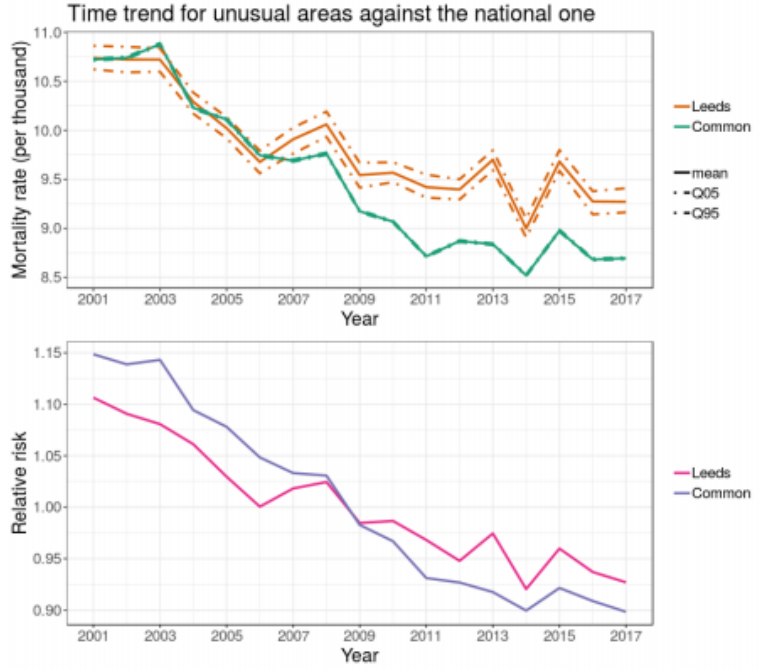


Figure 3: National time trend of mortality rate and relative risk plotted against the corresponding time trends for 'Leeds', detected as unusual with prob > 0.95.

References

1. Figure 2: National time trend of mortality rate plotted against the corresponding time trend for 'Southwark', detected as unusual with prob < 0.05.1. Hill, Amelia. (2018). UK *Life Expectancy Improvement Has Stalled, Figures Show*. Great Britain: The Guardian. Retrieved from https://www.theguardian.com/science/2018/sep/25/improvement-uk-life-expectancy-stalled-onsfigures-show
2. Therrien, Alex. (2018). UK *Life Expectancy Progress 'Has Stopped'*. Great Britain: BBC News. Retrieved from https://www.bbc.co.uk/news/health-45638646.
3. Olshansky, S. J., Passaro, D. J., Hershow, R. C., Layden, J., Carnes, B. A., Brody, J., ... & Ludwig, D. S. (2005). A potential decline in life expectancy in the United States in the 21st century. *New England Journal of Medicine*, 352(11), 1138-1145.
4. Hiam, L., Harrison, D., McKee, M., & Dorling, D. (2018). Why is life expectancy in England and Wales 'stalling'?. J Epidemiol Community Health, 72(5), 404-408.
5. Office for National Statistics. (2016). *Provisional Analysis of death registrations*: 2015. (p. 11). Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/pro visionalanalysisofdeathregistrations/2015
6. Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1), 1-20.
7. Rue, H., & Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.
8. Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology,* 4, 33-49.
9. Richardson, S., Thomson, A., Best, N., & Elliott, P. (2004). Interpreting posterior relative risk estimates in disease-mapping studies*. Environmental Health Perspectives*, 112(9), 1016.
10. Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC
11. Ugarte, M. D., Goicoa, T., Ibanez, B., & Militino, A. F. (2009). Evaluating the performance of spatio temporal Bayesian models in disease mapping. - *Environmetrics: The official journal of the International Environmetrics Society*, 20(6), 647-665.
12. Preston, S., Heuveline, P., & Guillot, M. (2000). *Demography: measuring and modeling population processes*. Blackwell Publishing, Oxford.
13. Woods, L. M., Rachet, B., Riga, M., Stone, N., Shah, A., & Coleman, M. P. (2005). Geographical variation in life expectancy at birth in England and

Wales is largely explained by deprivation. *Journal of Epidemiology & Community Health*, 59(2), 115-120.

14. Smith, M. P., Olatunde, O., & White, C. (2010). Inequalities in disability-free life expectancy by area deprivation: England, 2001–04 and 2005–08. *Health Statistics Quarterly*, 48(1), 36-57.

The modified Lee-Carter model with linearized cubic spline parameter approximation for Malaysian mortality data

Chin Tsung Rern, Dharini Pathmanathan, Shafiqah Azman, Nurul Aityqah Yaacob

Institute of Mathematical Sciences, Faculty of Science, University of Malaya, Malaysia

Abstract

We examine the application of the Lee-Carter (LC) model with parameters approximated using linearized cubic splines. The original LC model produces severely fluctuating predicted age-specific mortality. The model is applied to the Malaysian mortality data (1991 to 2015) to select the model which suits best to represent Malaysian mortality which was obtained from the Department of Statistics Malaysia. The base period of this data is only 25 years compared to other data available in the Human Mortality Database. The forecasts obtained based on the LC model are unstable when the base period is short. Hence, a modification of the LC model with linearized cubic spline parameter approximation for short base periods which was initially applied to the Chinese mortality data was applied to the Malaysian mortality data. Several models were studied based on the selection of knots for male and female mortality data and the best models to represent male and female mortalities in Malaysia were selected. The modified models attained smaller estimation errors (with respect to mean absolute percentage error, MAPE and mean squared error, MSE) compared to the LC model.

Keywords

cubic spline; mortality; Lee-Carter; forecast

1. Introduction:

In recent years, it has been observed that there is a decline in mortality rates. The most significant effect of this decline is the aging of the population.
The model proposed by Lee and Carter (1992) is as follows:

$$ln(m(x,t)) = a(x) + b(x)k(t) + \varepsilon(x,t), \quad x = 1, \dots, \omega \qquad (1)$$

where $m(x,t)$ is the mortality rate of age group x in year t, ω is the beginning of the last age interval, $a(x)$ is the average of $ln(m(x,t))$ over time, $b(x)$ determines which rates change in response to the changes in k_t for age x, $k(t)$ is the mortality index in year t and $\varepsilon(x,t)$ reflects particular age-specific historical influences not fully captured by the model which is independent and identically distributed and follows the $N(0,\sigma^2)$ distribution. Lee and Carter

(1992) estimated the parameters of (1) using a two-stage method with restrictions $\sum_t \hat{k}\,(t) = 0$ and $\sum_t \hat{b}\,(x) = 1$ to ensure a unique solution for the system of equations of the model. The singular value decomposition (SVD) approach was applied to the matrix of centered age profiles $ln(m_{x,t}) - \hat{a}(x)$, which allows a first estimation of parameters $\hat{b}(x)$ and $\hat{k}(t)$. A second step based on the refitting of $\hat{k}(t)$ on the number of deaths is usually suggested to assure a better convergence between the estimated and observed deaths. The aim is to find the $\hat{k}(t)$ such that:

$$\sum_{x=x_1}^{\omega} D(x,t) = \sum_{x=x_1}^{\omega} N(x,t)\, esp(\hat{a}(x) + \hat{b}(x)\,\hat{k}(t)), \qquad (2)$$

where $D(x,t)$ is the number of deaths of age x in time, and $N(x,t)$ is the exposure to risk of age x in time t. For this method, $a(x)$ and $b(x)$ are fixed. The adjusted $\hat{k}(t)$ is then extrapolated using autoregressive integrated moving average (ARIMA) models. The LC model uses a random walk with drift model, which can be expressed as:

$$k(t) = k(t-1) + d + e(t), \qquad (3)$$

where d is known as the drift parameter and measures the average annual change in the series, and $e(t)$ is an uncorrelated error.

Zhao (2012) introduced a modified approach to LC model for analysing short base period age-specific data using linearized cubic splines and other additive functions to estimate the unknown functions, $a(x)$ and $b(x)$. Zhao observed that predicted mortality curves were not smooth and fluctuate significantly over the age range. In this study, we applied the modified model by Zhao (2012) for the Malaysian mortality data (Department of Statistics Malaysia) for the period 1991-2015 for males and females.

2. Methodology

According to Zhao (2012), a piecewise cubic spline function with knots $\Omega_m = \{x_1, x_2, \ldots, x_m\}$ can be expressed as

$$f_{\Omega_m}(x) = f_0(x) I_{x_0 \le x \le x_1} + f_1(x) I_{x_1 \le x \le x_2} + \cdots + f_m(x) I_{x_m \le x \le x_{m+1}} \qquad (4)$$

where $f_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i$, $i = 0,1, \ldots m$ and $I_{x_i \le x \le x_{i+1}}$ is the indicator function with value 1 on the interval $[x_i, x_{i+1}]$ and 0 elsewhere. It is assumed that $f_{\Omega_m}(x)$ and its first and second derivatives are continuous and therefore we have $f_{i-1}(x_i) = f_i(x_i)$, $f'_{i-1}(x_i) = f'_i(x_i)$, $f''_{i-1}(x_i) = f''_i(x_i)$ for $i = 1,2,\ldots, m$. Hence $f_{\Omega_m}(x)$ can be expressed as:

$$f_{\Omega_m}(x) = \beta_{00} + \beta_{01}Z_{01}(x) + \beta_{02}Z_{02}(x) + \sum_{i=0}^{m}\beta_i Z_i(x) \quad (5)$$

where $Z_{0i}(x) = (x - x_0)^i{}_+$; i = 1,2 and $Z_i(x) = (x - x_i)_+{}^3, i = 0,1,\dots m$, (6)

and $(x - x_j)_+^i$ is $(x - x_j)^i$ for $x \geq x_j$ and 0 for $x < x_j$. Hence, the cubic spline can be written as a linear combination of some non-linear functions and that $f_{\Omega_m}(x)$ has $m + 4$ parameters and is cubic below the first knot x_1 and above the last knot, x_m. Additionally, Zhao (2012) proposed that further quadratic and linear restrictions could be applied to points above the last knot. For quadratic restriction, we set a_m = 0 and for the linear restriction, we set $a_m = b_m = 0$ where a_m and b_m are as per the polynomial $f_m(x)$

Under quadratic restriction above the last knot, we have:

$$\begin{gathered} Z_{0j}(x) = (x - x_0)_+^j, j = 1,2; \\ Z_i(x) = (x - x_i)_+^3 - (x - x_m)_+^3, \ i = 0,1, \dots, m-1; \\ Z_m(x) = 0. \end{gathered} \quad (7)$$

Under the linear restriction above the last knot we have (Zhao, 2012):

$$\begin{gathered} Z_{01}(x) = (x - x_0)_+; \\ Z_{02}(x) = (x - x_0)_+^2 - \frac{1}{3(x_m - x_{m-1})}\left((x - x_{m-1})_+{}^3 - (x - x_m)_+{}^3\right) \\ Z_i(x) = (x - x_i)_+^3 - (x - x_{m-1})_+^3 \frac{x_m - x_i}{x_m - x_{m-1}} + (x - x_m)_+^3 \frac{x_{m-1} - x_i}{x_m - x_{m-1}}, \ i = 0,1, \dots, m-2 \\ Z_{m-1}(x) = Z_m(x) = 0. \end{gathered} \quad (8)$$

The number of parameters corresponding to the quadratic and linear restrictions is $m + 3$ and $m + 2$ respectively. Furthermore, we can also apply linear or quadratic restrictions below the first knot. For the quadratic restriction, we set $Z0\ (x) = 0$ and for the linear restriction we set $Z_{02}(x) = Z_0\ (x) = 0$.

To obtain the modified LC model, we assume that $D(x,t)$ in (2) has a binomial distribution than we have $D(x,t) \sim Bin(n(x,t), p(x,t))$ where $p(x,t)$ is the mortality rate of an individual of age group x in year t. Hence the proposed model is of the form:

$$\ln\left(\frac{p(x,t)}{1-p(x,t)}\right) = a(x) + b(x)k(t), \quad (9)$$

where $a(x), b(x)$ and $k(t)$ are unknown.

We follow the methodology proposed by Zhao (2012) that is to approximate these functions as linear combination of a cubic spline with possible restrictions on the left and right tails. Additionally, other additive functions, $Z(x)$ such as $1/x, log(x)\ and\ 1/\sqrt{x}$ is added to the estimated function. According to Zhao (2012), for short based period it is sufficient assume that $k(t) = t$. We let a fixed set of knots Ω_m and write $a(x) = f_{\Omega_m}(x)$ and $b(x) = g_{\Omega_m}(x)$ and hence the model can be expressed as:

$$\ln\left(\frac{p(x,t)}{1-p(x,t)}\right) = f_{\Omega_m}(x) + g_{\Omega_m}(x)t + \gamma Z(x). \quad (10)$$

More specifically, an m-knot model can be expressed as

$$\ln\left(\frac{p(x,t)}{1-p(x,t)}\right) = \beta_{00,f} + \beta_{01,f}Z_{01}(x) + \beta_{02,f}Z_{02}(x) + \sum_{i=0}^{m}\beta_{i,f}Z_i(x) + t\beta_{00,g} + t\beta_{01,g}Z_{01}(x) + t\beta_{02,g}Z_{02}(x) + t\sum_{i=0}^{m}\beta_{i,g}Z_i(x) + \gamma Z(x). \quad (11)$$

where $p(x,t)$, γ, $\beta_{0j,f}(j = 0,1,2)$, $\beta_{i,f}(i = 0,1,\dots,m)$, $\beta_{0j,g}$ $(j = 0,1,2)$ and $\beta_{i,g}(i = 0,1,\dots,m)$ are the parameters to be estimated. The functions $Z(x)$, Z_{0j} (x) $(j = 1,2)$ and $Z_i(x)$ $(i = 0,1,\dots,m)$ are defined in equations (6), (7) and (8) depending on the restriction above the last knot, x_m. Depending on application of linear or quadratic restrictions below the first knot, we set Z_{02} $(x) = Z_0$ $(x) = 0$ or Z_0 $(x) = 0$ respectively.

3. Results and Discussion

The modified LC model was applied to the Malaysian mortality data (1991-2015) obtained from the Department of Statistics Malaysia. We used the middle point of the age period to represent the age group since the data is only available in intervals. To select a suitable model, we will select the model from a selection of candidate models with the lowest residual deviance. For ease of computation, we added 1 to x as it is undefined at 0. The number of knots available in the data is small due to the lack of granularity of the age groups. As such, 3 sets of knots(see Figure 1 for shape of curve) were selected i.e. {7,12,17,22,27}, {12, 22, 27,32} and {17,22,27}. The models listed in Table 1 are based on gender, additive function, restriction on the tail, number of parameters and the residual deviance for each model. We used the 'glm' and 'logit' link in R to fit the logistic regressions.

In general, it is observed (Figure 1) that the mortality rate decreases from birth (age group 0) until age group 12 and increases thereafter. It also appears that the change in death rates for younger age groups is more significant than the change in mortality rates for older ages. Compared to the female mortality rate, male mortality is generally higher for all age groups. Additionally, it is observed that there is a significant increase in mortality rate between age groups 17 and 22 for male mortality data compared to female mortality data.

We considered Male 1 and Female 1 to be the best models for each gender. The parameters and their corresponding p-values are listed in Table 2. From Table 2, it is observed all the parameters of Male 1 and Female 1 are significant.

Table 1. Some selected models with corresponding knots and residual deviance, Male 1 to Male 15 for males and Female 1 to Female 15 for females for mortality data in Malaysia (1991-2015)

Model	$Z(X = x + 1)$	**Left tail**	**Right tail**	**Knots**	**Number of parameters**	**Residual deviance**
Male 1	$1/X$	cubic	cubic	7,12,17,22,27	19	8658
Male 2	$1/X$	quadratic	cubic	7,12,17,22,27	17	9779
Male 3	$log\ X$	linear	cubic	12,22,27,32	13	10410
Male 4	$log\ X$	quadratic	cubic	12,17,22	13	10754
Male 5	$1/\sqrt{X}$	linear	cubic	12,22,27,32	13	11738
Female 1	$1/X$	cubic	cubic	7,12,17,22,27	19	6897
Female 2	$1/X$	quadratic	cubic	7,12,17,22,27	17	7125
Female 3	$1/\sqrt{X}$	linear	cubic	12,17,22,32	13	7525
Female 4	$1/\sqrt{X}$	linear	cubic	17,22,27	11	7477
Female 5	log X	linear	cubic	17,22,27	11	7628

Table 2. Parameter estimates and p-values for Male 1 and Female 1

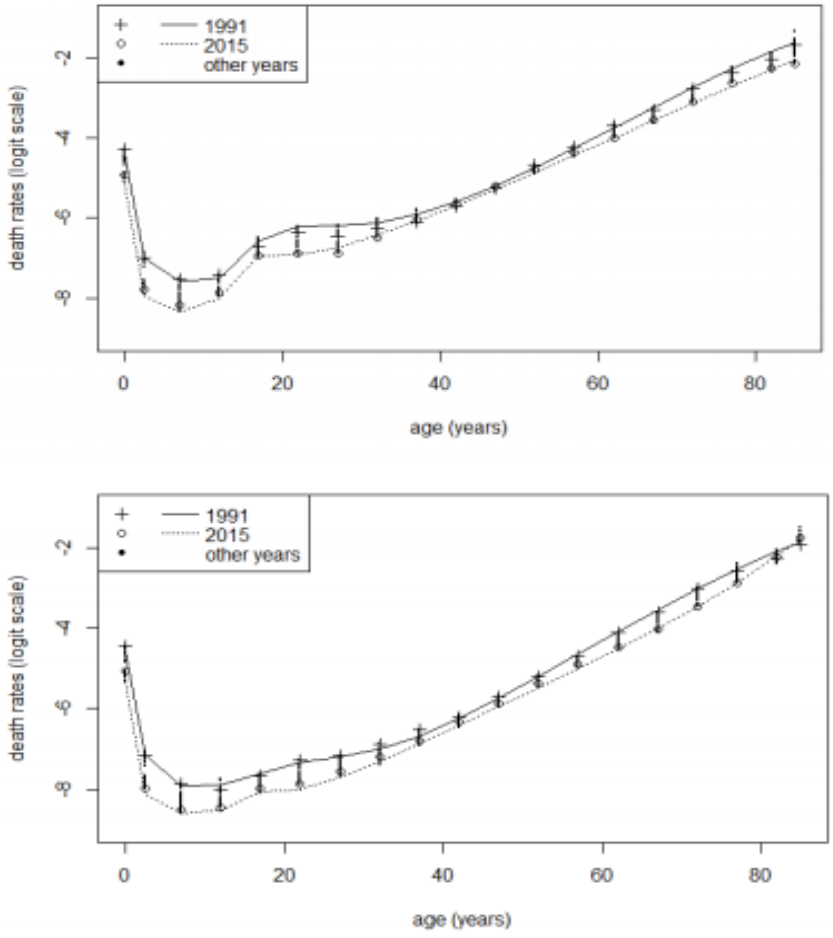
	Model Male 1		**Model Female 1**	
Function	Estimate	p-value	Estimate	p-value
Intercept	40.09	<0.0001	-41.19	<0.0001
$Z(x) = 1/(x + 1)$	18.59	<0.0001	98.89	<0.0001
$Z_{01} = x_+$	20.29	<0.0001	60.14	<0.0001
$Z_{02} = x_+^2$	-4.293	<0.0001	-11.76	<0.0001
$Z_0 = x_+^3$	0.2334	<0.0001	0.6746	<0.0001
$Z_1 = (x-7)_+^3$	-0.2882	0.0005	-0.9282	<0.0001
$Z_2 = (x-12)_+^3$	0.1692	0.0093	0.4581	<0.0001
$Z_3 = (x-17)_+^3$	-0.3204	<0.0001	-0.4142	<0.0001
$Z_4 = (x-22)_+^3$	0.3229	<0.0001	0.3227	<0.0001
$Z_5 = (x-27)_+^3$	-0.1185	<0.0001	-0.1156	<0.0001
t	-0.03168	<0.0001	-0.003129	<0.0001
$tZ_{01} = tx_+$	-0.006892	<0.0001	-0.009127	<0.0001
$tZ_{02} = tx_+^2$	0.001624	<0.0001	0.002504	<0.0001
$tZ_0 = tx_+^3$	-0.00008946	<0.0001	-0.0001608	<0.0001
$tZ_1 = t(x-7)_+^3$	-0.0001132	0.0045	0.0002577	<0.0001
$tZ_2 = t(x-12)_+^3$	-0.00008143	0.0121	-0.0001899	<0.0001
$tZ_3 = t(x-17)_+^3$	0.0001614	<0.0001	0.0001961	<0.0001
$tZ_4 = t(x-22)_+^3$	-0.0001621	<0.0001	-0.0001589	<0.0001
$tZ_5 = t(x-27)_+^3$	0.00005913	<0.0001	0.00005714	<0.0001

We used MAPE and MSE to examine the goodness-of-fit of the original LC model and modified models for the Malaysian mortality data. The modified method produced smaller MAPE and MSE compared to the LC model for the Malaysian data (see Table 3).

Table 3. MAPE and MSE of the original LC and the modified LC models

Gender		**LC**	**Modified LC**
Male	MAPE	0.0411	0.0155
	MSE	0.4874	0.0081
Female	MAPE	0.0232	0.0134
	MSE	0.0054	0.0007

Figure 1. Observed (points) and fitted (lines) rates for males in Malaysia (logit scale) in 1991 and 2015 and other years using (a) Male 1 and (b) Female 1 models



4. Conclusion

A major advantage of estimating mortality rates using cubic spline is that it allows for age-specific mortality estimates compared to the original LC model. This allows for more specific age group mortality estimation that is we can estimate mortality rates for ages which are not the midpoint of an age group. Additionally, compared to the LC Model which is not smooth, the cubic spline estimate is continuous up to the second derivative. Another advantage of the cubic spline method is the potentially lower number of model parameters. In comparison, the number of parameters using the spline approach would be determined by the number of knots and restrictions used. One of the disadvantages in fitting a cubic spline to the mortality dataset of Malaysia is the lack of granularity of the age groups as it limits the choice of knots. However, as the estimated mortality models the error is relatively small, it is reasonable to say that the model provides adequate estimates.

References

1. Hyndman, R.J. (2006). R package, demography: Forecasting Mortality, Fertility, Migration and Population Data.
2. Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 87(419), 659-671.
3. Zhao, B. B. (2012). A modified Lee–Carter model for analysing short-base-period data. *Population studies*, 66(1), 39-52.

The Lee-Carter Model: Extensions and applications to Malaysian mortality data

Nurul Aityqah Yaacob[1,2], Dharini Pathmanathan[1], Ibrahim Mohamed[1], Siti Haslinda Mohd Din[3]

[1]Institute of Mathematical Sciences, Faculty of Science, University of Malaya, Kuala Lumpur
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Negeri Sembilan, 72000 Kuala Pilah, Negeri Sembilan
[3]Department of Statistics Malaysia, Federal Government Administrative Centre, Putrajaya

Abstract

This study examines the application of Lee-Carter (LC) model and some of its extensions to Malaysia mortality data. The parameters were estimated by using the singular value decomposition (SVD) method while ARIMA (p,d,q) was used to forecast the mortality index. We find that, the log mortality rates for all populations decreased and the female population in Malaysia is expected to have longer life compared to the male population.

Keywords

Lee-Carter model; ARIMA; mortality; age-specific death rates; forecast

1. Introduction

Mortality forecasts have a long history in demography for population projections and actuarial science. Actuaries applied mortality forecasts for cash flow projections and assessment of premium and reserves in life insurance and pension. Various models have been proposed since Gompertz published the law of mortality in 1825. Commonly used methods in demographic forecasting such as extrapolation, explanation and expectation. Extrapolation is the most popular approach in demographic forecasting. The LC model which has been widely applied in mortality forecasting uses the extrapolation approach. The model is developed by Lee & Carter in 1992 to forecast mortality rates in the United States from 1990 to 2065 (Lee & Carter, 1992).

Lee & Miller (2001) found that the LC model did not perform well for United States when using the fitting period 1900-1989 to forecast the period 1990-1997. The pattern of change in mortality was not fixed over time. Due to the different age patterns of change for 1900-1950 and 1950-1995, the fitting period is reduced to commence in 1950 for the Lee-Miller (LM) variant (Booth et al., 2005).

The Booth-Maindonald-Smith (BMS) variant was used to fit Australian data from 1907 to 1999 and addressed two main issues in the original LC model; linearity in estimated parameter k_t and invariance in b_x (Booth et al., 2002). Thus, the optimal fitting period was applied, so the assumption of invariant b_x

was better met and the problem associated with violation of this assumption was avoided.

2. Methodology

The data sets used in this study were obtained from the Department of Statistics Malaysia (DOSM) and it consists of Malaysia mortality (Peninsular Malaysia, Sabah and Sarawak) for the period of 1991 to 2015. The mortality data for female, male and total population will measure by age-specific death rates (ASDR). The ASDR is calculated as $m_{x,t} = d_{x,t}/p_{x,t}$, where $d_{x,t}$ is the number of deaths for a group of x ages in year t and $p_{x,t}$ is the observed population (exposure to risk) for a group of x ages in year t. The data set consist of ASDR for 19 age groups which are organized into 5-year intervals such as 0, 1-4, 5-9... 85+.

The LC model is as follows:

$$\ln(m_{x,t}) = a_x + b_x k_t, + \varepsilon_{x,t} x = 0, \ldots\ldots, nt = 1, \ldots, n \tag{1}$$

where $m_{x,t}$ is the age-specific death rate for the x interval and the year t, k_t is the mortality index in the year t, a_x is the average age-specific mortality, b_x is a deviation in the mortality due to changes in the k_t index and $\varepsilon_{x,t}$ is the random error.

As the solution is not possible to be unique, the following constraints were imposed (Lee. & Carter, 1992):

$$a_x = \frac{1}{T} \sum_{t=t_1}^{t_n} ln(m_{x,t}) \tag{2}$$

$$\sum_{x=x_1}^{x_n} b_x = 1 \tag{3}$$

$$\sum_{t=t_1}^{t_n} k_t = 0 \tag{4}$$

Applying SVD to the matrix

$$Y_{xt} = ln(m_{x,t}) - \tilde{a}_x \tag{5}$$

produced

$$ULV' = SVD(Y_{xt}) = L_1 U_{x1} V_{t1} + \cdots + L_X U_{xX} V_{tX} \tag{6}$$

where U represents the age component, L is the singular value and V represents the time component. The first step in forecasting mortality via LC model is estimating a_x , b_x and k_t using historical age specific mortality rates. The estimates of $\tilde{a}_x$ can be obtained by finding the average over time of $ln(m_{x,t})$, $\tilde{a}_x = \frac{1}{T}\sum_{t=t_1}^{t_n} ln(m_{x,t})$. The estimates of $\tilde{b}_x = U_{x1}$ and $\tilde{k}_t = L_1 V_{t1}$ can be obtained by approximating the first term (Wang, 2007).

As the first stage of estimation is based on logs of death rates rather than the death rates themselves, there will be a fairly large disparity between predicted and actual deaths. Therefore, re-estimation of k_t is necessary, by taking thea_x and b_x estimates. In order to search for k_t such that:

$$\sum_{x=x_1}^{x_n} D_{xt} = \sum_{x=x_1}^{x_n} e^{(a_x + b_x k_t)} N_{x,t} \quad (7)$$

Where $\sum_{x=x_1}^{x_n} D_{xt}$ is the total number of deaths in year t and $N_{x,t}$ is the population (exposure to risk) of age in year t . The estimated k_t was adjusted to ensure equality between the observed and estimated number of deaths in a certain period. Lee and Carter found an appropriate ARIMA time series model for the mortality index k_t . They proposed the standard univariate ARIMA (0,1,0) time series model which is a random walk with drift, as an appropriate model to forecast. The model is as follows: (8)

$$\tilde{k}_t = \tilde{k}_{t-1} + \theta + \varepsilon_t$$

where θ is known as the drift parameter and

$$\tilde{\theta} = \frac{\tilde{k}_t - \tilde{k}_{t-1}}{T - 1} \quad (9)$$

Finally, the forecasted values of adjusted k_t and the estimated a_x and b_x had substituted into equation (1) to get the forecasted values of $log(m_{x,t})$ in order to get forecast mortality rates. The forecasts for age-specific death rates, $m_{x,t}$ can be obtained by using the equation below: (10)

$$m_{x,n+h} = m_{x,n}\, exp\{b_x\,(k_{n+h} - k_n)\}, h = 1,2,\ldots\ x1,2,\ldots.n$$

where n is the last year from which data are available; h is the forecast horizon, and x represents the age group (Andreozzi, Blaconá, & Arnesi, 2011).

The LM variant differ from the LC model by constraining the model such that the jump-of rates are the observed rates in the jump-of year instead of the fitted rates and k_t passes through zero in the jump-off year to avoid jump-off bias. While the BMS variant varies from the LC model when the fitting period is chosen based on statistical goodness-of-fit criteria under the assumption of linear k_t and the jump-off rates are taken to be the fitted rates based on this fitting methodology (Booth et al., 2005). The model also makes an adjustment of k_t by fitting to the age distribution of deaths $D_{x,t}$ using the

Poisson distribution (Booth et al., 2006). Booth, Maindonald, & Smith (2002), also suggested to adopt a maximum likelihood approach in adjustment of k_t.

3. Results

The parameters a_x , b_x and k_t were estimated by using the SVD method. Figure 1 shows the comparison of the parameter estimation for the LC model and its extensions. As expected the average mortality (a_x) grows when age increases. The parameter b_x describes the tendency of mortality at age x to change as the general level of mortality (k_t) changes. This indicates that when b_x is large for some x , the death rate at age x varies a lot than the general level of mortality changes and when b_x small, then the death rate is at that age varies a little.

From Figure 1, a similar trend of parameters was observed for the LC model and its extensions. b_x appears to be highest at younger ages for male, female and total population, which means that the mortality varies when the general mortality index k_t changes.

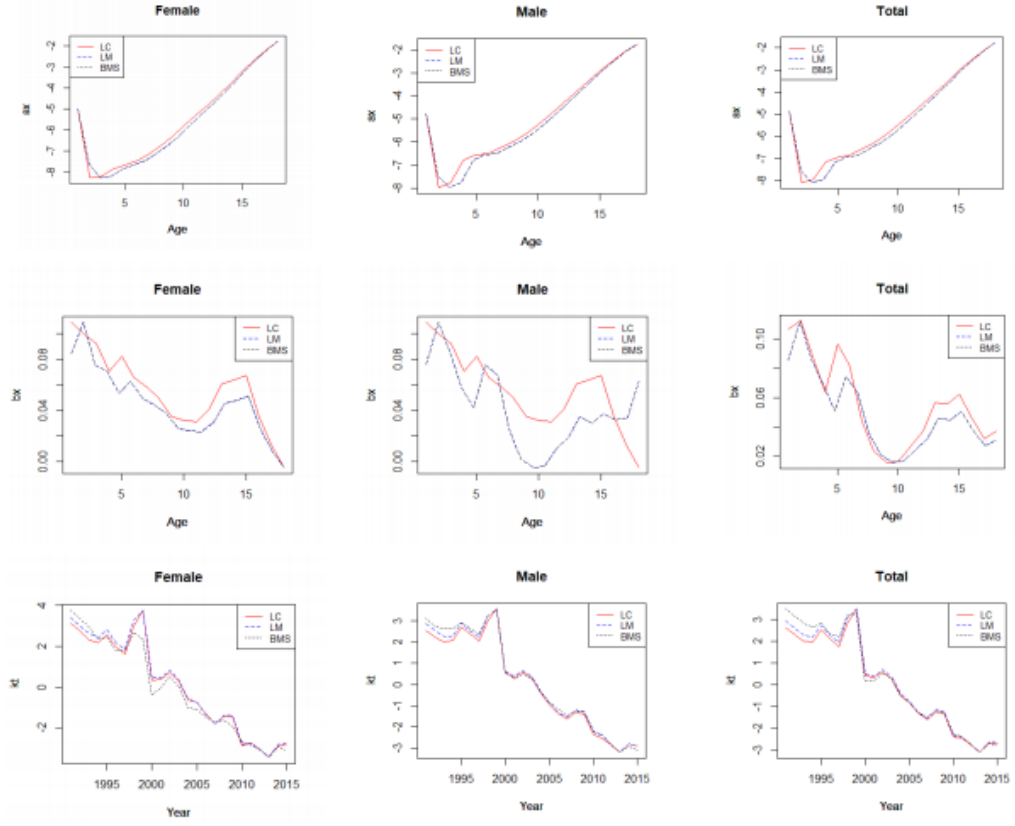


Figure 1: Comparison of parameter estimations

With regard to older age groups, the value of this parameter is lower, which means that the mortality slightly varies in that period of time. Finally, as expected, k_t shows a decreasing trend with increment of time, t.

Table 1 shows the life expectancies of male, female and total population for the next 10 years. From the table, we can see that the life expectancy of both male and female population increases for the next 10 years. The female population in Malaysia is expected to outlive the male population.

Table 1: Estimation of life expectancies for the next 10 years

	Male			Female		
Year	**LC**	**LM**	**MMS**	**LC**	**LM**	**MMS**
2016	68.29	73.13	73.14	72.24	77.05	77.14
2017	68.45	73.30	73.32	72.35	77.16	77.26
2018	68.62	73.47	73.50	72.45	77.27	77.38
2019	68.78	73.65	73.69	72.56	77.37	77.50
2020	68.95	73.82	73.87	72.67	77.48	77.62
2021	69.12	74.00	74.06	72.77	77.59	77.74
2022	69.29	74.18	74.25	72.87	77.69	77.86
2023	69.46	74.36	74.44	72.98	77.80	77.97
2024	69.64	74.55	74.64	73.08	77.90	78.09
2025	69.82	74.73	74.83	73.18	78.00	78.20

Figure 2 shows the estimated mortality index for male, female and total population in Malaysia for 2016 to 2025. It shows a downward trend, which are similar to the observed years (1991 to 2015). ARIMA (0,1,0) is used to forecast the mortality index. We obtain the drift parameter by using equation (9). Estimated k_t can be obtained by finding the sum of drift parameter and k_{t-1} which is shown in equation (8).

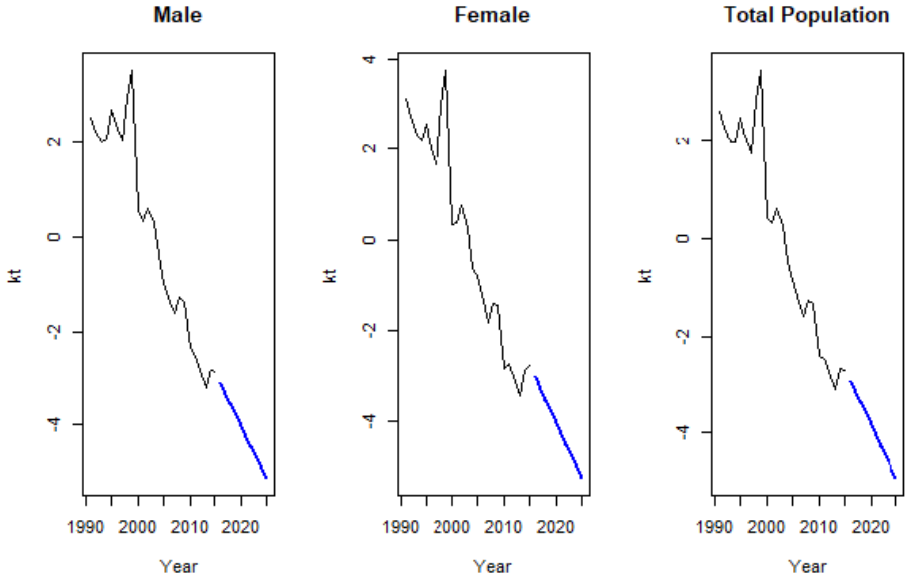


Figure 2: Forecast of $\mathbf{k_t}$ from 2016 to 2025 with ARIMA (0,1,0)

Tables 2, 3 and 4 present the performance of out-of-sample prediction of LC model and its variants for female, male and total populations. We used mean error (ME), root mean squared error (RMSE), mean absolute error (MAE), mean

absolute percentage error (MAPE) and mean absolute scaled error (MASE) to evaluate the accuracy of the LC model and its variants. The result shows that BMS method performed well for female, male and total population for the Malaysian data. Booth et al. (2006) also found that the BMS method have been found to be more accurate than original LC model and the LM variant. The decomposition of error has demonstrated that jump-off bias is a significant source of error for LC (Booth et al., 2005).

Table 2: Evaluation performance of LC method and its variant (Female)

Method	*ME*	*RMSE*	*MAE*	*MAPE*	*MASE*
LC	o.82	0.97	0.82	28.28	1.22
LM	0.87	1.06	0.87	27.37	1.23
BMS	**0.63**	**0.81**	**0.63**	**19.78**	**1.06**

Table 3: Evaluation performance of LC method and its variant (Male)

Method	*ME*	*RMSE*	*MAE*	*MAPE*	*MASE*
LC	0.28	0.44	0.31	10.88	0.56
LM	0.24	0.44	0.34	10.29	0.57
BMS	**0.15**	**0.31**	**0.27**	**7.88**	**0.51**

Table 4: Evaluation performance of LC method and its variant (Total Population)

Method	*ME*	*RMSE*	*MAE*	*MAPE*	*MASE*
LC	0.44	0.57	0.44	16.27	0.77
LM	0.47	0.65	0.47	15.64	0.76
BMS	**0.42**	**0.58**	**0.42**	**13.75**	**0.78**

4. Conclusion

Based on our observation, we can see that the log mortality rates for male, female and total population decreased as time passed and females in Malaysia are expected to outlive the male population. The BMS variant performed best for Malaysian data compared to its two counterparts examined in this study. In conclusion, all the methods with actual rates taken as jump-off rates performed better than fitted rates (Zakiyatussariroh et al. 2014).

References

1. Andreozzi, L., Blaconá, M. T., & Arnesi, N. (2011). The Lee Carter Method for Estimating and Forecasting Mortaity: An Application for Argentina. In *International Symposium on Forecasting-2011, Prague Proceedings*, 2050(5), 1–17.

2. Booth, H., Hyndman, R. J., Tickle, L., & De Jong, P. (2006). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. https://doi.org/10.4054/DemRes.2006.15.9
3. Booth, H., Maindonald, J., & Smith, L. (2002). Applying Lee – Carter Under Conditions of Variable Mortality Decline. *Population Studies*, 56(3), 325–336. https://doi.org/10.1080/00324720215935
4. Booth, H., Tickle, L., & Smith, L. (2005). Evaluation of the variants of the Lee-Carter method of forecasting mortality: A multi-country comparison. *New Zealand Population Review*, 31(1), 13–34.
5. Lee, R., & Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, 38(4), 537–549.
6. Lee, Ronald D. & Carter, L. R. (1992). Modeling and Forecasting U. S. Mortality. *Journal of the American Statistical Association*, 87(419), 659. https://doi.org/10.2307/2290201
7. Wang, J. Z. (2007). *Fitting and Forecasting Mortality for Sweden: Applying the Lee-Carter Model*. Mathematical Statistics, Stockholm University.

Multi-aspect permutation methods for cytomorphometric data under multivariate directional alternatives with application to comparative neuroanatomy

Livio Corain[1], Bruno Cozzi[1], Jean-Marie Graïc[1], Ludovica Montanucci[1], Luigi Salmaso[1], Antonella Peruffo[1], Ruben Carvajal-Schiaffino[2], Enrico Grisan[3]

[1]University of Padova, Italy
[2]University of Santiago de Chile, Chile
[3]King's College London, United Kingdom

Abstract

When bio-medical imaging shape data refer to multiple single-cell morphological features and the goal is to inferentially compare different populations, it appears that traditional statistical shape analysis methods are not suitable for handling multivariate directional alternatives. This is actually a central issue because it does not allow to draw conclusions on whether some populations have cells that are smaller/more regular/denser vs. larger/irregular/sparse. After organizing the neural cell descriptors in multidimensional domains such as size, regularity and density, we propose a data representation model in the form of a two-way multivariate linear effects model. On the related location and scatter parameters, i.e. by using a multi-aspect approach, and under multivariate directional alternatives, we propose to apply the union-intersection combination-based methodology as inferential method to separately test and rank the possible equality vs. dominance of two or more populations. We numerically prove the effectiveness of the proposed methodology through a simulation study where cell shape data were obtained by simulating slices from randomly generated geometric solids within a volume. Finally, we applied the proposed procedure to a comparative neuroanatomy study aimed at quantifying possible morphometric structural differences in the brain cytoarchitecture of three sex-related bovine populations, i.e. male, female and natural intersex.

Keywords

Nonparametric combination; multivariate ranking; permutation tests

1. Introduction

Morphometrics or morphometry is a quantitative way of addressing the shape comparisons that have always interested biologists. In neuroscience structural differences in the brain cytoarchitecture represent the anatomical substrate underlying the functional differences. Especially in the field of neurodegenerative pathologies, studying structural changing in brain tissue could be a powerful instrument to carry out morphometric analysis providing

robust bases for objective tissue screening. In this view, cell shape analysis is a powerful proxy of cellular function (Lobo et al., 2016).

Cell-oriented shape analysis differs from classical shape analysis supporting traditional bio-medical morphometric studies in one important issue: because of their relatively small sizes (at most a few tens of microns), cell shapes are generally much more regular than traditional anatomical shapes (objects of size usually in the scale of centimetre, such as bones, anatomical regions, etc.). As a result, instead of focusing on spatial landmarks, the endpoints of interest in cell shape analysis are cytomorphometric descriptors such as area, perimeter, axis length, etc. In this view, there is no need for pre-processing morphometric data by translation, scaling or rotation methods such as procrustes superimposition.

As about the statistical hypothesis testing methods for geometric shape analysis, it turns out that traditional testing approaches refer to quadratic-like statistics (Rohlf, 2000), so that they are not suitable to handle multivariate directional alternatives. This is a central issue in cell shape analysis because formalizing the departure from the null hypothesis by one specific direction allows us to draw conclusions on whether some populations have smaller vs. larger or more regular vs. irregular cells, or finally denser vs. sparser cell density. According to the so-called multi-aspect permutation paradigm (Brombin and Salmaso 2009), in this paper we face the cell shape testing problem by focusing the attention into two different aspects: the location-aspect, based on the comparison of location-related statistics, and the scatter-aspect, based on the comparison of variability-related statistics.

2. Methodology

Let us assume that we are facing a comparative neuroanatomy problem involving a number of $C \geq 2$ populations that are defined according to the levels of one factor of interest such as sex, age, specie, etc. In order to formalize the comparison between the C populations, we assume a two-way experimental cytomorphometric data representation model where we represent as $\mathbf{Y}$ a dataset of size $n = \Sigma j\, n_j$, where p morphometric features (belonging to a given morphometric domain such as size, regularity and density) have been measured on the i-th cell, located in the l-th brain region layer, and belonging to the j-th population. Let us assume that the p-variate response variable the can be modelled as

$$\mathbf{Y}_{ijl} = \boldsymbol{\mu} + \boldsymbol{\tau}_j + \boldsymbol{\beta}_l + (\boldsymbol{\tau\beta})_{jl} + \boldsymbol{\varepsilon}_{ijl}, \qquad (1)$$

where $\boldsymbol{\varepsilon}_{ijl}$ are i.i.d. possibly non-Gaussian error terms with null mean and population/region-dependent scale coefficients $\boldsymbol{\sigma}_{jl}^2$ and unknown distribution P_{ε}, $\boldsymbol{\mu}$ is a population-invariant constant, coefficients $\boldsymbol{\tau}_j$ represent the main population effects, $\boldsymbol{\beta}_l$ and $(\boldsymbol{\tau\beta})_{jl}$ refer to location effects due to the

brain region/layer and the interaction between population and region, and $\boldsymbol{\sigma}_{jl}{}^2$ are population-and-region varying scale coefficients which may depend, through monotonic functions, on main treatment effects $\boldsymbol{\tau}_j$ and brain region/layer location effects $\boldsymbol{\beta}_l$. Basically, according to the so-called multivariate generalized Behrens-Fisher problem (Yanagihara and Yuan, 2005), the proposed data representation model is a quite general less-demanding two-way linear effect nonparametric model where specific location and scale effects are both allowed to differ across populations and brain regions.

Since the study's main goal is to compare populations, both jointly across all regions and separately within any given region, we are actually inferring on the sum of location coefficients $\boldsymbol{\tau}_j + (\boldsymbol{\tau\beta})_{jl}$. By using the Roy's Union-Intersection testing approach (Roy, 1953; Pesarin and Salmaso, 2010), let us formalize, separately for the location and scatter parameters, the comparison between the *j*-th and the *h*-th population with the null and alternative hypotheses as follow

$$\begin{cases} H_{0(jh)}^{loc}: \cap_l \cap_k Y_{jlk} \overset{loc}{=} Y_{hlk} \equiv \cap_l \cap_k [\tau_{jlk} = \tau_{hlk}] \\ H_{1(jh)}^{loc}: \cup_l \cup_k [(Y_{jlk} \overset{loc}{<} Y_{hlk}) \cup (Y_{jlk} \overset{loc}{>} Y_{hlk})] \\ \equiv \cup_l \cup_k [(\tau_{jlk} < \tau_{hlk}) \cup (\tau_{jlk} > \tau_{hlk})] \equiv H_{1(jh)}^{loc,-} \cup H_{1(jh)}^{loc,+} \end{cases} \quad \begin{cases} H_{0(jh)}^{sca}: \cap_l \cap_k Y_{jlk} \overset{sca}{=} Y_{hlk} \equiv \cap_l \cap_k [\sigma_{jlk}^2 = \sigma_{hlk}^2] \\ H_{1(jh)}^{sca}: \cup_l \cup_k [(Y_{jlk} \overset{sca}{<} Y_{hlk}) \cup (Y_{jlk} \overset{sca}{>} Y_{hlk})] \\ \equiv \cup_l \cup_k [(\sigma_{jlk}^2 < \sigma_{hlk}^2) \cup (\sigma_{jlk}^2 > \sigma_{hlk}^2)] \equiv H_{1(jh)}^{sca,-} \cup H_{1(jh)}^{sca,+} \end{cases} \quad (2)$$

where $\tau_{\boldsymbol{jlk}} = \tau_{jk} + (\tau\beta)_{jlk}$., and $l = 1, \ldots, L, k = 1, .., p$, are the reference indexes for each brain region and univariate response, i.e. cell morphometric indicator, respectively. It is worth noting that hypotheses (2) refers to a general version of the so-called generalized Beherens-Fisher problem (Yanagihara and Yuan, 2005). Under the null hypotheses of joint equality, either in location and in scatter, actual data are exchangeable random components that can be permuted between combinations of groups and strata in order to derive, separately for the location and scatter problems, two multivariate directional p-values. Once we removed the nuisance individual location effect $\boldsymbol{\tau}_j + (\boldsymbol{\tau\beta})_{jl}$ by computing suitable individual-free residuals, as univariate location and scatter test statistic we respectively used the difference of sample means and squared deviations along with the Fisher's combining function, as combination strategy to derive the multivariate p-values (Pesarin and Salmaso, 2010). In this view, when just one between $H_{0(jh)}^{loc}$ and $H_{0(jh)}^{loc}$ are true, the permutation approach can be considered as an approximated nonparametric testing solution. For a more in depth understanding of the testing procedure we shortly sketched here we refer the readers to Pesarin and Salmaso (2010).

By exploiting the multivariate one-sided alternatives in expression (2), we may derive two sorts of location and scatter rankings using the ranking methodology proposed by Arboretti et al. (2014). In fact, by suitable combining information from directional multivariate p-values, the possible underline latent ordering among $\boldsymbol{\tau}_{\boldsymbol{jl}}$ and $\boldsymbol{\sigma}_{jl}{}^2$ parameters can be properly estimated. In a nutshell, the rationale behind the ranking within a multivariate

setting is the following: if not all $H_{0(jh)}$ in (2) are true, it must exist an ordering [1],[2],...,[*C*] among $\boldsymbol{\tau_{jl}}$ and/or $\boldsymbol{\sigma}_{jl}^2$ such that

$$\boldsymbol{\tau}_{[1]} \leq \boldsymbol{\tau}_{[2]} \leq \ldots \leq \boldsymbol{\tau}_{[C]} \text{ and } \boldsymbol{\sigma}^2_{[1]} \leq \boldsymbol{\sigma}^2_{[2]} \leq \ldots \leq \boldsymbol{\sigma}^2_{[C]}, \quad (3)$$

where "≤" should be intended as "<" when there exists at least one univariate component of $\tau_{\boldsymbol{jl}}$ and $\boldsymbol{\sigma}_{jl}^2$ for which the strict inequality holds, and at the same time, there isn't any univariate component of $\tau_{\boldsymbol{jlk}}$ and $\boldsymbol{\sigma}_{jlk}^2$ for which the opposite strict inequality holds. When the last condition is not true, "≤" is intended as "=", meaning that the two $\tau_{\boldsymbol{jl}}$ and $\boldsymbol{\sigma}_{jl}^2$ are tied in the ranking. Note that, within a multivariate setting, we state that two multivariate parameters are tied not only when all univariate $\tau_{\boldsymbol{jlk}}$ and $\boldsymbol{\sigma}_{jlk}^2$ are equal but also when that $H^{-}_{1(jh)}$ and $H^{+}_{1(jh)}$ + are *jointly* true. For more details on the ranking methodology within a multivariate setting we refer the reader to Arboretti et al. (2014).

3. Result

Simulation Study

In order prove the effectiveness of the proposed methodology we performed a simulation study where cell shape data were obtained by simulating artificial slices from randomly generated 3D geometric solids within a virtual volume (see Figure 2). By applying imaging routines for detection and outline of simulated cell contours (Grisan et al., 2018), we got a set of cell morphometric indicators (listed in Table 1). Subsequently, we added to those values some random variation according to three multivariate distributions: normal, Student's t with 3 d.f. and an *g*-and-*h* right skewed distribution, as examples of a heavy-tailed and skewed distributions respectively. The advantage of using this kind of strategy to simulate cell morphometric data instead of simply directly generate those data is twofold: first, we can naturally capture the underlying geometrical correlation among size and regularity-related descriptors; secondly, our virtual cuts induces a level of variability which simulated the actual laboratory process of brain samples processing and cutting to obtain the tissue slices. Due to editorial limitations, we cannot include here results of our simulation study and we refer the interested readers to the forthcoming extended version of the present short paper.

Application to Comparative Neuroanatomy

We applied the proposed procedure to a comparative neuroanatomy study aimed at quantifying possible morphometric structural differences in the brain cytoarchitecture of three sex-related bovine populations, i.e. male, female and intersex freemartins. The freemartin syndrome is the most common form of intersexuality in the bovine species, arising from vascular connections between the placentas of heterosexual twin foetuses (Graïc et al., 2018). A series of 10 male, 10 female and 8 freemartin adult bovine brain (24

months old), were obtained from local abattoirs in the Veneto region. Animals were treated according to the present European Community Council directive concerning animal welfare during the commercial slaughtering process and were constantly monitored under mandatory official veterinary medical care.

The posterior domain (lobules VIII and IX) of each specimen was cut into 8 µm thick parasagittal sections. For each cerebellar sample, one section every five was stained (a total of 10 slides per individual per sex) according to the Nissl protocol (Cozzi et al., 2017). Ten stained sections per subject per sex were scanned with a semi-automated microscope equipment at a magnification of 40x in fast mode with automatic focusing, saving the acquisition as Jpeg2000 images. The topographical organization of the bovine cerebellum was analyzed on the Nissl-stained coronal sections, and the histology of the cerebellar cortex resulted uniform over the entire lobules VIII-IX of the vermis (Figure 1.A,1.B,1.C). Three layers were distinguishable from the outer to the inner cerebellar cortex: (i) the superficial molecular layer, (ii) the Purkinje layer and (iii) the inner granular layer (Figure 1.D,1.E,1.F).

The complete analysis of the acquired imaged of cerebellar slices involves the detection and outline of tens of thousands of cells. This is not feasible by human annotation of the images, unless the procedure is carried out in small region of interest, potentially introducing bias in the procedure. To tackle the problem, we developed an automatic procedure (Grisan et al., 2018) that can process the images identifying the position and outline of most of the visible cells, taking care of the different sizes among the different cell populations, and at the same time addressing the packed and clustered appearance of cells in the different layers of the cerebellum, and in particular in the granule cells layer, where they are most packed.

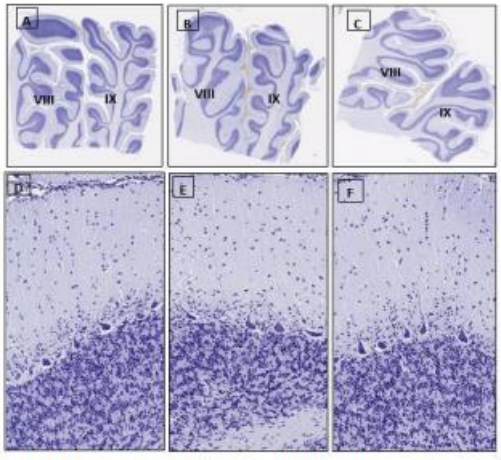


Figure 1: Photomicrographs of 8 µm thick parasagittal Nissl-stained sections of the bovine lobule VIII and lobule IX (Male: A,D; Female: B,E; freemartin: C.F). Purkinje cells are located in a monolayer between the inner granular layer and the outer molecular layer of the cortex.

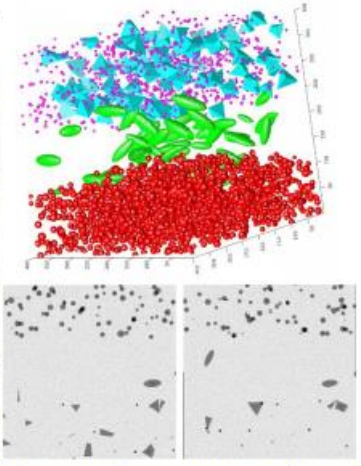


Figure 2: on the top, an example of virtual volume filled by four types of virtual cells that were randomly generated as 3D geometric solids. On the bottom, two simulated artificial slices.

Each identified cell was described by a set of morphometric descriptors characterizing its shape and its local relationship with surrounding cells (Table 1). These measures can be broadly assigned to three domains: size, regularity and density. Size and regularity address cell morphology and are composed of classic measures on shapes, while density attempts to characterize the context around each cell by counting the number of cells present within a radius of 50 μm or within 100 μm around it.

Morphological domain	Morphometric descriptor	Description / mathematical formula
Size	Area	Area of the cell body expressed in μm^2
	Perimeter	Total length of neural cell boundary expressed in μm
	Major axis length	Measure of the length of the major axis of the cell body expressed in μm
	Minor axis length	Measure of the length of the minor axis of the cell body expressed in μm
Regularity	Solidity	Proportion of pixels in the convex hull that are also in the region of the cell
	Extent	Area/(Area of the bounding box)
	Inv.AR (1/AR)	Inverse of the Aspect Ratio, defined as (Major axis length)/(Minor axis length)
	Convex circularity	$(4 \times \pi \times \text{Convex Area})/(\text{Convex Perimeter}^2)$
Density	Ngb_50	No. of neighbor cells counted within a radius of 50 μm all around a given cell
	Ngb_100	No. of neighbor cells counted within a radius of 100 μm all around a given cell

Table 1: morphological domains and morphometric descriptors, along with their description and/or mathematical formula. Actual data were obtained by using corresponding Matlab functions. Convex circularity was used instead of traditional circularity in order to avoid meaningless values that can result in case of very small cells

It is worth noting that for all size-related morphometric measures the rule "the larger they are, the larger is the neuron dimension" applies. Seemingly, for all regularity-based descriptors the rule "the larger they are, the more regular is the neuron" takes place; note that all regularity-based descriptors are measured as dimensionless ratios bounded in the closed interval [0;1]. Finally, both density-related descriptors refer to the less or large amount of neighbour cells that are placed all around a given cell.

We applied the proposed multi-aspect permutation testing and ranking method for cytomorphometric data either jointly across all three cortical layers or separately for each layer (external molecular, Purkinje and granular layer). As in Table 2, we analysed morphometric data by describing the location and scatter results separately for each one morphological domain, i.e. size, regularity and density respectively. Results of multi-aspect permutation-based testing and ranking, are presented in Table 2.

Table 2: Testing and ranking results by layer, domain and aspects. Pairwise between-populations location and scatter one-sided adjusted permutation *p*-values are presented in squared matrices. In each cell the alternative hypothesis is "population-in-row is larger than population-in-column". The 5% significant p-values are highlighted in bold. Location and scatter rankings are

derived from dominance in pairwise comparisons. In red female population (F), in blue male population (M), in green freemartin population (FM)

All layers

LOCATION adjusted *p*-values	Domain: Size			Domain: Regularity			Domain: Density		
	FM	F	M	FM	F	M	FM	F	M
FM		.003	.001		.272	.003		.003	.001
F	1.000		.523	.423		.001	.351		.001
M	1.000	.003		1.000	1.000		1.000	1.000	
ranking=	1	3	2	1	1	3	1	2	3
	scatter ranking: F = M > FM			location ranking: FM > M = F			location ranking: FM > F > M		

SCATTER adjusted *p*-values	Domain: Size			Domain: Regularity			Domain: Density		
	FM	F	M	FM	F	M	FM	F	M
FM		1.000	1.000		1.000	1.000		1.000	.003
F	.003		.081	.002		1.000	.003		.001
M	.001	1.000		.003	.030		1.000	1.000	
ranking=	3	1	1	3	1	1	2	1	3
	scatter ranking: F = M > FM			scatter ranking: F = M > FM			scatter ranking: F > FM > M		

Molecular layer

LOCATION adjusted *p*-values	Domain: Size			Domain: Regularity			Domain: Density		
	FM	F	M	FM	F	M	FM	F	M
FM		.003	1.000		1.000	1.000		.003	.001
F	1.000		1.000	.006		1.000	1.000		.001
M	.001	.003		.015	1.000		1.000	1.000	
ranking=	2	3	1	3	1	1	1	2	3
	location ranking: M > FM > F			location ranking: M = F > FM			location ranking: FM > F > M		

SCATTER adjusted *p*-values	Domain: Size			Domain: Regularity			Domain: Density		
	FM	F	M	FM	F	M	FM	F	M
FM		.006	1.000		1.000	.003		.003	.001
F	1.000		1.000	.540		.001	1.000		.001
M	.084	.003		1.000	1.000		1.000	1.000	
ranking=	1	3	1	1	1	3	1	2	3
	scatter ranking: M = FM > F			location ranking: FM = F > M			scatter ranking: FM > F > M		

Purkinje cells

LOCATION adjusted *p*-values	Domain: Size			Domain: Regularity			Domain: Density		
	FM	F	M	FM	F	M	FM	F	M
FM		1.000	.652		1.000	1.000		1.000	.003
F	.003		.003	.377		1.000	.003		.001
M	.858	1.000		.003	.001		1.000	1.000	
ranking=	2	1	2	2	2	1	2	1	3
	location ranking: F > FM = M			location ranking: M > F = FM			location ranking: F > FM > M		

SCATTER adjusted *p*-values	Domain: Size			Domain: Regularity			Domain: Density		
	FM	F	M	FM	F	M	FM	F	M
FM		1.000	1.000		1.000	1.000		1.000	1.000
F	.648		.594	1.000		1.000	1.000		1.000
M	.939	1.000		.653	1.000		.653	1.000	
ranking=	1	1	1	1	1	1	1	1	1
	scatter ranking: F = FM = M			scatter ranking: F = FM = M			scatter ranking: F = FM = M		

Granular layer

LOCATION adjusted *p*-values	Domain: Size			Domain: Regularity			Domain: Density		
	FM	F	M	FM	F	M	FM	F	M
FM		1.000	.003		.003	.001		.003	.001
F	.003		.001	1.000		1.000	1.000		.001
M	1.000	1.000		1.000	.003		1.000	1.000	
ranking=	2	1	3	1	3	2	1	2	3
	location ranking: F > FM > M			location ranking: FM > M > F			location ranking: FM > F > M		

SCATTER adjusted *p*-values	Domain: Size			Domain: Regularity			Domain: Density		
	FM	F	M	FM	F	M	FM	F	M
FM		1.000	1.000		1.000	.003		.003	.001
F	.003		.003	.003		.001	1.000		.001
M	.001	1.000		1.000	1.000		1.000	1.000	
ranking=	3	1	2	3	1	2	1	2	3
	scatter ranking: F > M > FM			scatter ranking: F > FM > M			location ranking: FM > F > M		

4. Discussion and Conclusion

As an extension of the permutation and combination-based testing methodology (Bonnini et al., 2014; Corain and Salmaso, 2015), the main goal of this paper is to propose a multi-aspect testing and ranking method to support shape analysis of cytomorphometric data under multivariate directional alternatives. The main advantage of the proposed multivariate inferential approach is in that it allows to separately analyse in either all-in-one or each brain layer, two distributional aspects, i.e. the mean (inference on location), and the variance (inference on scatter). Distinctly for each domain (size, regularity and density), it turns out that the data analytics we propose is

effective for directionally testing on the possible equality vs. dominance across populations. The results of our approach are multivariate in nature, allowing for a general view on the problem at hand. As about any possible limitation, note that our procedure is implicitly assuming that there is just one type of cell within each region/layer. This assumption may represent an oversimplification under some practical circumstances.

As about the real case study in comparative neuroanatomy, our results revealed that in the molecular layer in females have smaller, more irregular and denser cells than in males, while the freemartins showed an intermediate cell density values between females and males. Interestingly, the Purkinje neurons and the underlying granule cells revealed the same morphological pattern: females possessed larger, more irregular and dense neurons than males. In freemartins, Purkinje neurons showed an intermediate setting between males and females and the granule cells were the largest, most regular and densest. We think that this methodology could be a powerful instrument to carry out cell-oriented morphometric analysis providing robust bases for objective tissue screening, especially in the field of neurodegenerative pathologies where structural differences between genders in the brain cytoarchitecture represent the anatomical substrate underlying the functional differences between the two sexes.

As about some directions for future research, it is well known that resampling-based statistical methods, such as permutation testing and ranking, are quite demanding in computational power and time. In this view, there is a compelling need for trying to optimize computational algorithms to make them more efficient and suitable for practical use. Concerning possible future developments, we aim to apply the solution proposed in the paper by Carvajal-Schiaffino and Firinguetti (2016), where authors discuss a comparison of execution times with parallel implementations in R versus C languages and prove the much better computational performance of C-based parallel programming. The main reason behind this result is in that the compiled version of the algorithm is faster than the interpreted version.

References

1. Arboretti Giancristofaro R., Bonnini S., Corain L., Salmaso L., 2014, A Permutation Approach for Ranking of Multivariate Populations, *Journal of Multivariate Analysis*, 132, pp. 39–57.
2. Bonnini S., Corain L., Marozzi M., Salmaso L. Nonparametric Hypothesis Testing: Rank and Permutation *Methods with Applications* in R. 2014, Wiley: Chichester.
3. Brombin C., Salmaso L. Multi-aspect permutation tests in shape analysis with small sample size. *Computational Statistics & Data Analysis*, 2009, 53, 12, pp. 3921-3931.

4. Carvajal-Schiaffino R., Firinguetti L. A. Bootstrap Confidence Intervals for the Correlation Coefficient: A comparison of Execution Times with Parallel Implementations in R and C languages. *Book of Abstracts of the I Latin American Conference on Statistical Computing*, Gramado, Brazil, July 2016.
5. Corain L, Salmaso L. Improving Power of Multivariate Combination-based Permutation Tests. *Statistics and Computing*, 2015; 25 (2), pp. 203–214.
6. Cozzi B., De Giorgio A., Peruffo A., Montelli S., Panin M., Bombardi C., Grandis A., Pirone A., Zambenedetti P., Corain L., Granato A., 2017, The Laminar Organization of the Motor Cortex in Monodactylous Mammals: a Comparative Assessment Based on Horse, Chimpanzee and Macaque, *Brain Structure and Function*, 222(6), pp. 2743-2757.
7. Graïc J.M., Corain L., Peruffo A., Swaab D.F., 2018 The Bovine Anterior Hypothalamus: Characterization of the Vasopressin-Oxytocin Containing Nucleus and Changes in Relation to Sexual Differentiation, *Journal of Comparative Neurology*, 526, pp. 2898–2917.
8. Grisan E., Graïc J.M., Corain L., Peruffo A., Resolving single cells in heavily clustered nissl-stained images for the analysis of brain cytoarchitecture, 15th IEEE International Symposium on Biomedical Imaging, ISBI 2018, Volume 2018-April, 23 May 2018; 2018, pp. 427-430.
9. Lobo J., See E.Y., Biggs M., Pandit A. An insight into morphometric descriptors of cell shape that pertain to regenerative medicine. *Journal of Tissue Engineering and Regenerative Medicine*, 2016; 10(7), pp. 539-53.
10. Rohlf F.J. Statistical power comparisons among alternative morphometric methods, *American Journal of Physical Anthropology*, 2000, 111, pp. 463-478.
11. Pesarin F., Salmaso L., 2010. *Permutation tests for complex data: theory, applications and software*. Wiley, Chichester.
12. Roy S.N., 1953, On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics*, 24, pp. 220-238.
13. Yanagihara H, Yuan KH. 2005. Three approximate solutions to the multivariate Behrens–Fisher problem. *Communications in Statistics - Simulation and Computation*, 34 (4), pp. 975–988.

Structural breaks in Nonparametric Models via Atomic Pursuit Methods

Matúš Maciak[1], Ivan Mizera[2]
[1]Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Czech Republic
[2]Department Mathematical and Statistical Sciences, University of Alberta, Canada

Abstract
We propose a new innovative approach to changepoint detection and estimation in nonparametric regression: standard methods either require a prior knowledge of the locations of changepoints (discontinuities or higher order structural breaks) or, instead, the estimation process is performed in multiple stages where firstly relevant changepoints are detected and, later, the final model is estimated using the information about the existing locations already gained from the previous stage. This can be, however, limiting in some practical situations. We propose an effective alternative—a fully data-driven all-at-once approach where neither additional prior knowledge for changepoint locations nor their overall number is needed. Our method combines the nonparametric regression estimation with various concepts of sparse atomic pursuit techniques— the L_1 −norm regularization in particular. In addition, various hierarchical forms of structural breaks in the model can be accounted for and the final model can be obtained as a solution to a convex optimization problem. We discuss different model alternatives, and important theoretical properties and changepoint inference tools are derived. A finite sample performance is investigated in terms of an extensive Monte-Carlo simulation study and some practical examples.

Keywords
Nonparametric regression; changepoints; L_1 regularization; LASSO; consistency

1. Introduction

Let $\{(X_i, Y_i);\ i = 1, \dots, N\}$ be a random sample given from some unknown population $\mathcal{F}_{(Y,X)}$. The underlying dependence of Y given X can be expressed as

$$Y_i = m(X_i) + \varepsilon_i,\ for\ i = 1, \dots, N, \tag{1}$$

where ε_i's are assumed to be independent and centered error terms. Such nonparametric setup attracts a lot of interest in statistic: assuming no parametric shape restrictions for the baseline function m turns out to be a very favorable quality. On the other hand, most nonparametric regression methods

still assume some level of smoothness at least. However, relaxing also the smoothness assumption introduces additional flexibility in the model but the classical estimation methods are not applicable in a straightforward way any more.

Nonparametric regression models with changepoints in location were firstly discussed in Müller (1992) and a simultaneous estimation of discontinuities in the function itself the its first derivative was introduced in Loader (1996). Since then, changepoints became very popular and hot topic and huge literature can be found on this matter. Some authors pay attention specifically to changepoint detection problem (see, for instance, Horváth and Kokoszka, 2002; Qiu and Yandell, 1998), others discuss strategies to test for a significance of the changepoint presence (Hušková and Maciak, 2017; Antoch and Jarušková, 2013; Aue et al., 2008; Antoch et al., 2006; Csörgö and Horváth, 1997), but they all rather focus on detecting changepoints then estimating the underlying function itself.

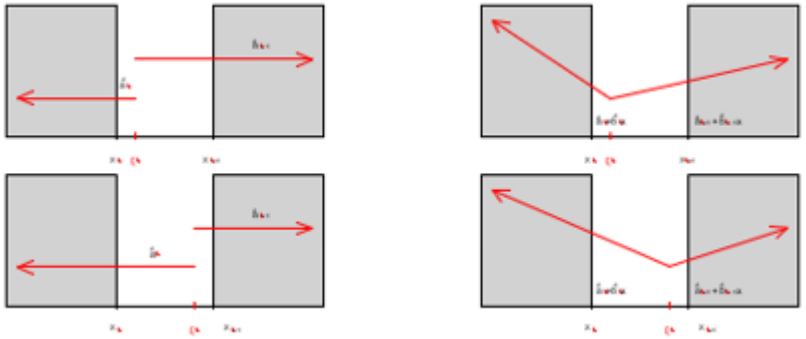
Our approach is motivated by Harchaoui and Lévy-Leduc (2010) and Maciak and Mizera (2016) where piece-wise constant models and continuous piece-wise linear models respectively are estimated using sparsity and the L_1 −norm regularization. Similar approach was recently also considered in Tibshirani (2014) and Sadhanala and Tibshirani (2018) where the authors utilized the L_1 regularization concept for additive models within a trend filtering framework. In addition to their work we deal with a complex nonparametric structure with structural breaks in the underlying function itself and, at the same time, in its derivatives and the overall model is estimated simultaneously. Moreover, using various regularization concepts we can account for an arbitrary changepoint hierarchical structure in the model. Finally, in order to have a complex insight into the data generating model in (1) conditional quantiles can be estimated instead of the conditional mean which only offers a very limited information about the data.

2. Methodology

From the theoretical point of view it is assumed that function m in (1) is sufficiently smooth almost everywhere, but some finitely many points—changepoints (jumps in the function itself or its derivatives). Moreover, no prior knowledge about the locations of changepoints neither their overall number is given. The underlying regression function m from (1) can be decomposed as

$$m(x) = m_0(x) + \sum_{j=0}^{p-1} s_j(x) \tag{2}$$

for all $x \in \mathcal{D}$, where $D \subset \mathbb{R}$ is some compact domain of interest (for instance, interval [0,1]). Function m_0 is a *smooth function* of the order $p \in \mathbb{N}$ at least (for instance, $p = 3$), and functions s_j's are so called *background shock processes* of the lower orders $j \in \{0, \ldots, p-1\}$, having all the same domain as m, or m_0 respectively. In other words, function s_0 is a piece-wise constant function generating jumps in m, function s_1 is a continuous piece-wise linear function generating jumps in the first derivative of m, etc. For every location where j-th order polynomial pieces of s_j functions "join" together, there is a j-th order structural break in the underlying regression function m (a jump in the j-th order derivative). The idea is to develop a direct estimation approach that can fully reconstruct the underlying function m, while also estimating all locations where the changepoints occur (thus, defining also the orders and magnitudes of all jumps, or breaks respectively).



(a) Piece-wise constant trend (b) Piece-wise linear trend

FIGURE 1. An illustration of the optimal changepoint location gained from the underlying data: two piece-wise constant models in (a) are equivalent if there are no other data points available in between observations X_i and X_{i+1}. On the other hand, two models in (b) are clearly different even though there are no more data points between X_i and X_{i+1}. Thus, there is no information about the optimal changepoint location given within the data for a jump in the model but there is sufficient information obtained implicitly in the data for the optimal changepoint location for higher order derivatives.

From the computational point of view, splines are used to estimate the underlying function m and its components—the smooth part m_0 and the shock processes $s_0, \ldots, s_{p-1}$. However, in order to stay within the convex optimization framework, the shock processes are only allowed to be active at the observational points X_i's and, moreover, they are assumed to be active only at some very few points—changepoints. Therefore, the *sparsity principle* is employed when estimating the shock processes to achieve this property. Restricting changepoint occurrences onto the set of observational points

$\{X_i\}_{i=1}^N$ does not play any role when estimating discontinuities in the function itself, however, it effects the situation when estimating jumps in the corresponding derivatives (see Figure 1 for an illustration).

Let's assume that $m_0(x) = \sum_{\ell=1}^K \beta_\ell h_\ell(x)$, where $\{h_\ell\}_{\ell=1}^K$ is the set of standard B-spline basis functions of the order $p \in \mathbb{N}$ and $\boldsymbol{\beta}_S = (\beta_1, \ldots, \beta_K)^T$ is the vector of unknown parameters which needs to be estimated to obtain the estimate of m_0. Similarly, for the shock processes$\{ s_0, \ldots, s_{p-1}\}$ we define $s_j(x) = \sum_{i=1}^N \gamma_i^{(j)}(x - X_i)_+^j$, for $j = 0, \ldots, p-1$, where $(z)_+$ denotes the positive part of $z \in \mathbb{R}$. It is easy to see that the set of functions $\{h_i^{(j)}(x)\}_{i=1}^N = \{(x - X_i)_+^j\}_{i=1}^N$ defines a truncated power spline basis of order $j \in \{0, \ldots, p-1\}$ over a set of knot points$\{X_i\}_{i=1}^N$, which are, without any loss of generality, assumed to be unique.

Finally, the *sparsity principle* is introduced with respect to $\gamma_i^{(j)}$ parameters, such that $\gamma_i^{(j)} = 0$ holds for almost all indexes $i = 1, \ldots, N$ and $j = 0, \ldots, p-1$ (thus, there is no structural break at the location X_i), but some few exceptions—existing changepoints. Moreover, for some identifiability reasons, it is also assumed that all shock processes are inactive at the beginning (for instance, $s_j(X_1) = 0$, and thus, $\gamma_1^{(j)} = 0$, for all $j = 0, \ldots, p-1$).

The estimate for the underlying regression function m in (1) can be now obtained as a solution of the minimization problem

$$(3) \qquad \underset{(\beta_S^\intercal, \gamma_0^\intercal, \ldots, \gamma_{p-1}^\intercal) \in \mathbb{R}^d}{\text{Minimize}} \left\| \boldsymbol{Y} - \left(\boldsymbol{B}_S \boldsymbol{\beta}_S + \sum_{j=0}^{p-1} \boldsymbol{B}_J^{(j)} \boldsymbol{\gamma}_i \right) \right\|^2 + \lambda_S \left\| \boldsymbol{P} \boldsymbol{\beta}_S \right\|_2^2 + \lambda_J \mathcal{P}_{L_1}(\boldsymbol{\gamma}_0, \ldots, \boldsymbol{\gamma}_{p-1})$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^\mathsf{T}$ is the response vector, $\beta_S = (\beta_1, \ldots, \beta_K)^\mathsf{T}$, and $\boldsymbol{\gamma}_j = (\gamma_1^{(j)}, \ldots, \gamma_N^{(j)})^\top$, for $j = 0, \ldots, p-1$ are the unknown vectors of parameters to be estimated Parameters $\lambda_S > 0$ and $\lambda_J > 0$ are some tuning parameters, which controls for the overall amount of smoothness in the final estimate (parameter λ_S) and the amount of sparsity in the parameter vectors $\boldsymbol{\gamma}_j = (\gamma_1^{(j)}, \ldots, \gamma_N^{(j)})^\top$, for $j = 0, \ldots, p-1$ (tuning parameter λ_J) which are included in the L_1 type penalty $\mathcal{P}_{L_1}(\boldsymbol{\gamma}_0, \ldots, \boldsymbol{\gamma}_{p-1})$. The penalty can take various forms (see below). Matrices $B_S = \left(h_\ell(X_i)\right)_{i,\ell=1}^{N,K}$ and $B_J^{(j)} = \left(h_\ell^{(j)}(X_i)\right)_{i,\ell=1}^{N,N}$ stand for a classical smoothing spline design matrix and a j-order jump generating bases with sparse vectors of parameters $\boldsymbol{\gamma}_j = (\gamma_1^{(j)}, \ldots, \gamma_N^{(j)})^\top$, for $j = 0, \ldots, p-1$. The L_2 – norm penalty in (3) controls for the overall smoothness of the estimate of m_0 and $\boldsymbol{P}$ is such that $\boldsymbol{P}^\mathsf{T}\boldsymbol{P} = \boldsymbol{W}$, where $\boldsymbol{W}$ is a matrix of mutual products of second derivatives of the basis functions$\{h_\ell\}_{\ell=1}^K$.

In general, for $\lambda_J \to 0$ we expect changepoints to occur in each X_i observation and every level of smoothness while for $\lambda_J \to \infty$ no changepoints are expected and the final fit is fully determined by the vector of parameters $\beta_S \in \mathbb{R}^K$ (the smooth function m_0).

The minimization problem (3) is a convex problem and it can be solved using standard optimization tools. The smoothing parameter $\lambda_S > 0$ and the sparsity parameter $\lambda_J > 0$ can be selected, for instance, by some Cross-Validation technique. Alternatively, one can compute the whole solution paths for$\lambda_J > 0$ using LARS algorithm (Efron at al., 2004) and to choose the final model from a set of plausible models along the whole solution path.

2.1 Independent Changepoints: The L_1 type penalty term $\mathcal{P}_{L_1}(\gamma_0, \ldots, \gamma_{p-1})$. in (3) can take various forms. Let us firstly mention the simplest scenario where there is no hierarchical restriction imposed on the changepoint occurrences: any discontinuity point in the function itself or its derivatives can occur on its own. This property can be expressed by a specific penalty form, where

$$\mathcal{P}_{L_1}(\gamma_0, \ldots, \gamma_{p-1}) = \sum_{j=0}^{p-1} \sum_{i=1}^{N} \left|\gamma_i^{(j)}\right| \tag{4}$$

Alternatively, one can consider a whole set of regularization parameters $\lambda_J = \left(\lambda_{J0}, \ldots, \lambda_{J(p-1)}\right)^{\mathsf{T}}$

$\lambda_J = (\lambda_{J0},\ldots,\lambda_{J(p-1)})^{\top}$ to control for the sparsity in each smoothness level $j \in \{0, \ldots, p-1\}$ separately.

2.2 Simultaneous Changepoints: Unlike the previous situation it can be suitable for some scenarios to link the changepoint at some location across all different levels of $j \in \{0, \ldots, p-1\}$. The motivation comes from some practical examples where the shock processes in (2) are expected to become all active at the same point. This quality can implemented into (3) by replacing the standard LASSO penalty in (4) with the group LASSO penalty

$$\mathcal{P}_{L_1}(\gamma_0, \ldots, \gamma_{p-1}) = \sum_{i=1}^{N} \sqrt{\left(\gamma_i^{(0)}\right)^2 + \cdots + \left(\gamma_i^{(p-1)}\right)^2} \tag{5}$$

which either selects the whole group of parameters $\gamma_i^{(0)}, \ldots, \gamma_i^{(p-1)}$ for some $i \in \{1, \ldots, N\}$ to be nonzero, or all parameters within this group are set to zero exactly.

2.3 Hierarchical Changepoints: An innovative approach to changepoints in the nonparametric regression models can be obtained by using the overlap

group LASSO proposed in Jacob et al. (2009) which allows to implement any arbitrary hierarchical structure into the model. The idea is to replace $\sqrt{\left(\gamma_i^{(0)}\right)^2 + \cdots + \left(\gamma_i^{(p-1)}\right)^2}$ in the sum in (5) by its decomposition into latent parameters, such that $\sqrt{\left(\gamma_i^{(0)}\right)^2 + \cdots + \left(\gamma_i^{(p-1)}\right)^2} = \sqrt{\left(\gamma_{i(1)}^{(0)}\right)^2 + \cdots + \left(\gamma_{i(1)}^{(p-1)}\right)^2} + \cdots + \sqrt{\left(\gamma_{i(p-1)}^{(0)}\right)^2 + \cdots + \left(\gamma_{i(p-1)}^{(p-1)}\right)^2}$ and with some well defined restrictions of the form $\gamma_{i(k)}^{(j)} = 0$, for some $j, k \in \{0, \ldots, p-1\}$ and all $i = 1, \ldots, N$ one can enforce the required form of the changepoint hierarchy in the model (for instance, if there is a jump in the *j*-th order derivative of m, changepoints will also occur in higher order derivatives but not in the lower order derivatives).

3. Results

The nonparametric regression models with changepoints being detected and estimated by using the L_1 -norm regularization approaches are investigated for various L_1-type penalty forms. Theoretical results are derived with respect to the quality of the final estimate and also with respect to the quality of the changepoint detection performance. To be specific:

- under some necessary regularity assumptions, some technical conditions, and some minor changepoint restrictions, the consistency of the model estimation is proved such that
$$\|\mathbb{X}\widehat{\boldsymbol{\beta}}(\lambda_N) - \mathbb{X}\boldsymbol{\beta}\|_2^2 \leq C_N \cdot \sqrt{\frac{\log N}{N}},$$
for a well defined constant C_N and the vector of unknown param $\boldsymbol{\beta} = (\boldsymbol{\beta}_S^\top, \boldsymbol{\gamma}_0^\top, \ldots, \boldsymbol{\gamma}_{p-1}^\top)^\top$;
- under some necessary regularity assumptions and some technical conditions on the number of estimated changepoints the consistency of the detection is proved such that
$$P\left(\max_{1 \leq k \leq K} |\hat{t}_k - t_k^*| \leq N\delta_N\right) \longrightarrow 1, \text{ for } N \to \infty,$$
for some well defined non-increasing positive sequence $\delta_N > 0$, where $K \in \mathbb{N}$ is the total number of true changepoint locations t_k^* in the model with their corresponding estimates $\hat{t}_k$;
- under some necessary regularity assumptions and some technical conditions it is proved that the proposed methodology recovers all existing changepoints with probability tending to one as sample size increases (in a sense that the number of estimated locations is at least K);
- the consistency of the model estimation and changepoint detection is also proved (in an analogous sense as above) for estimating the conditional quantiles thus, for the L_2-norm objective function in (3) being replaced with the quantile check function $\rho_\tau(u) = u(\tau - \mathbb{I}_{\{u<0\}})$, for some $\tau \in (0,1)$;

- the robustness property of the conditional quantile estimation is proved with respect to outlying observations and heavy tailed distributions (for instance, Cauchy distribution);

Due to the strict page limitation all theoretical assumptions, technical details and complete proofs can be found in Maciak and Mizera (2016), Ciuperca and Maciak (2018) and Maciak and Mizera (2019).

4. Discussion and Conclusion

Similarly as standard LASSO approaches the proposed methodology does not yield the *oracle properties*. Indeed, different values of the regularization parameter $\lambda_J > 0$ in (3) are needed to achieve the consistency of the changepoint detection or the consistency of the model estimation. From the empirical point of view, various techniques can be used to improve the finite sample properties especially for small sample sizes (for instance, de-biasing, relaxed LASSO, or two-stage estimation). However, the proposed methodology can effectively deal with changepoint detection and estimation problem within nonparametric regression setups and it is especially suitable for situations where no prior knowledge on any changepoint structure is known in advance. An extensive Monte Carlo simulation study is proposed to closely investigate finite sample properties and various model selection options. The overall suitability of the proposed regularized changepoint detection in nonparametric models is also illustrated on some practical examples.

References

1. Antoch, J., Gregoire, G., and Huˇskov´a M. (2006), "Test for Continuity of Regression Function." *Journal for Statistical Planning and Inference* 137, 753 – 777.
2. Antoch, J. and Jaruˇskov´a, D. (2013), "Testing for Multiple Change-points." *Journal of Computational Statistics* 28, 2161 – 2183.
3. Aue, A., Hotv´ath, L., Huˇskov´a M. and Kokoszka, P. (2008), "Testing for Changes in Polynomial Regression." *Bernoulli*, Volume 13, No.3, 637 – 660.
4. Ciuperca, G. and Maciak, M. (2018), "Change-point detection in a linear model by adaptive fused quantile method." *Scandinavian Journal of Statistic*, (submitted).
5. Cs¨org¨o, M., and Hotv´ath, L. (1997), "Limit Theorems in Change-Point Analysis." *Wiley Series in Probability & Statistics*, Chichester, England.
6. Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), "Least Angle Regression." *The Annals of Statistics*, 32, No.2, 407 – 499.

7. Harchaoui, Z. and L´evy-Leduc, C. (2010), "Multiple Change-Point Estimation With a Total Variation Penalty." *Journal of the American Statistical Association*, 105, No.492, 1480 – 1493.
8. Hotv´ath, L. and Kokoszka, P. (2002), "Change-Point Detection With Non-Parametric Regression." *Statistics* 36, No.1, 9-31(23).
9. Hu˘skov´a, M. and Maciak, M. (2017), "Discontinuities in Robust Nonparametric Regression with A-mixing Dependence." *Journal of Nonparametric Statistics* 29, No.2, 447-475.
10. Jacob, L., Obozinski, G. and Vert, J.P. (2009), "Group Lasso with Overlap and Graph Lasso." *Proceedings of he 26th International Conference on Machine Learning (ICML 26)*, Montreal, Canada.
11. Loader, C. (1996), "Change Point Estimation Using Nonparametric Regression." *Annals of Statistics* 24, 1667– 1678.
12. Maciak, M. and Mizera, I. (2016), "Regularization Techniques in Joinpoint Regression." *Statistical Papers*, 1-17.
13. Maciak, M. and Mizera, I. (2019), "Splines with Changepoints: Additive Models for Functional Data." (*to be submitted*).
14. M¨uller, H. (1992), "Change Points in Nonparametric Regression Analysis." *Annals of Statistics* 20, 737 – 715.
15. Qiu, P. and Yandell, B. (1998), "A Local Polynomial Jump Detection Algorithm in Nonparametric Regression." *Technometrics* 40, 141 – 152.
16. Sadhanala, V. and Tibshirani, R. (2018), "Additive Models with Trend Filtering." arXiv:1702.05037, 1–63.
17. Tibshirani, R. (2014), "Adaptive Piece-wise Polynomial Estimation via Trend Filtering." *The Annals of Statistics*, 42(1), 285–323.

Investigating dissimilarity in spatial area data using Bayesian Inference: The case of voter participation in the Philippine National and Local Elections of 2016

Francisco N. de los Reyes
School of Statistics, University of the Philippines Diliman Quezon City Philippines

Abstract
A commonly studied characteristic of area data is the assessment of similarity (or absence thereof) among neighboring areal units. However, most methodologies do not measure uncertainties which are likely outcomes of sampling variation and do not consider spatial autocorrelation. This paper explores the ability of Bayesian modeling to address the said situations. It attempts to apply this modeling technique to the voting participation statistics in the Philippine National and Local Elections of 2016.

Keywords
conditional autoregressive (CAR); proximity matrix; dissimilarity; voter turnout

1. Introduction

Many inquiries in statistics are interested in determining heterogeneity in some population. Dissimilarity is one such measure. It is the extent to which two or more groups are integrated or isolated. The most popular metric is the Dissimilarity Index. However, the Dissimilarity Index has the following inadequacies in spatial data: it does not measure uncertainties which could potentially be a result of random sampling variation, and it does not consider spatial autocorrelation which could be present in the data.

This paper aims at detecting dissimilarity in a specific spatial area data: voter participation. In the Philippines, voter turnout is intuitively spatially autocorrelated. There are strong bailiwicks in various corridors in Philippine geography: Northern Luzon is one, Bicol region is another. There is also a strong solid vote in Panay and Negros Islands, another in Cebu and then the Davao region. Voter turnout in nearby barangays (the Philippine basic geopolitical unit) tends to be similar. The same may be opined for larger units like cities and municipalities and even up to the level of the province or region. This paper shall first present a classical method in establishing dissimilarity. However, in consideration of the spatial nature of voter turnout, a Bayesian model will be used to introduce smoothing in the presence of spatial autocorrelation.

2. Methodology

The Dissimilarity Index. Let the spatial areal data be denoted by $\underline{Y} = (Y_1, \ldots, Y_n)$ and $\underline{N} = (N_1, \ldots, N_n)$, which respectively denote the number of people who voted and the number of registered voters for each of the n areal units. Here, the areal units are the provinces, both regular and special provinces as determined by the Commission on Elections (COMELEC), as well as the districts in the National Capital Region. Define voter turnout as the proportion of registered voters who actually voted. Let $\underline{p} = (p_1, \ldots, p_n)$ denote the true voter turnout in each areal unit. The Dissimilarity index is given by Lee et al (2015) as

$$D = \sum_{k=1}^{n} \frac{N_k |p_k - p|}{2N_p(1-p)}$$

where $N = \sum_{k=1}^{n} N_k$ and p are the total population of registered voters and overall voter turnout in 2016 for the entire Philippines. The value of D lies in the interval [0, 1], where 0 conveys parity and 1 means full disparity (or segregation). The unknown true proportions p are typically estimated by their sample equivalents, that is $\hat{p}_k = Y_k/N_k$ and $\hat{p} = (\sum_{k=1}^{n} Y_k)/\sum_{k=1}^{n} N_k$. Sampling variation is clearly present if (Y_k, N_k) emanates from a survey, since they are based on a random sample in areal unit k. For the elections data, variation may be alluded to measurement errors due to misreporting, misrecording or computation as in the case of manual tallying.

Bayesian Modelling. The estimator $\hat{p}_k$ is both the method of moments estimator and the maximum likelihood estimator under the model $Y_k \sim Binomial(N_k, p_k)$. However, this model assumes that data among areal units are independent, something which is not valid in the presence of spatial autocorrelation. To accommodate this dependence, a Conditional Autoregressive (CAR) model will be used to model the spatial autocorrelation in the data. In this study, the methodology proposed by Lee, Minton and Pryce (2015) was followed. Lee, et al. proposed a global smoothing model for spatially autocorrelated data using a binomial generalized linear mixed model (GLMM), where the random effects are spatially autocorrelated. The full model is given by Lee et al (2015) as follows:

$$Y_k \sim Binomial(N_k, p_k)$$

$$ln\left(\frac{p_k}{1-p_k}\right) = \beta_0 + \varphi_k;\ \underline{\varphi} \sim N(\underline{0},\ \tau^2 Q(\rho, W)^{-1})$$

$$\beta_o \sim N(0, C), C\ constant$$

$$\tau^2 \sim Inverse\ Gamma(a, b)$$

$$\rho \sim Uniform(0,1)$$

The random effects $\underline{\varphi} = (\varphi_1, \ldots, \varphi_n)$ shall account for the spatial dependence in the data, and are represented by a CAR prior distribution. Moran's Index was used to confirm if spatial autocorrelation exists. The CAR priors shall induce the spatial autocorrelation by a binary $n \times n$ proximity matrix $W = (w_{ki})$, which is computed from the contiguity structure of the n areal units. Based on W, the CAR priors take the form of a zero-mean multivariate Gaussian distribution, where spatial autocorrelation is induced via the precision matrix that depends on W. Leroux et al. (1999) proposed that the strength of the autocorrelation be estimated from the data. The precision matrix for this model involves an autocorrelation parameter and the proximity matrix and is given by

$$Q(\rho, W) = \rho(diag(W\,\underline{1}) - W) + (1 - \rho)I,$$

where I is an $n \times n$ identity matrix, $\underline{1}$ is an $n \times 1$ vector of ones, and $diag(W\underline{1})$ is a diagonal matrix with elements equal to the row sums of W. The matrix $Q(\rho, W) = \rho(diag(W\underline{1}) - W) + (1 - \rho)I$ is proper if $\rho \in [0, 1)$, and the spatial structure amongst φ can be observed more clearly from the univariate full conditional distributions

$$\varphi_k | \underline{\varphi}_{-k} \sim Normal\left(\frac{\rho \sum_{i=1}^{n} w_{ki}\varphi_i}{\rho \sum_{i=1}^{n} w_{ki} + 1 - p}, \frac{\tau^2}{\rho \sum_{i=1}^{n} w_{ki} + 1 - p}\right)$$

where $\underline{\varphi} - k$ denotes the vector of random effects except for φ_k. The parameter ρ controls the spatial autocorrelation structure, with $\rho = 1$ corresponding to strong spatial autocorrelation, while $\rho = 0$ corresponds to independent random effects according to Besag et al (1991). The effects have constant mean and have constant variance. Weakly informative prior distributions are assigned to the other hyperparameters so as to allow for estimates of these parameters to be determined from the observed data and not to skew the analysis. The intercept coefficient in the logit function was assigned a univariate Normal distribution with mean zero and homoskedastic. The hyperparameter τ^2, accounting for the variation in the spatial effects are assigned the inverse-gamma. The spatial autoregression parameter is assumed to be uniform over (0,1) since it is over this support that the precision matrix is also deemed proper. According to Lee et al (2015), the posterior distribution for the dissimilarity index D can be computed using M Markov Chain Monte Carlo (MCMC) samples from the posterior distribution

$$\left\{\boldsymbol{\Theta}^{(j)}\right\}_{j=1}^{M} where\ \boldsymbol{\Theta}^{(j)} = \left(\phi^{(j)}, \beta_0^{(j)}, \tau^{2(j)}, \rho^{(j)}\right)$$

In the analysis, three values of M were used: twenty thousand, thirty thousand and forty thousand all with 50% burn-in. The posterior samples are then used to construct samples $\boldsymbol{p}^{(j)} = (p_1^{(j)}, \dots, p_n^{(j)})$, using the inverse logit transform

$$p_k^{(j)} = exp(\beta_0^{(j)} + \varphi_k^{(j)}) / [1 + exp\left(\beta_0^{(j)} + \varphi_k^{(j)}\right)].$$

The jth sample from the posterior distribution of D is constructed as

$$D^{(j)} = \sum_{k=1}^{n} \frac{N_k |p_k^{(j)} - p^{(j)}|}{2Np^{(j)}(1-p^{(j)})}, j = 1, \dots, M \text{ where } p^{(j)} = (\sum_{k=1}^{n} N_k\, p_k^{(j)})/(\sum_{k=1}^{n} N_k).$$

Finally, D can be estimated by the median of $\{D^{(1)}, \dots, D^{(M)}\}$, while a 95% credible interval is obtained from the 2.5th and 97.5th quantiles of $\{D^{(1)}, \dots, D^{(M)}\}$.

Voter Participation in the Philippine National and Local Elections (NLE) of 2016, official data from the Commission on Elections was used in the research. Since voter turnout is viewed here in a spatial data analysis paradigm, and therefore contiguity-sensitive, turnout from overseas voting was not included. Provincial level information on number of registered voters and actual voter turnout for 86 areal units comprise the entirety of the dataset. This includes special readings for the cities of Isabela and Cotabato, which are labeled special provinces, and the four districts of Manila. Proximity due to common-border cannot be used since there are 15 provinces which islands. These are Batanes, Biliran, Bohol, Camiguin, Catanduanes, Cebu, Dinagat, Guimaras, Marinduque, Masbate, Palawan, Romblon, Siquijor, Sulu and Tawi-Tawi. Here, connectivity was based on a nominated critical distance. Special consideration arose for the island of Palawan since a large critical distance was needed for it to have just one neighbor. Thus for this specific province, indication of geographic integration like presence of boat routes and trade with a nearby province was used. This led to Iloilo being set as Palawan's neighbor. The proximity matrix was then revised to force a neighbor for Palawan. Areal centroids were identified via the Universal Transverse Mercator (UTM) coordinate system. Inter-unit distance was computed via these coordinates. Created a proximity matrix W based on L_1 distance of at most a nominated δ.

$$W = \{w_{ik}\} \text{ has } w_{ik} = \begin{cases} 1, & d_{L_1}(A_i, A_j) \leq \delta \\ 0, & d_{L_1}(A_i, A_j) \leq \delta \end{cases} \quad ; \quad w_{ii} = 0$$

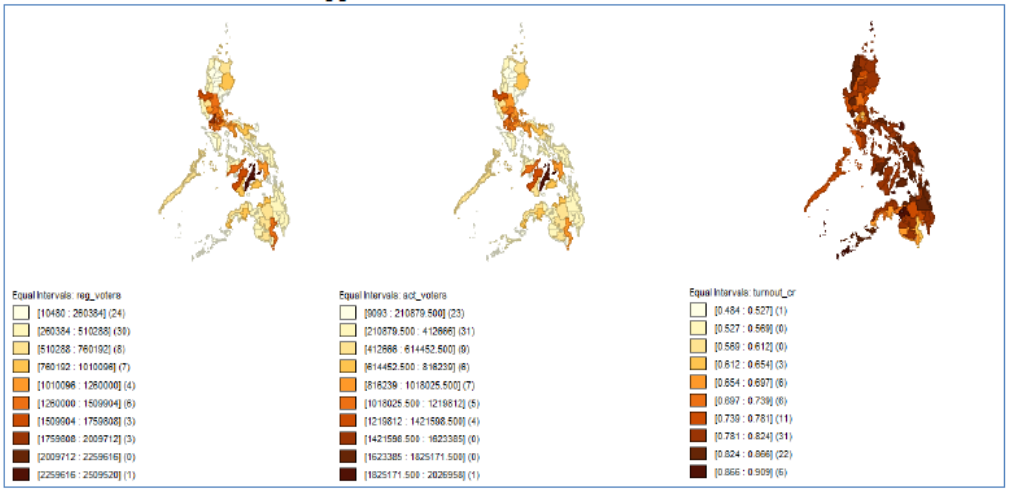
Several values were tried for δ but only in $\delta = 500$ UTM units did all areal units, except Palawan, had at least one neighbor. Iloilo was forced to become Palawan's neighbor by virtue of transportation and trade relations. A distance

decay term was introduced to the spatial variance matrix in accordance to Tobler's principle. The term is $g(d_{ij}) = e^{-3d_{ij}}$ where d_{ij} is the inter-centroid distance between areal units i and j. A piecewise mean function was also generated for two clusters: voter turnout > 80% where the mean turnout is 83% and voter turnout is 73%. The logit transform of the proportions are generated from a multivariate Gaussian distribution with a mean of 0.99 (logit corresponding to the mean turnout of 73%) or 1.73 (logit corresponding to the mean turnout of 83%). Low, moderate and high (on account of Cotabato City) spatial variation scenarios were investigated. The Bayesian model was fitted at M = 20,000 , 30,000 and 40,000 MCMC samples. Parameter estimation proceeded after 50% burn-in. Data integration, computation of proximity matrix and testing for spatial autocorrelation were done in Geoda. Modelling was done in R CARBayes package with extensive use of the **S.CARleroux** function.

3. Results

Voter turnout was generally high in the NLE of 2016 as indicated by an average of 79% across the 86 areal units (Figure 1). The special province of Cotabato City was outlying with a turnout of only 48%. There was a significant positive spatial autocorrelation in voter turnout (Moran's I = 0.22, p = 0.0067). This indicates that areas with relatively higher voter turnout are spatially close. A similar conclusion can be said of areas with relatively lower voter turnout. There is suggestion of parity in voter participation across provinces as evidenced by a dissimilarity index of D=0.15. The 95% confidence interval is (0.113, 0.175) based on 10,000 bootstrap samples.

Figure 1. Registered Voters, Actual Voters and Voter Turnout in the Philippine National and Local Elections of 2016



Voter turnout was generally high in the NLE of 2016 as indicated by an average of 79% across the 86 areal units (Figure 1). The special province of Cotabato City was outlying with a turnout of only 48%. There was a significant positive spatial autocorrelation in voter turnout (Moran's I = 0.22, p = 0.0067). This indicates that areas with relatively higher voter turnout are spatially close. A similar conclusion can be said of areas with relatively lower voter turnout. There is suggestion of parity in voter participation across provinces as evidenced by a dissimilarity index of D=0.15. The 95% confidence interval is (0.113, 0.175) based on 10,000 bootstrap samples.

The hierarchical model had poor fit under the assumption of high spatial variation scenario within clusters given the official election voter turnout data (Table 1). Here, the standard deviation of residuals and Deviance Information Criterion (DIC) are highest within tiers of MCMC sample. The width of the 95% credible interval for the dissimilarity index, intercept term for the logit expression, τ^2 and ρ are generally largest.

The hierarchical model had relatively better fit under the assumption of moderate spatial variation scenario as compared to the model given a high spatial variation assumption. In this scenario, the standard deviation of residuals and Deviance Information Criterion (DIC) are lower within tiers of MCMC sample. The width of the 95% credible interval for the dissimilarity index, intercept term for the logit expression, τ^2 and ρ are tend to be narrower. The hierarchical model showed best fit under low spatial variation scenario within clusters given the official election turnout data. The 95% confidence intervals are narrowest within each group of MCMC samples. Residual variability is at its least and so is the DIC. Estimates seem to have good precision at M=30,000 MCMC samples (50% burn-in). Here, the spatial autocorrelation parameter can reasonably be expected to fall in the interval (0.001, 0.181).

The dissimilarity indices generated across all scenarios are relatively small. These values signify that variability in voter participation in indeed small across provinces. There is generally high turnout nationwide with provincial rates which are not far from this general average.

Table 1. Results of Bayes Modelling of Voter Turnout in the Philippine National and Local Elections of 2016

POSTERIOR QUANTITIES AND MODEL FIT												
	Scenario A: Low spatial variation											
	M=20,000 (50% Burn In)				M=30,000 (50% Burn in)				M=40,000 (50% Burn in)			
	Median	L95	U95	width	Median	L95	U95	width	Median	L95	U95	width
Dissimilarity, D	0.103	0.097	0.121	0.024	0.141	0.137	0.142	0.005	0.120	0.119	0.120	0.001
Intercept	1.097	1.095	1.109	0.014	1.099	1.098	1.101	0.003	1.114	1.113	1.115	0.002
tau-square	0.191	0.113	0.619	0.506	0.221	0.118	0.875	0.757	0.194	0.114	0.874	0.760
rho	0.014	0.000	0.109	0.109	0.022	0.001	0.181	0.180	0.015	0.001	0.171	0.170
SD residuals	184.67				168.22				181.97			
DIC	3011698				2515216				2914018			
	Scenario B: Moderate spatial variation											
	M=20,000 (50% Burn In)				M=30,000 (50% Burn in)				M=40,000 (50% Burn in)			
	Median	L95	U95	width	Median	L95	U95	width	Median	L95	U95	width
Dissimilarity, D	0.098	0.097	0.099	0.002	0.092	0.092	0.098	0.006	0.103	0.102	0.103	0.001
Intercept	1.292	1.291	1.293	0.002	1.286	1.283	1.287	0.004	1.305	1.304	1.306	0.002
tau-square	0.250	0.139	1.424	1.285	0.223	0.124	1.069	0.945	0.253	0.143	1.165	1.022
rho	0.017	0.001	0.240	0.239	0.019	0.000	0.205	0.205	0.018	0.001	0.184	0.183
SD residuals	222.96				204.17				216.36			
DIC	4313926				3629979				4053158			
	Scenario C: High spatial variation											
	M=20,000 (50% Burn In)				M=30,000 (50% Burn in)				M=40,000 (50% Burn in)			
	Median	L95	U95	width	Median	L95	U95	width	Median	L95	U95	width
Dissimilarity, D	0.223	0.222	0.228	0.006	0.207	0.206	0.208	0.002	0.209	0.203	0.213	0.010
Intercept	1.210	1.208	1.212	0.004	1.223	1.214	1.225	0.011	1.217	1.215	1.218	0.003
tau-square	1.406	0.449	7.197	6.748	1.784	0.545	6.326	5.781	0.864	0.327	4.425	4.098
rho	0.113	0.014	0.743	0.729	0.169	0.026	0.669	0.643	0.063	0.005	0.452	0.447
SD residuals	238.28				229.23				219.91			
DIC	4884536				4514759				4165975			

4. Conclusions, Recommendations and Learning's

Both Bayesian and non-Bayesian models revealed that there is low dissimilarity in voter turnout among the 86 areal units contained in the official Comelec dataset. When spatial variation is taken into account, there is sufficient basis to say that the spatial variation is low. Thus, it is clear that Filipinos participated well in the National and Local Elections of 2016 and quite consistently homogeneous in pattern if taken spatially. As to the statistical specification of the model, the case of the Philippines requires critical distance of 500,000 UTM units to assure that areal units have at least one neighbor based on inter-centroid. A proximity matrix can still be constructed for a critical distance lower than this value but the algorithm fails to converge due to provinces without neighbors. For the case of Palawan, one needs to override the generated proximity matrix to force a neighbor under some special criterion (here, transportation and trade relation). Localized smoothing is beyond the scope of this study and is a welcome improvement moving forward. Moving forward, the dissimilarity index in voter turnout should be tracked over time to gather insight if the Philippine electorate is indeed participative. The technique presented here may be also be applied to other areal information with inherent spatial variation like poverty and health statistics where strong spatial components are expectedly inherent.

References

1. Bailey, Trevor C. and Gatrell, Anthony C. (1995). Interactive Spatial Data Analysis. New York. Longman Group Limited.
2. Besag,J.,York,J.,Mollie,A. (1991). Bayesian image restoration with two applications in spatial statistics. Ann.Inst.Statist.Math. 43,1–59.
3. Bivand, Roger S., Pebesma, Edzer J. and Gomez-Rubio, Virgilio (2103). Applied Spatial Data Analysis with R. New York. Springer.
4. Cressie, Noel, A. (1993). Statistics for Spatial Data. New York. John Wiley & Sons, Inc.
5. Duncan Lee, Jon Minton, Gwilym Pryce (2015). Bayesian inference for the dissimilarity index in the presence of spatial autocorrelation. Spatial Statistics 11 (2015) 81–95.
6. Leroux, B., Lei, X., Breslow, N. (1999). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Halloran, M., Berry, D. (Eds.), Statistical Models in Epidemiology, the Environment and Clinical Trials. New York. Springer-Verlag.
7. R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org.

Traditional and newly emerging data quality problems in countries with functioning vital statistics: experience of the human mortality database

Dmitri Jdanov[1,2], Domantas Jasilionis[1]
[1]Max Planck Institute for Demographic Research, Rostock, Germany
[2]National Research University Higher School of Economics, Moscow, Russia

Abstract
The Human Mortality Database (HMD, www.mortality.org) is the world´s leading data resource on mortality in developed countries. This paper summarizes the experience of the HMD project in the area of assessment of mortality and population data for countries with functioning vital statistics. Data and methodological challenges and their solutions are illustrated using empirical examples of country-specific cases. Several methodological approaches allowing enhancing the utility of the population-level mortality data even if data quality is problematic are discussed. Unfortunately, standard direct and indirect demographic estimation methods which are widely applied for the data from developing countries or historical data series often do not work in solving data-related problems for developed countries with functioning statistical systems. The HMD proposes a comprehensive approach for these cases which is based on combining the application of advanced demographic and statistical methods and extensive usage of additional or alternative data sources. The major challenge for this approach is to follow the fundamental principle of the HMD methodology ensuring comparability of mortality data series across time and space.

Keywords
data quality; mortality; databases; vital statistics; population estimates

1. Introduction

The Human Mortality Database (HMD, www.mortality.org) is the world´s leading data resource on mortality in developed countries. This unique open-access collection provides detailed mortality and population data for 38 countries with relatively complete and reliable vital registration and census data. It is a collaborative project by the Department of Demography at the University of California at Berkeley (USA) and the Max Planck Institute for Demographic Research in Rostock (Germany). The main goal of the HMD is to document the longevity revolution of the modern era and to facilitate research into its causes and consequences by providing high quality data to researchers, students, journalists, policy analysts, and others interested in the history of human longevity.

All data series in the HMD are updated on a rolling basis. One of the main principles of the HMD is to include countries with reliable population statistics, especially requiring a full coverage of registration of vital events. The countries and areas included thus are relatively wealthy and for the most part highly industrialized. The Human Mortality Database contains original calculations of death rates and life tables for national populations (countries or areas), as well as the original input data used for constructing those tables and an extensive documentation. More details about the HMD project can be found in (Barbieri et al., 2015). The HMD has more than 45,000 registered users and has been cited in more than 1500 scientific publications. The average number of citations during the last two years is 200 per year.

One of the main advantages of the HMD is quality of provided data and their comparability across time and space. The constructed and updated series are carefully checked and reviewed for internal and external consistency before publication. Every data series in the HMD have to meet strong requirements. One of the most important criteria is vital registration system which cover more than 90 percent of population.

Principles, recommendations and guidelines for evaluation of quality of vital registration systems and vital data were discussed many times (Setel et al., 2007; UN, 2014; WHO, 2010). The HMD relies on external researches by preliminary selection of acceptable countries. Unfortunately, for most of the world population, complete and accurate data on mortality are not available. To produce such data an expensive and well-organized system for registration of vital events and also censuses or population registers to count population is needed. This is something that majority of developing nations have been unable to achieve. Death registration does not exist or is very fragmentary in most of the developing world including its most populated parts (China, India, Indonesia) and also in countries that are facing the greatest health challenges (Sub-Saharan Africa).

Vital registration that can be used to calculate life tables over the whole range of ages exist in about 60 countries. But existing of nearly complete vital registration system does not guarantee that quality of population statistics is sufficient. In about 15 to 20 of these countries, quality of these data is a serious concern. For other 40-45 countries, data quality can still be problematic during some time periods or at some ages. Below we summarize most important aspects of the work on evaluation of data quality within the HMD project. We do not touch problems related to some of historical populations' experience, quality of population censuses or population statistics in developing countries. We focus on the newly emerging data quality problems. These problems are largely related to growing uncertainty about the population denominator due to unregistered migration.

2. Inter-Censal Estimates and Migration

Contrary to common beliefs, international migration has been low for decades. In 2015, about 3.3% of the world's population, or 244 million people, lived in a country other than the country of birth (Willekens et al., 2016).

However, the distribution of migration is often highly uneven. First, it affects working ages and have significant influence on fertility estimates. For countries with a long history of erroneous statistics of migrants it also has serious consequences for estimation of mortality at old ages because the proportion of underestimated or overestimated number of migrants becomes higher with age. Second, uneven distribution across countries might be a serious problem for countries with substantial in- or out-migration and small populations.

Population censuses provide the most reliable basis for retrospective estimation of both population estimates and international migration. In many cases, current annual migration statistics are unreliable due to a lack of accurate flow data. In general, arrival data are more reliable than information on out-migration. The last round of censuses in 2010-2012 allowed to produce reliable (at least in most of the cases) population estimates around the time of the census. In many countries the new census-based population estimates significantly differ from the post-censal estimates based on the updating the previous census . The next step would be to recalculate annual population estimates back using the last census data. Unfortunately, not all countries do it. Moreover, a number of countries showing disruption in annual population estimates has increased in comparison with the previous round of censuses in 2000.

The standard HMD methodology (Wilmoth et al., 2007) for the cases when population estimates between two neighboring censuses are either not available or unreliable is based on the assumption of uniform distribution across the entire inter-censal period. This assumption works well in many conventional situations, but may be violated in the case of special events. For example, the collapse of the USSR and abrupt social-economic changes in Eastern Europe produced several migration waves at the end of the 1980s and over the 1990s. Huge and irregular migration waves followed the EU enlargement in 2004 and the financial crisis in 2008-2009. In such cases direct application of the basic HMD approach would not yield satisfactory results.

For example, the official population estimates for Bulgaria show a sudden drop in the total population count for the census years 1985, 1992, and 2001 (Figure 1). According to the official data, the total number of males decreased by about 252,000 between 1991 and 1992 (the census year). A similar notable discrepancy is found when one compares the official post-censal estimates as of December 31st, 2010 to the 2011 census counts (as of February 1st, 2011).

Discontinuities in the population trends are due to several factors. First, it seems that population estimates have not been recalculated backward using the latest censuses. Thus, the available data series consist of post-censal estimates rather than the inter-censal population estimates. Second, there was a significant amount of unregistered emigration (especially during the period from 1989 to 2010) not accounted for in the official statistics (Glei et al., 2015).

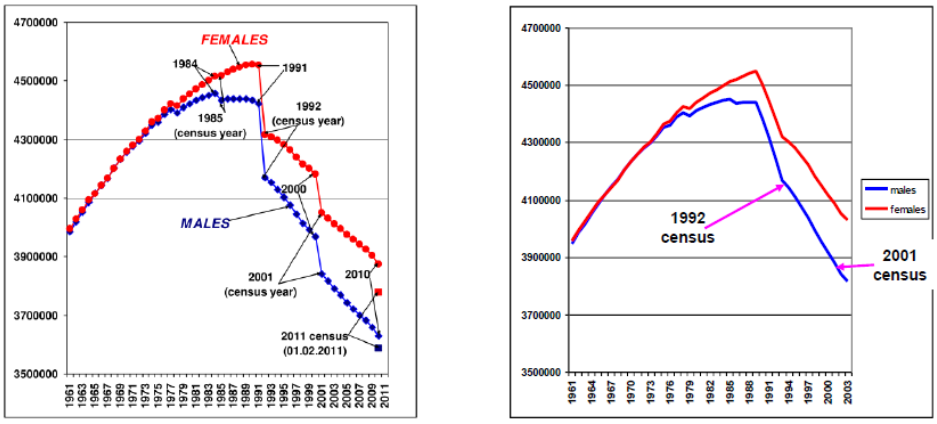


Figure 1. Trends in the official (left panel) and HMD (right panel) population by sex, Bulgaria, 1961-2011. Source: (Glei et al., 2015)

The standard HMD inter-censal method is not applicable to the period from 1985 to 1992 because of an irregular pattern of out-migration. During 1985 to 1988, international migration was very restricted in Bulgaria. After the collapse of Communism in 1989, the Government lifted the ban on free movement abroad, which led to mass emigration (mostly of the Turkish minority) over the next several years (IOM, 2003). After analyzing additional data (including indirect migration estimates), the HMD team decided to split the 1985-1992 inter-censal period into two sub-periods: 1985-1988 (a period of stable and negligible migration) and 1989-1992 (a period of substantial out-migration). The official population estimates were used for the former period, but new population estimates were calculated for the latter period. Specifically, the year 1988 was treated as a "pseudo-census point" for the start of the interval and the 1992 census counts were used to end the interval. The standard HMD inter-censal method was then used to derive adjusted annual population estimates (Glei et al., 2015).

3. Numerator-denominator bias

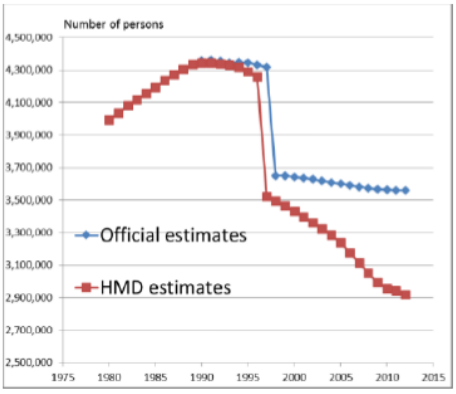
Problems with registration of out-migration together with relatively well functioning system of vital registration within country and inability to get information about vital events related to citizens living abroad may lead to a numerator-denominator bias. In this case a formally de facto population turns

into a de jure population, since it includes people registered in the country but living abroad, while vital events produced by these people are being registered only if these events occurr within the country.

Moldova provides an example of substantial numerator-denominator bias. This is despite a functioning population register. The registration of deaths and births in Moldova covers only the events that occur within the territory of the country (the de facto population), whereas the population estimates include Moldovan citizens who live abroad (Penina et al., 2015). In addition, Moldova experiences a very high level of out-migration. About 490,000 people left the country between the censuses of 1989 and 2004, and 322,000 people left between the censuses of 2004 and 2014. Thus, Moldova lost close to 800,000 its residents due to out-migration, or more than 20% of the total population, as enumerated by the 1989 census. The official net migration numbers are far more modest: 206,000 and 17,000 for the two inter-censal periods, respectively (Penina et al., 2015). All this (hidden) migrants are still registered in Moldova but the corresponding vital events are not included in the official population statistics of the country. This bias produces substantial differences between the de facto and the de jure population numbers.

(Penina et al., 2015) proposed an alternative population estimates based on unofficial data of the 2004 census which refer to the de facto population instead of the official counts of people registered as residents of Moldova. According to the alternative population estimates, net migration was equally redistributed over the inter-censal years 1989-2004 and 2004-2009. From 2009 onward, the annual net migration was estimated from the border crossing migration statistics. As one may expect, the crucial point by this reconstruction is availability of border crossing data, which is quite unusual for modern statistics. It allows obtaining more reliable estimates of migration.

It is often assumed that error due incorrect counting of migration has a rather minor impact on the aggregate mortality indicators such as life expectancy at birth because it accumulates around the youngest and the most mobile population groups, which have relatively low mortality. In Moldova, corrected population estimates are 18 percent lower than the official estimates (figure 2). The adjusted estimates of life expectancy at birth in 2014 were 64.94 years for males and 73.74 years for females. Compared to the official estimates these figures are by 2.58 years and 1.65 years lower, respectively.



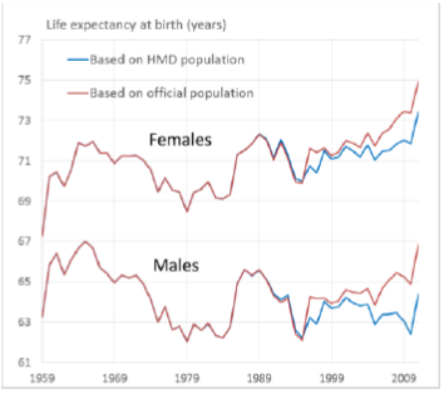


Figure 2. Official and alternative population estimates* of the Moldovan population (left panel) and life expectancy at birth in Moldova based on official and adjusted population estimates (right panel).
Source: (Penina et al., 2015)

* Since 1998 official population counts do not include Transnistria region

4. Changes in definition of population

Another important issue concerns changes in population definition. Procedure of implementation of the new definition is decisive in such cases. In 2000s, Poland faced a massive out-migration that followed the EU enlargement of 2004. This flow was understated by the official statistics after the census of 2002. It was expected that the population counts will be corrected downward after the next population census of 2011. But Statistics Poland has unexpectedly decided to change the official definition of the population status from the permanently resident (2010 and earlier) to the usually resident (from 2011 and later) (Fihel and Jasilionis, 2016). It led to the upward correction of the population counts. Unfortunately, Statistics Poland did not make any attempt to reestimate age-specific population counts back to previous censuses. Only population estimates for 2010 were recalculated using the new concept. As a result, there is a clear rupture in the official data (Figure 4). Due to irregular migration pattern the standard HMD inter-censal method for reconstruction of annual population estimates is not applicable. The HMD team at MPIDR had to use unofficial age-specific population estimates recalculated back to 2000 by Tymicki et al. (2015) and to apply a special adjustment factor in 2000 to take into account the change of the population concept from the permanently resident to a broader usually resident concept.

5. Infant mortality

In general, the HMD does not apply any correction to infant or early childhood mortality, but in many cases quality of registration of infant death can be considered as a useful indicator of the general quality of mortality data.

Even in countries with nearly complete vital statistics this problem may exist. The main source of underestimation of infant mortality in such cases is a restrictive definition of live birth. Perhaps the most cited example of such definition is the Soviet definition of live birth that existed with some modifications till (Anderson and Silver, 1986).

6. Impact of migration at working/reproductive ages on mortality and fertility estimates

As we discussed above, accurate annual population estimates are still a problem in the demographic statistics. While reliable census data are available in almost all developed countries, the annual estimates require much more efforts to reach sufficient data quality. Usually each new census brings inconsistency between the census results and earlier post-census population estimates based on the previous census and vital and officially registered migration events over the inter-censal period. In the previous section we discussed situations when statistical offices do not recalculate the annual inter-censal populations back from the newly available census. In some cases, however, official inter-censal estimates exist but their quality is much worse that one may expect. If there no additional or alternative annual data are available, it may be better to use the inter-censal estimates calculated using the standard HMD method based on the assumption of a uniform distribution of migration across time. If the overall level of migration is not too high, such estimates seem to be an optimal choice.

Noteworthy, mortality data are relatively insensitive to biases in the population denominator due to migration since mortality at ages of maximum migration is low. But this is not a case for fertility data. Quality of the population exposure is much more important for fertility indicators.

7. Mortality estimates at old ages

The old age population in developed countries has increased very rapidly throughout the second half of the 20th century. Improvements in survival are pushing it up to new limits: today more than half of all males and two thirds of all females born in Western countries may reach their 80^{th} birthday. The proportion of centenarians has increased by about ten times over the last thirty years, and more and more people celebrate their 100^{th} birthday (Robine and Vaupel, 2001). The importance of high quality mortality data increases with every decade. Despite it, internationally comparable high quality demographic data on old-age populations remain insufficient. The HMD is the only demographic database which provides such data. Population estimates for ages 80+ in the HMD are recalculated using extinct/almost extinct cohort and survival ratio methods (Wilmoth et al., 2007). But even such extensive

correction of old age mortality does not solve all problems (Jdanov et al., 2008).

Evaluation of data quality at old ages was extensively discussed elsewhere (Jdanov et al., 2008; Kannisto, 1994). The standard set of methods includes tests on age overstatement (e.g. the ratio of the total person-years lived above age 100 to the total person-years lived above age 80), precision of age reporting with the UN age-sex accuracy index, age heaping with the Whipple's Index of age accuracy. The comparison to other countries with reliable statistics may be also used for evaluation of overall quality of mortality estimates.

8. Conclusion

There is no perfect data in the world, but it is enough to have high quality data. Data are of high quality if they are "Fit for Use" in their intended operational, decision-making and other roles (Juran and Godfrey, 1999). This is why the understanding of problems hidden in the data is important in any demographic estimation, forecast or study. We discussed several approaches which allow us to increase significantly utility of the data even if data quality is problematic. Unfortunately, standard demographic methods which work well with data from developing countries or historical data series are often not applicable to problematic data from countries with functioning modern statistical systems. Such data lead to new challenges and new problems. To solve these problems more laborious approaches in combination with usage of additional and alternative data sources are needed. Country-specific approach should be combined with certain general principles that are applied in all countries to ensure comparability of data series across time and space.

References

1. Anderson, B.A., Silver, B.D., 1986. Infant Mortality in the Soviet Union: Regional Differences and Measurement Issues. Popul. Dev. Rev. 12, 705–738. https://doi.org/10.2307/1973432
2. Barbieri, M., Wilmoth, J.R., Shkolnikov, V.M., Glei, D., Jasilionis, D., Jdanov, D.A., Boe, C., Riffe, T., Grigoriev, P., Winant, C., 2015. Data Resource Profile: The Human Mortality Database (HMD). Int. J. Epidemiol. 44, 1549–1556. https://doi.org/10.1093/ije/dyv105
3. Fihel, A., Jasilionis, D., 2016. About mortality data for Poland (Background and Documentation). Human Mortality Database.
4. Glei, D.A., Lundstrom, H., Wilmoth, J., Borges, G., Barbieri, M., 2015. About mortality data for Sweden (Background and Documentation). Human Mortality Database.

5. IOM, 2003. Migration Trends in Selected Applicant Countries - Volume I - Bulgaria - The social impact of seasonal migration. International Organiztion for Migration.
6. Jdanov, D.A., Jasilionis, D., Rau, R., Vaupel, James, 2008. Beyond the Kannisto-Thatcher Database on Old Age Mortality: an assessment of data quality at advanced ages (Working Paper No. WP-2008-013). MPIDR, Rostock.
7. Juran, J.M., Godfrey, A.B., 1999. Juran's Quality Handbook. McGraw Hill.
8. Kannisto, V., 1994. Development of oldest-old mortality, 1950-1990: Evidence from 28 developed countries, Monograph on Population Aging. Odense University Press, Odense.
9. Penina, O., Jdanov, D.A., Grigoriev, P., others, 2015. Producing reliable mortality estimates in the context of distorted population statistics: the case of Moldova. Max Planck Institute for Demographic Research, Rostock, Germany.
10. Robine, J.-M., Vaupel, J.W., 2001. Supercentenarians: Slower ageing individuals or senile elderly? ResearchGate 36, 915–30. https://doi.org/10.1016/S0531-5565(00)00250-3
11. Setel, P.W., Macfarlane, S.B., Szreter, S., Mikkelsen, L., Jha, P., Stout, S., AbouZahr, C., 2007. A scandal of invisibility: making everyone count by counting everyone. The Lancet 370, 1569–1577. https://doi.org/10.1016/S0140-6736(07)61307-5
12. UN, 2014. Principles and Recommendations for a Vital Statistics System, Statistical Papers, Series M. United Nations.
13. WHO, 2010. Improving the quality and use of birth, death and cause-of-death information: guidance for a standards-based review of country practices. World Health Organization.
14. Willekens, F., Massey, D., Raymer, J., Beauchemin, C., 2016. International migration under the microscope. Science 352, 897–899. https://doi.org/10.1126/science.aaf6545
15. Wilmoth, J.R., Andreev, K., Jdanov, D.A., Glei, D.A., 2007. Methods protocol for the Human Mortality Database. University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. URL: http://mortality.org [version 31/05/2007].

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

Data analytics for better statistics

Chong Ning, Benjamin Ng, Jeremy Heng
Ministry of Manpower, Singapore

Abstract

Modern technology has revolutionized the way official statistics is produced and consumed. With large amounts of data available, Singapore's Ministry of Manpower seeks to tap on data analytics to improve the quality of statistics produced and formulate better policies. In the Manpower Research and Statistics Department, data analytics is used in three areas to aid in the statistical production process: (1) automated-classification system, (2) sentiment analysis and (3) predictive modelling.

An automated-coding system is developed to automate the conversion of raw occupation and industry data into standard occupation and industry codes. It ensures consistency among interviewers and respondents who may have their own understanding of occupation and industry definitions. A speech-to-text analytics tool facilitates sentiment analysis to provide insights into the behaviour of interviewers and respondents. By analysing telephone conversations, it is able to flag out uncooperative respondents and under-performing interviewers for follow-up action. Lastly, a fieldwork predictive model is used to predict the optimal dates and times to conduct survey interviews with various demographics of respondents, thereby reducing the likelihood of non-response and refusal cases.

Through these initiatives, the Ministry is able to improve operational efficiency and data quality. The paper discusses the challenges facing official statistics, the idea behind the data analytics initiatives, how they are able to tackle the challenges, and ultimately take official statistics into the future.

Keywords

Efficiency; Quality; Prediction; Optimization; Classification

1. Introduction

Singapore's official labour statistics is produced and compiled by the Manpower Research and Statistics Department (MRSD) of the Ministry of Manpower. The department conducts regular national surveys to collect a wide range of labour-related data from households, individuals and businesses. The survey data is cleansed and processed into usable information which is then analysed to provide labour market insights. The statistics and accompanying publications are eventually disseminated to the public and

policymakers for their use. The labour market indicators that track the state of the economy include, but is not limited to, employment, unemployment, income, job vacancies, labour turnover, retrenchment, employment conditions and hours worked.

With increasing demands for more granular and timely statistics, MRSD has leveraged on technology in recent years to implement various data analytics initiatives to meet those demands.

In the following sections, the data analytics initiatives are explored in greater detail, and how they have helped MRSD to improve operational efficiency and data quality.

2. Methodology

In this section, the methodology behind the data analytics initiatives is explained in greater detail.

a) Predictive Modelling

With limited resources and increasing demands for large amounts of data at short notice, it is important to optimise resources to collect as much survey data as possible. Survey data is collected online or through phone or face-to-face interviews. Predictive modelling is done by making use of demographic information such as age, gender and household composition of the respondents. A random forest model can be trained with past information to predict optimal timings to contact the selected respondents.

A random forest model is built from growing multiple decision trees, with each tree depending on the values of a random predictor sampled independently (James, Witten, Hastie & Tibshirani, 2013). At each split, only one of m randomly selected predictors are considered from the full set of p predictors (James et al., 2013). This method prevents high correlation among trees from the influence of strong predictors (James et al., 2013). The outcome is taken from the average of all predictions to reduce variance (James et al., 2013).

Assuming each household has a responsible adult, the simplified decision tree shows the most appropriate timing of contact. For instance, if a household comprises a couple of working age, it would be recommended to only contact them outside normal working hours. However, for a family with young children, establishing contact with them during working hours has a higher probability of success as an adult is more likely to be at home to care for the children (Diagram 1).

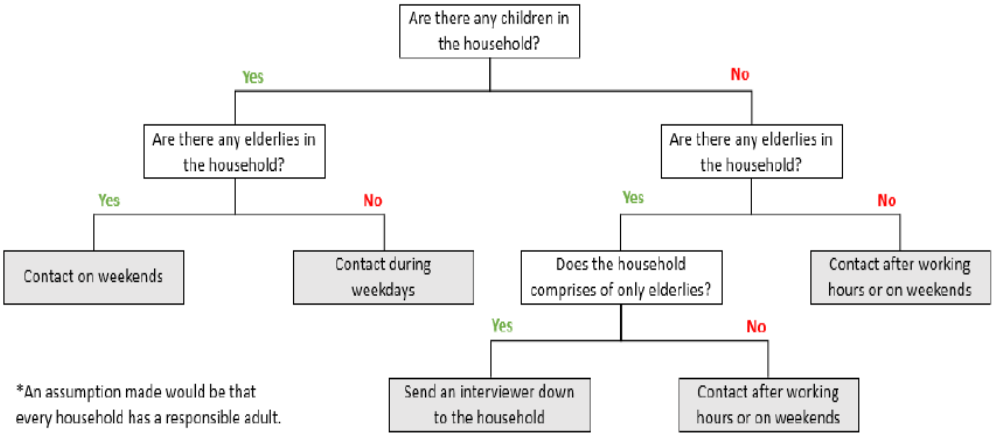


Diagram 1: Decision Tree on when to visit a household

In addition to the optimal timing of contact, the mode of contact is equally important. In certain cases, a call to remind the respondents to complete the survey online would suffice, but for others a visit to their house may be more suitable. For instance, if the household comprises of only elderlies, they may have difficulty completing the survey online or over the phone (Diagram 1). Therefore, sending an interviewer to visit the household for a face-to-face interview may yield a more successful result. Hence, this tool is able to help MRSD make calculated decisions on the manpower and workload distribution while maximising survey returns.

b) Sentiment Analysis

In every business, it is important to understand the public sentiment towards the business. It is no different with survey operations. With over 30 surveys conducted each year, MRSD sees value in gaining a better understanding of survey respondents to alleviate their burden and cater to their needs.

As phone interviews comprise a significant proportion of survey responses, MRSD has adopted the use of sentiment analysis for every telephone conversation. This would help to filter out difficult or uncooperative respondents. Through a speech-to-text analytics tool, telephone conversations are converted to text data. Text mining of contextual keywords is subsequently applied. Respondents who depict negative sentiments are then flagged out for follow-up action by a more experienced interviewer or Team Leader (Diagram 2).

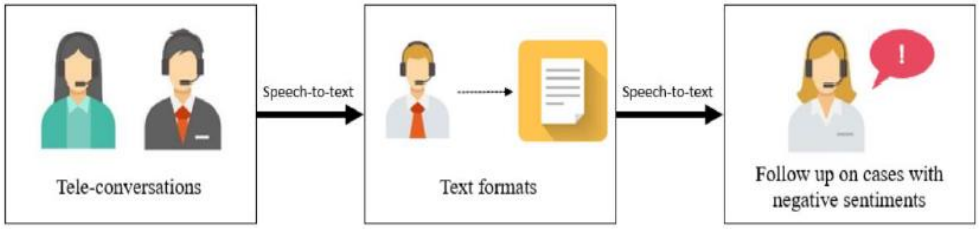


Diagram 2: Process flow of Sentiment Analysis

Sentiment analysis is also utilized when assigning cases to interviewers performing fieldwork. Cases can be accurately matched to interviewers who possess the necessary skillset to complete the case. For example, a new interviewer will be assigned cases which previously had positive or neutral sentiments while a more experienced interviewer will handle cases that had history of negative sentiments. This ensures that each interviewer will be able to conduct effective interviews with the respondents, hence increasing the survey response rate.

Sentiment analysis can also help to measure the performance of interviewers. An interviewer who continuously has negative connotations in their conversations with survey respondents will be flagged out for retraining on customer service skills. Similarly, interviewers with good customer service skills can also be identified. As such, interviewers are not only monitored quantitively on their output, but also on the qualitative side such as soft skills. This ensures that all survey respondents go through an optimized survey journey experience.

c) Automated Classification System

As MRSD compiles statistics on the labour market, information on occupation and industry are key data items that will aid policymakers have an accurate sensing of the labour market. It is also the most tedious data item to collect, requiring large amounts of time and resources to classify the textual information. In the past, respondents would just provide some details of their occupation and interviewers have to classify each of them into one of the 1,202 codes of the Singapore Standard Occupational Classification (SSOC).

The Automated Classification System (ACS) was developed to cope with the increasing textual data being collected. It utilizes text analytics algorithms to convert unstructured data into meaningful structured data that can be used for analysis. The data is cleaned through a process of word tokenisation, customised stemming, removal of stop words and punctuations (Diagram 3).

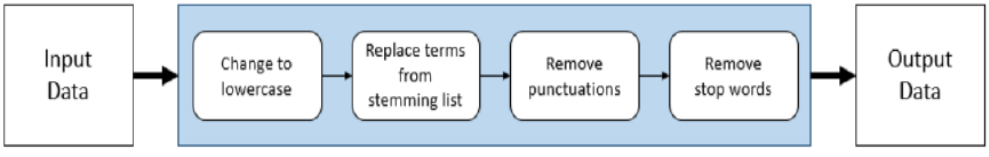


Diagram 3: Process flow of Data Cleaning

Subsequently, feature extraction is used to convert textual data into vectors. The term frequency-inverse document frequency ($\mathrm{tf}-\mathrm{idf}$) is an information retrieval model used to weigh the relevance of the term in a document (Manning, Raghavan & Schütze, 2008).

The term frequency (tf) is defined by the frequency of the occurrence of a term t in the document d as

$$\mathrm{tf}_{t,d} = \frac{n_{t,d}}{\sum_k n_{t,d}}.$$

The inverse document frequency (idf) of term t is defined as

$$\mathrm{idf}_t = \log \frac{N}{df_t}.$$

The idf ensures that the weight of rare terms increases while it decreases for commonly used terms (Manning et al., 2008). The $\mathrm{tf}-\mathrm{idf}$ weighting gives the term t a weight in document d by

$$tf - idf_{t,d} = tf_{t,d} \times idf_t.$$

Past data that has been labelled with SSOC is used to train the model.

Cosine similarity is then used to score and rank the relevance between two documents d_1 and d_2. To offset the effect of difference in document length, cosine similarity of their vector representation $\vec{V}(d_1)$ and $\vec{V}(d_2)$ is used for computation

$$sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}.$$

The most suitable output is derived based on the highest cosine similarity for the term

Job Title
Accountant
Chef
Waitress

Job Title	SSOC
Accountant	24111
Chef	34341
Waitress	51312

Diagram 4: The output based on the highest cosine similarity score.

Through the use of ACS, time spent on manual classification is reduced and resources can be further deployed more effectively.

3. Results

The implementation of these initiatives has helped improve operational efficiency and data quality. Before predictive modelling was implemented, interviewers were unaware of the optimal timings to contact respondents and as a result, the successful completion rate for fieldwork was only around 20%. This proportion has more than doubled after the implementation of the model as it provides interviewers with the knowledge on when and how to contact the survey respondents.

The completion rate is further boosted to around 60% with sentiment analysis as case assignment is done more strategically, by allowing more experienced interviewers to handle difficult respondents. Sentiment analysis has also helped interviewers to improve on their interviewing skills.

After the data is collected, ACS has helped to reduce the time needed for data verification. More cases can be verified to ensure data consistency and robustness with the same amount of time. More importantly, ACS eliminates the human variable factor when classifying the occupational information into their respective codes. This ensures consistency for each code and eliminates any subjectivity on the part of interviewers.

4. Conclusion

With increasingly widespread use, data analytics will emerge as an important tool in the statistical production process. The three initiatives described in this paper is the start of revolutionizing the way official statistics is produced in future. Encouraged by the positive results, MRSD will continue to explore new initiatives in data analytics to aid in its survey operations.

References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. New York: Springer.
2. Manning, C.D., Raghavan, P., Schütze, H. (2008). Introduction to Information Retrieval. Cambridge (England): Cambridge University Press.

Trusted official smart statistics - Challenges for official statistics in using data sources coming from private data producers

Markus Zwick
Federal Statistical Office of Germany

Abstract

The Federal Statistical Office of Germany engaged in the Eurostat financed project 'Smart business cycle statistics (SBCS)' from the beginning of 2018 until March 2019. The aim of the project was to use sensor data coming from satellite and flyover observations for the production of economic indicators. The project SBCS has shown that business cycle indicators based on satellite data will nearly be possible in real-time and cross-border. With these characteristics, SBCS will expand the information content fundamentally compared to currently used business cycle indicators. In general, National Statistical Institutes (NSIs) will be able to produce SBCS in the future, but probably with external partners and sometimes even in cooperation with private data producers. The paper discusses the production of official statistics in the IoT age, on the example and the experience made at the project SBCS. IoT based official statistical products will probably mostly be work-sharing products, with specialists in the three working fields 'conception and quality', 'data engineering' and 'data visualization'. For getting high quality statistical products, the rules for integrating private data products into Official Statistics have to be developed.

Keywords

New digital data sources; Official Statistics; Privately held data; Quality; Skills

1. Introduction

The opportunities and challenges National Statistical Offices (NSIs) are facing are growing with the digital revolution. In general, official statistics will be faster, more precise for small groups and all of that with fewer burdens for the respondents in the future. Therefore, statistical offices have to be transformed in a modern information provider. Otherwise private data producers will do more of what we call official statistics.

The paper will describe some of the challenges of the digital transformation process for statistical offices. Big Data, meanwhile called as 'new digital data' or 'non-traditional data' allow a deeper and more precise picture of the society and the economy. Mobile phone, satellite as well as data from the internet add information to official statistical products, in the meanwhile and even more in the future. In some cases, these kinds of new

data sources could even replace survey data. But in general, the assumption in the European Statistical System (ESS) is that new digital data will be an additional data source and official statistical products will be multiple source products. Linked survey, administrative as well as non-traditional data will be the basis for official statistical products in the future.

Besides legal and ethical challenges, the access to the often privately held digital data sources as well as to the necessary skills for developing the algorithm in the new digital statistic production could be a central limitation for NSIs (ESS 2017a).

With the Big Data Roadmap of the ESS a lot of initiatives started in the year 2015. With the ESSnet Big Data 2016-2018 eight projects were finalized in May 2018 (https://ec.europa.eu/eurostat/cros/content/essnet-big-data_en) .The following ESSnet Big Data 2018-2020 has started on 1st November 2018. The Federal Statistical Office of Germany (Destatis) is engaged in these initiatives. Furthermore, Destatis ran the project 'Smart business cycle statistics (SBCS)' financed by Eurostat. This project tried to use satellite images for describing the change of economic activities over the time. Besides this, Destatis works on different projects by using mobile phone data. One result of these activities is that NSIs probably will use more semi-finished statistical products coming from private data producer in the future.

This paper is focusing on the experience with the project SBCS in respect of using semi-finished statistical products for official statistics.

2. Smart business cycle statistics

In 2017, Eurostat, the statistical office of the European Union, published the call for tender 'Smart Statistics'. The aim of the call was to get 'Services for developing "Proofs of Concept (PoC)" of leveraging data within an extended Internet of Things (IoT) ecosystem for the production of smart statistics in Commission policy areas.' A consortium of SOGETI Luxembourg, a data consulter, the Jožef Stefan Institute from Slovenia and the Federal Statistical Office of Germany (Destatis) won the competition of this tender and conducts the project since March 2018 until March 2019. The Jožef Stefan Institute did the two PoCs 'Smart mobility for Smart Cities' and 'Smart labour market statistics'. Destatis was responsible for the PoC 'Smart business cycle statistics'. SOGETI organised the administrative frame of the project and the 'Workshop on Trusted Smart Statistics: policymaking in the age of the IoT' as part of the project in Germany in January 2019 (https://ec.europa.eu/eurostat/cros/content/workshop-trusted-smart-statistics-policymaking-age-iot-0_en) . More detailed articles of the three PoCs are in preparation.

The main idea of the PoC 'Smart business cycle statistics (SBCS)' was to use sensor data coming from satellite and flyover observations for producing

economic indicators. Different existing satellite data were proofed for describing observable economic activities. High resolution satellite and non-space (e.g. airborne and seaborne) data contains valuable information for business cycle analysis. For example, activities at harbours, airports or industrial facilities can be observed. A set of statistical indicators based on these data can give first signs of the possible changes in the economy and therefore to provide information to the government to take appropriate measures, for example, against overheating of the economy at the earlier stage.

The PoC SBCS has shown that nearly real-time business cycle statistics based on satellite data are possible. Furthermore, the indicators describing business cycles could be developed as cross-border indicators. A lot of economic activities do not stop at national frontiers. Connected economic areas are cross-border (see Taubenböck et al 2017). With these characteristics, SBCS will expand the information content fundamentally compared to currently used business cycle indicators.

The starting point of the PoC SBCS was data coming from the Copernicus Programme of the European Space Agency (https://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus). Especially the images of the Sentinel 2 satellites were of key interest at the beginning. Copernicus is an Earth observation program directed by the European Commission in partnership with the European Space Agency, which includes the Sentinel missions. Sentinel-2 delivers multispectral satellite images, which have a spatial resolution of 10 m. The Sentinel-2 mission consists of two satellites through which the Earth can be observed every 5 days. The main advantage of using Sentinel-2 is its free and open data policy. However, the spatial resolution of 10 m does not allow detecting some of the objects necessary, such as cars or containers. The lessons learned from the PoC SBCS is first of all that the current free of charge available satellite data are not good enough to detect smaller objects related to economic activities. On the other hand, there are a lot of high quality data at the information market but this data are still quite expensive (Cao et al 2016). An overview of some of the more commonly used Earth Observation satellites in terms of spatial resolution and revisit time is shown in Figure 1. With the further technical progress the price for nearly real time and high-resolution satellite images will probably decrease.

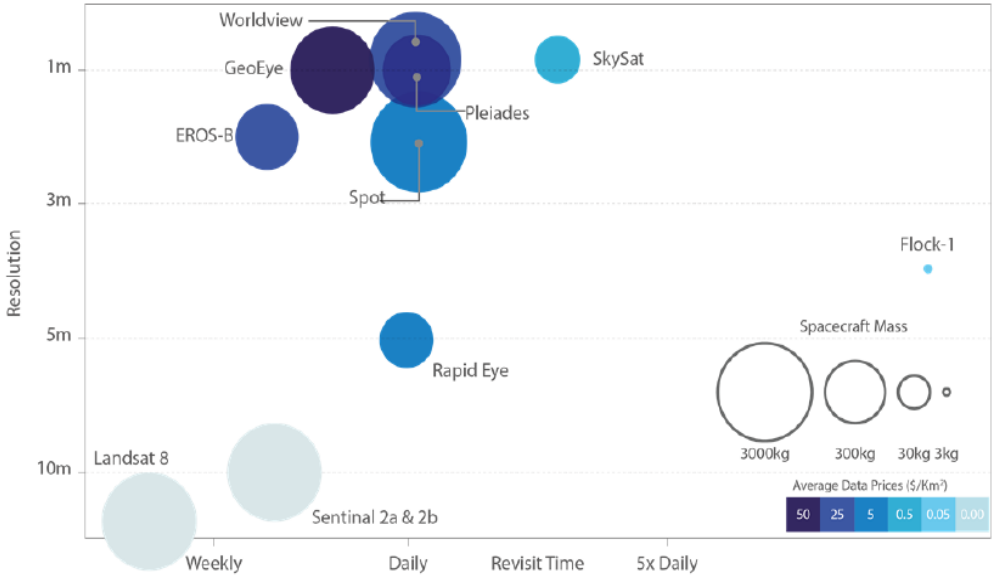


Figure 1: Spatial Resolution vs Revisit Time for various example satellites
(Source: Satellite Applications Catapult 2017, https://sa.catapult.org.uk/).

As a conclusion at this point, we can summarize that in principle NSIs are able to produce smart business cycle statistics nearly in real time and for cross-border regions. But for regular SBCS products, statistical offices have to do a lot of own research and to establish cooperations with other specialized institutions, sometimes even with private data producers.

3. Challenges for NSIs by producing trusted smart statistics

Producing the entire SBCS in statistical offices is challenging. The challenges include the IT infrastructure, the soft- as well as at the hardware. High resolution satellite data are really big data. For example the sentinel satellites, from the quality point of view not good enough for SBCS, collect images of the whole earth surface every five days. Apart from the complexity of the data volume, the skills necessary for detecting objects in satellite images, based on machine learning algorithms (Xu et al 2018), are not yet available in most statistical offices. It is to discuss whether the necessary infrastructure should be established in statistical offices. As an alternative, semi-finished statistical products, coming from partners outside of the official statistical system, could be integrated into official statistics because there exists a break-even where the use of semi-finished statistical products from private enterprises will be cheaper compared to the self-production of NSIs. Furthermore, in some cases the self-production could be impossible because the limitations of access to the granulated data and to the necessary skills to transform these data into statistical products. In the case of using private data,

clear rules have to be developed, which semi-finished statistical products have to fulfil, before NSIs can integrate this data in official statistical products.

This issue shows that SBCS is an example of the fundamental question of the role of NSIs in IoT times. The necessary IT infrastructure to produce smart statistics based on data coming from the IoT has to be capable of handling much larger amounts of data than the infrastructure currently available in NSIs. The required skills to transform IoT data into official statistical products do not exist in most statistical offices. Due to the high demand of these skills on the labour market and the conditions of government institutions, the recruitment is especially challenging, if not impossible. The data scientists who are able to do the job are in institutions that have the IT infrastructure to work with IoT data because they pay better compared to government institution. During the PoC SBCS Destatis got a lot of support from the German Space Agency. This was necessary because the knowledge for object deduction based on sensor data doesn't exist at Destatis. Furthermore, for analysing the sentinel data the Copernicus Open Access Hub were used (https://scihub.copernicus.eu/). This platform provides full access to all Sentinel-1,-2 and -3 user products. The Copernicus Open Access Hub is a service for analysing a big amount of data financed by the German government. It would be economic inefficient to build up the same expensive IT infrastructure also at Destatis. Therefore, Destatis has to work with semi-finished statistical products if Copernicus products should use. As mentioned above, Sentinel data were not well enough for producing SBCS.

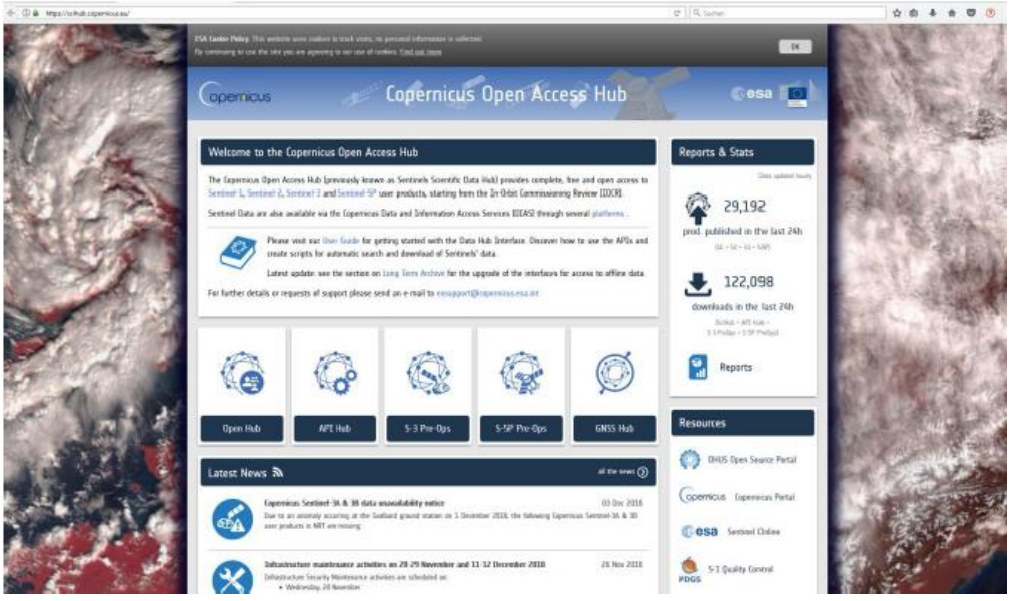


Figure 2: The Copernicus Open Access Hub allows free and open access to satellite images (https://scihub.copernicus.eu/)

During the project, private data from SpaceKnow (https://www.spaceknow.com/) were used as well. SpaceKnow is a start-up company from the United States and has analysed - for the PoC - high resolution satellite data for 50 parking places in Germany at more than 20 dates in 2018. Furthermore SpaceKnow analysed container at different harbours in Europe. In general, the data were tailor-made for the PoC SMCS, but not sufficient for a permanent official data production. Besides, the price, especially the access to the meta-data, is an issue. For Destatis, it was not clear how the algorithm had transformed the data as neither the algorithm nor the meta-data were accessible.

Based on the experience done by the PoC SBCS as well as the projects based on mobile phone data it can be concluded that the quality issue will be the core business for statistical offices in times of IoT. It is of high importance to have a view on the data generating processes and to understand, not to develop, the algorithm which transforms the data. Some of the new products based on digital data look very impressive especially if the results are visualized in a dynamic presentation, sometimes in real time. But the most important issue for official statistics in IoT times should be to have a quality influence on the statistical products that are the basis for decision-making.

Figure 3: Small object detection from satellite images by algorithm (https://www.spaceknow.com/)

That means that in the case of the PoC SBCS, the statistical offices are able to produce satellite-based business cycle indicators, but with the support of other institutions or by buying semi-finished statistical products. They need strong partners who are able to process the big volume of nearly real time data in combination with the necessary experience and skills. The statistician's job is conceptual design and to guarantee the quality of trusted smart statistics.

4. Conclusions

Until now, the experience made in the different projects gives a first hint to the further work of NSIs in times of IoT. Probably not all production steps

to get high quality official statistics will be done in the statistical offices in the future in some cases semi-finished statistical products coming from private date production will used. The Generic Statistical Business Process Model (GSBPM) (https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0) has to be rethought in the context of IoT. The experience has shown that generally NSIs aren't able to produce real-time statistics based on sensor data alone currently. A lot of investment is necessary for the IT infrastructure and regarding the skills of the staff in the NSIs for this task. But the discussion has only started whether this is efficient and possible.

The IT infrastructures in NSIs have to be further developed, but the question is in which format. In some cases the necessary IT infrastructure is established by the government, but in other official institutions. In some cases buying data services could be cheaper compared to a situation in which NSIs do the whole production by themselves. As discussed above, it is not clear whether NSIs are able to hire data scientist at the labour market with the right skills and in the quantity who is needed. The competition for getting these skills is intensive and therefore this production factor is expensive, maybe too expensive for NSIs. The next challenge is the access to the often privately held data of the IoT. Current experience has shown that it is often more realistic to buy the whole product as service; that means the product coming from the combination of adequate IT infrastructure, skills and data. Often, private data producers do not have the interest to sell only the detailed data. Generally, buying this kind of semi-finished statistical products could be a good solution for NSIs, because self-production is also expensive. The questions are, where the break-even is, which quality the data have and how permanent the data access is.

Future official data production probably will be done as work-sharing for some products. Therefore, the rules for integration private data products into official statistics are to define. Another important issue is the sharing of responsibilities between the private and the official data production. First of all, the statistical conception, including economic and sampling issues, is needed for statistics production. This will still be the task of the statistician. But statistical offices have to further develop their staff's skills for doing these parts of the statistical production. Probably it will not be the task of NSIs to do the data engineering production steps, like developing algorithms to detect small objects based on machine learning as example, but it will be necessary that NSIs are able to understand the methods that are used. Besides, the conception and the collection of the data and the publication of the results in an interactive and intuitive format will be the third process step. The aim should be to get a system of automated processes to steer the whole production process free of media disruption and based on a further developed GSBPM.

One of the results of the PoC SBCS is that most of the statistical offices will not be able to produce statistical products of the IoT age by themselves. IoT based statistical products will be a work-sharing product, with specialists in the three identified fields 'conception and quality', 'data engineering' and 'data visualization'. For getting a high quality statistical product, the rules for the interfaces between the work-sharing partners have to be developed.

References

1. Cao, L., Wang,. C., Li, J., 2016: Vehicle detection from highway satellite images via transfer learning. Information Science 366, pp. 177-187.
2. European Statistical System (2017a) Position paper on access to privately held data which are of public interest, https://ec.europa.eu/eurostat/documents/7330775/8463599/ESS+Position+Paper+on+Access+to+privately+held+data+final+-+Nov+2017.pdf/6ef6398f-6580-4731-86ab-9d9d015d15ae (accessed on 29.01.2019) .
3. European Statistical System (2017b) European Statistics Code of Practice, https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice (accessed on 29.01.2019).
4. Taubenböck H, Ferstl J & Dech S (2017): Regions set in stone – Classifying and categorizing regions in Europe by settlement patterns derived from EO-data. ISPRS Internatl. Journal of Geo-Information, 6(2), pp. 1-27.
5. Xu, Y., Zhu, M., Xin, P., Li, S., Qi, M., Ma,S., 2018: Rapid Airplane Detection in Remote Sensing Images Based on Multilayer Feature Fusion in Fully Convolutional Neural Networks. Sensors 2018, 18, 2335.

Unreliable retrial queue with two types customers, two delays and vacations

Saggou Hafida Hafida, Lach Lachemot Tassadit, Sadeg Ines Ines, Ourbih Tari Megdouda
Universite Des Sciences Et E La Technologie Houari Boumedienne Alger

1. Introduction

The queueing theory is used in various applications, this classical theory soon proved ineffective in the face of real system more and more complex. The limitations of the classical theory of queues that did not allow to explain the stochastic behavior of telephone systems where subscribers repeated their calls by redialling the number several times to obtain the communication, this phenomenon is called retrial queues.

Queueing systems with retrials are characterized by the next customer who finds the server busy is obliged to leave and joins the retrial group (called"Orbit") and repeat this demand for service after some random time. The retrial queue have been widely used in the telephone switching systems, computer networks and computer systems, maintenance and repair problems. For more details on these models, see (Yand and Templeton, 1987; Falin, 1990; Kulkani and Liang, 1997) and monographs by (Falin and Templeton 1997; Artalejo and Gomez-Corral, 2008).

(Fayolle, 1986) considered an *M/M/*1 retrial queue, in which a customer who finds the server busy joins the tail of a retrial queue. Only the customer at the head of the orbit is allowed to attempt to receive service after an exponentially distributed retrial time in competition with a new primary customer. (Farhamand, 1996) called this discipline a retrial queue with FCFS orbit. (Choi and al, 1992) generalized this retrial policy by considering an *M/M/*1 retrial queue with general retrial times.

In this paper, we consider an $M^{[X]}/G/1$ retrial queue with general retrial time with two classes of customers, transit and recurrent customers, service subject to random actives breakdowns and repairs. The customer whose service is interrupted stays in the service, waiting for the first delay of verification, repair and second delay of verification. After the second delay of verification, this customer completes his service. We assume that after every service completion, the server has the option to leave for a vacation of random length with probability $(1 - r)$ or to can serve the next customer with probability *r*. The batch arrival who finds the server busy or failed or under the 1^{st} or 2^{nd} delay of verification or repair or vacation are allowed to balk with probability $1 - p$ or to stay in the system with probability p in according with *F.C.F.S.*

The paper is structured as follows. In the next section, we give the mathematical description of our queueing model. In section 3, we analyze the steady-state distribution of the queueing system under consideration. The different probabilities and important performance measures and reliability indices this model are given in section 4. In section 5, we present the stochastic decomposition property.

2. The mathematical model

We consider an $M^{[X]}/G/1$ retrial queueing model with two types of customers: transit (also called ordinary) customers and a fixed number K ($K \geq 1$) of recurrent (also called permanent) customers. Our model is based on the following hypotheses.

1. Transit Customers arrives at the system according to a compound Poisson process with rate λ. Let X be the Upon arrival, if a batch of transit customers finds the server busy, it can joins the retrial group (orbit) in accordance with an *F.C.F.S* discipline with probability p or leaves the system with probability $1-p$. We assume that only the transit customer at the head of the orbit is allowed to access the server. Successive inter-retrial times of any transit customer follow an arbitrary probability distribution function $A(x)$, with a corresponding density function $a(x)$ and a Laplace-Stieldjes transform $L_A(s)$.

2. A single server can serve only one-by-one transit customer at a time based on the *F.C.F.S* discipline of service concerning the batch transit arrival. Successive service times are independent with a common probability distribution function $B_1(x)$, a density function $b_1(x)$, a Laplace-Stieltjes Transform (*LST*) $L_{B1}(s)$, n^{th} moments β_{1n}.

3. There is a fixed number K of recurrent customers in the system. Once served, recurrent customers immediately return to the retrial group in accordance with an *F.C.F.S* discipline. We assume that only the recurrent customer at the head of the orbit is allowed to access the server. The recurrent customer at the head of the group repeats his call after an amount of time following an exponential distribution with the parameter γ.

4. The service time of recurrent customers are *i.i.d* with a probability distribution function $B_2(x)$,, a density function $b_2(x)$,), a Laplace-Stieldjes transform $L_{B2}(s)$ and n^{th} moments β_{2n}

5. Once the service time of the transit customer is completed, the server can go on vacation with probability $1-r$ or stays in the system for serving

another transit customer with probability r. We assume that the vacation times are *i.i.d* random variables with a probability distribution function $V(x)$, a density function $v(x)$, an LST $L_V(s)$ and n^{th} moments ν_n.

6. The server is subject to active breakdowns, i. e. the server fails if and only if it is rending service. It fails after an exponential amount of time with rate $\mu > 0$. The customer receiving service during a breakdown has to wait until the service is recovered. Once the system breaks down, its repairs is not supported immediately, there is a first delay time verification which has a general distribution with a distribution function $D_1(x)$, a density function $d_1(x)$, an LST $L_{D1}(s)$and n^{th} moments $\emptyset_{1,n}$.

7. After a first verification delay, the repair process starts immediately. The repair times are *i.i.d* random variables according to a general distribution function $R(x)$, a density function $r(x)$, an LST $L_R(s)$ and n^{th} moments η_n.

8. As soon as the repair time is over, the server will have a second verification delay, the delay times are independent with a commonprobability distribution function $D_2(x)$, a density function $d_2(x)$, an LST $L_{D2}(s)$and n^{th} moments $\emptyset_{2,n}$.

The customer whose service is interrupted remains in service. Once the second verification delay completed, the server takes over the service of this customer. The server is not authorized to accept another customer until the customer in service leaves the system. The customer at the head of the orbit enter in competition with a new transit customer for an attempt to receive service.

The state of the system at time t can be described by the following Markov process

$$\{X(t), t \geq 0\} = \{(S(t), S^*(t), N(t), \xi_0(t), \xi_1(t), \xi_2(t), \xi_3(t), \xi_4(t), \xi_5(t), \xi_6(t)); t \geq 0\}$$

where $S(t)$ takes its values in the $set\ \{0,1,2,3,4,5,6\}$ according to the server being idle, busy with a transit customer, busy with a recurrent customer, under first delay of verification, under repair, under vacation, or under second delay of verification. If $S(t) = 3, 4\ or\ 6$, we define $S*(t)$ as a type of customers in service ($S*(t) = 1\ or\ 2$ according to the occupancy of the server by a transit or a recurrent customer). If $S(t) = 0$ and $N(t) > K$, then $\xi_0(t)$ represents the elapsed retrial time of the transit customer. If $S(t) = 1$, we define $\xi_1(t)$ as the elapsed service time of the transit customer. If $S(t) = 2$, we define $\xi_2(t)$ as the elapsed service time of the recurrent customer. If $S(t) = 3$, we define $\xi_3(t)$ as the elapsed first verification delay time. If $S(t) = 4$, we define $\xi_4(t)$ as the elapsed repair time; If $S(t) = 5$, we define $\xi_5(t)$ as the elapsed

vacation time and if $S(t) = 6$, we define $\xi_6(t)$ as the elapsed second verification delay time.

3. Analysis of the steady-state probabilities

We investigate in this section the steady-state distribution of the system. The conditional completion rates for the repeated attempts of transit customers, for the service of transit customers, for the service of recurrent customers, for the first verification delay, for repair, for vacation and for second verification delay times are given.

Theorem 1. Under the stability condition, the marginal Probability Generating Functions of the server's state queue size distribution are as follow:

$$P_0(z) = \frac{(\lambda+\gamma)[z(\lambda+\gamma)-\emptyset(z)P_{0K}z^K]}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-L_A(\lambda+\gamma)}{(\lambda+\gamma)}$$

$$P_1(z) = \frac{(\lambda+\gamma)L_A(\lambda+\gamma)[\lambda(1-g(z))+\gamma(1-k_2(z))]P_{0K}z^K}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-k_1(z)}{K(z)}$$

$$P_2(z) = \frac{\gamma(\lambda+\gamma)L_A(\lambda+\gamma)[q(z)-z]P_{0K}z^{K-1}}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-k_2(z)}{K(z)}$$

$$D_{1,1}(z) = \mu\frac{(\lambda+\gamma)L_A(\lambda+\gamma)[\lambda(1-g(z))+\gamma(1-k_2(z))]P_{0K}z^K}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-k_1(z)}{K(z)}\times\frac{1-L_{D_1}(b(z))}{b(z)}$$

$$D_{1,2}(z) = \mu\frac{\gamma(\lambda+\gamma)L_A(\lambda+\gamma)[q(z)-z]P_{0K}z^{K-1}}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-k_2(z)}{K(z)}\times\frac{1-L_{D_1}(b(z))}{b(z)}$$

$$D_{2,1}(z) = \mu\frac{(\lambda+\gamma)L_A(\lambda+\gamma)[\lambda(1-g(z))+-\gamma(1-k_2(z))]P_{0K}z^K}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-k_1(z)}{K(z)}\times\frac{1-L_{D_2}(b(z))}{b(z)}L_R(b(z))L_{D1}(b(z))$$

$$D_{2,2}(z) = \mu\frac{\gamma(\lambda+\gamma)L_A(\lambda+\gamma)[q(z)-(z)]P_{0K}z^{K-1}}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-k_2(z)}{K(z)}\times\frac{1-L_{D_2}(b(z))}{b(z)}L_R(b(z))L_{D1}(b(z))$$

$$R_1(z) = \mu\frac{(\lambda+\gamma)L_A(\lambda+\gamma)[\lambda(\mathbf{1}-\boldsymbol{g})(z))\gamma(1-k_2(z))]P_{0K}z^K}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-k_1(z)}{K(z)}\times\frac{1-L_R(b(z))}{b(z)}L_{D_1}(b(z))$$

$$\mathrm{R}_2(z) = \mu\frac{\gamma(\lambda+\gamma)L_A(\lambda+\gamma)[\boldsymbol{q}(z)-z)]P_{0K}z^{K-1}}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-k_2(z)}{K(z)}\times\frac{1-L_R(b(z))}{b(z)}L_{D_1}(b(z))$$

$$\mathrm{V_s}(z) = (1-r)\frac{(\lambda+\gamma)L_A(\lambda+\gamma)[\lambda(1-g(z))+\gamma(1-k_2(z))]P_{0K}z^K}{(1-L_A(\lambda+\gamma))\emptyset(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}\frac{1-L_V(b(z))}{b(z)}k_1(z)$$

4. Performance Measures

In this section, we deduce some performance measures of the system under the study state condition.

1. Distribution of the system size in terms of generating function is given by

$$\Phi(z) = \frac{\lambda(1-g(z))[q(z)(p-z)+z(1-p)]+\gamma(1-k_2(z))q(z)(1-z)}{(1-L_A(\lambda+\gamma))\phi(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]} \times \frac{(\lambda+\gamma)L_A(\lambda+\gamma)P_{0K}z^K}{\lambda p(1-g(z))}$$

Distribution of the orbit size interms of generating function is given by

$$\Omega(z) = \frac{\lambda z(1-g(z))[(1-z)+(z-q(z))(1-p)]+\gamma(1-k_2(z))q(z)(1-z)}{(1-L_A(\lambda+\gamma))\phi(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]} \times \frac{(\lambda+\gamma)L_A(\lambda+\gamma)P_{0K}z^{K-1}}{\lambda p(1-g(z))}$$

2. Distribution of the number of transit customers in the orbit in terms of generating function is given by

$$\Upsilon(z) = \frac{\lambda(1-g(z))[(1-p)(z-q(z))+(1-z)]+\gamma(1-k_2(z))(1-z)}{(1-L_A(\lambda+\gamma))\phi(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]} \times \frac{(\lambda+\gamma)L_A(\lambda+\gamma)P_{0K}}{\lambda p(1-g(z))}$$

Distribution of the number of transit customers in the system interms of generating function is given by

$$\Psi(z) = \frac{\gamma(1-k_2(z))(1-z)q(z)+\lambda(1-g(z))[q(z)(p-z)+z(1-p)]}{(1-L_A(\lambda+\gamma))\phi(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]} \times \frac{(\lambda+\gamma)L_A(\lambda+\gamma)P_{0K}}{\lambda p(1-g(z))}$$

For p = 1 , we obtain $\Phi(z) = z^K \times \Psi(z)$ where, $\Psi(z) = \Upsilon(z) \times q(z)$
Given that

$$E(L) = \lim_{z\to 1} \Phi'(z) \quad and \quad E(N) = \lim_{z\to 1} \Omega'(z)$$

then the results are obtained by using the Hospital's rule.

For the mean waiting time of customers in the retrial queue and in the system, they can be obtained respectively by the following formulas $W_{T_q} = \frac{E(N)}{\lambda p m_1}$ and $W_{T_s} = \frac{E(L)}{\lambda p m_1}$

Corrolary

1. The probability that the server is busy is given by

$$P_B = \frac{\lambda m_1\beta_{11} + \gamma\beta_{21}(1-\lambda p m_1\nu_1(1-r))}{1+(1+\phi_{11}+\eta_1+\phi_{21})(\lambda m_1\beta_{11}(1-p)+\gamma\beta_{21})+\lambda m_1\nu_1(1-p)(1-r)}$$

2. The probability that the server is under first verification delay is given by

$$P_{D_1} = \frac{\mu\phi_{11}[\lambda m_1\beta_{11} + \gamma\beta_{21}(1-\lambda p m_1\nu_1(1-r))]}{1+(1+\phi_{11}+\eta_1+\phi_{21})(\lambda m_1\beta_{11}(1-p)+\gamma\beta_{21})+\lambda m_1\nu_1(1-p)(1-r)}$$

3. The probability that the server is under repair is given by

$$P_R = \frac{\mu\eta_1[\lambda m_1\beta_{11} + \gamma\beta_{21}(1-\lambda p m_1\nu_1(1-r))]}{1+(1+\phi_{11}+\eta_1+\phi_{21})(\lambda m_1\beta_{11}(1-p)+\gamma\beta_{21})+\lambda m_1\nu_1(1-p)(1-r)}$$

4. The probability that the server is under repair given that the transit customer is in service position is obtained by

$$P_{R_1} = \frac{\mu\beta_{11}\eta_1\lambda m_1(1+\gamma p\beta_{21}(1+\phi_{11}+\eta_1+\phi_{21}))}{1+(1+\phi_{11}+\eta_1+\phi_{21})(\lambda m_1\beta_{11}(1-p)+\gamma\beta_{21})+\lambda m_1\nu_1(1-p)(1-r)}$$

5. The probability that the server is under repair given that the recurrent customer is in service position is obtained by

$$P_{R_2} = \frac{\mu\beta_{21}\eta_1\gamma[1-\lambda pm_1(\beta_{11}(1+\phi_{11}+\eta_1+\phi_{21})+\nu_1(1-r))]}{1+(1+\phi_{11}+\eta_1+\phi_{21})(\lambda m_1\beta_{11}(1-p)+\gamma\beta_{21})+\lambda m_1\nu_1(1-p)(1-r)}$$

6. The probability that the server is under second verification delay is given by

$$P_{D_2} = \frac{\mu\phi_{21}[\lambda m_1\beta_{11}+\gamma\beta_{21}(1-\lambda pm_1\nu_1(1-r))]}{1+(1+\phi_{11}+\eta_1+\phi_{21})(\lambda m_1\beta_{11}(1-p)+\gamma\beta_{21})+\lambda m_1\nu_1(1-p)(1-r)}$$

7. The probability that the server is on vacation is given by

$$P_V = \frac{\nu_1(1-r)\lambda m_1(1+\gamma p\beta_{21}(1+\phi_{11}+\eta_1+\phi_{21}))}{1+(1+\phi_{11}+\eta_1+\phi_{21})(\lambda m_1\beta_{11}(1-p)+\gamma\beta_{21})+\lambda m_1\nu_1(1-p)(1-r)}$$

8. The probability of customer's loss is given by

$$P_L = (1-p)[P_B + P_{D1} + P_R + P_{D2} + P_V]$$

These probabilities are independent of K The steady-state availability and failure frequency of the server are respectively noted by (A_v) and (F_f) and given by

$$A_v = \frac{\lambda pm_1[(1-q'(1))(\lambda+\gamma L_A(\lambda+\gamma))+\gamma\beta_{21}(\lambda+\gamma)L_A(\lambda+\gamma)]}{(\lambda+\gamma)L_A(\lambda+\gamma)[\lambda m_1(1-(1-q'(1))(1-p))+\gamma k_2'(1)]} + \frac{[(\beta_{11}(\lambda+\gamma)L_A(\lambda+\gamma)-(1-L_A(\lambda+\gamma)))(\lambda m_1+\gamma k_2'(1))]}{(\lambda+\gamma)L_A(\lambda+\gamma)[\lambda m_1(1-(1-q'(1))(1-p))+\gamma k_2'(1)]}$$

$$F_f = \frac{\mu\lambda pm_1(\lambda+\gamma)L_A(\lambda+\gamma)[\beta_{11}(\lambda m_1+\gamma k_2'(1))+\beta_{21}\gamma(1-q'(1))]}{(\lambda+\gamma)L_A(\lambda+\gamma)[\lambda m_1(1-(1-q'(1))(1-p))+\gamma k_2'(1)]}$$

Let T_b and T_c be respectively the length of a busy period and the length of a busy cycle, their expectations under the steady state conditions are given respectively by

$$E(T_b) = \left[\frac{(\lambda+\gamma)L_A(\lambda+\gamma)(\lambda m_1 q'(1)+\gamma k_2'(1))+\lambda pm_1(1-L_A(\lambda+\gamma))(\lambda q'(1)-\lambda(1-m_1)+\gamma k_2'(1))}{\lambda pm_1[(\lambda+\gamma L_A(\lambda+\gamma))(1-q'(1))-(1-L_A(\lambda+\gamma))(\lambda m_1+\gamma k_2'(1))](\lambda m_1+\gamma)}\right]$$

and

$$E(T_c) = \frac{(\lambda+\gamma)L_A(\lambda+\gamma)[\lambda m_1(1-(1-q'(1))(1-p))+\gamma k_2'(1)]}{\lambda pm_1[(\lambda+\gamma L_A(\lambda+\gamma))(1-q'(1))-(1-L_A(\lambda+\gamma))(\lambda m_1+\gamma k_2'(1))](\lambda m_1+\gamma)}$$

5. Stochastic decomposition

We note the probability generating function of the system size distribution for $p = 1$ is given by

$$\Phi(z) = \frac{(\lambda+\gamma)L_A(\lambda+\gamma)P_{0K}z^K[q(z)(1-z)(\lambda(1-g(z))+\gamma(1-k_2(z))]}{(1-L_A(\lambda+\gamma))\phi(z)-(\lambda+\gamma)[z-L_A(\lambda+\gamma)q(z)]}$$

Let $\chi(z)$ be the probability generating function of the number of customers in the orbit, given that the server is idle or busy serving a recurrent customers or under first and second delay of verification or under repair. Under this definition, it can easily seen that

$$\chi(z) = \frac{z^K P_{0K} + P_0(z) + zP_2(z) + zD_{12}(z) + zR_2(z) + zD_{22}(z)}{P_{0K} + P_0(1) + P_2(1) + D_{12}(1) + R_2(1) + D_{22}(1)}.$$

$$\Phi(z) = \Phi_0(z) \times \chi(z)$$

We obtain,

$$\Phi_0(z) = \frac{(1-z)q(z)(1-q'(1))}{q(z)-z}$$

6. Conclusion

In this paper we have studied an $M^{[X]}/G/1$ retrial queue with recurrent customers, geometric loss and Bernoulli vacation, where the server is unreliable, which consists of breakdowns period, repairs, and two delays of verifications. We have obtained the probability generating function of various states in the steady state solution with performance measures of the system. Some numerical examples were given to show the effect of breakdowns, repairs, two delays of verifications and size of batch arrivals in the system on the main performance measures.

References

1. Aissani, A. (1988). On the *M/G/1/1* queueing system with repeated orders and unreliable, J. Technology.6, 93-123.
2. Artalejo, J.R. (1999). Accessible bibliography on retrial queues, Math. Comput. Model 30, 1-6.
3. Artalejo, J.R., *Gomez´ − Corral*, A. (2008). Retrial queueing systems, Springer-Verlag, Berlin.
4. Atencia, I., Moreno, P. (2005). A single retrial queue with general retrial times and Bernoulli schedule, Appl.Math´emat.Comp 162, 855–880.
5. Choi, B.D., Kulkarni, V.G. (1992). Feedback retrial queueing systems, Stochastic Model Relat Field,pp 93-105.
6. Djellab, N.V. (2002). On the *M/G/1* retrial queue subjected to breakdowns, RAIRO Oper. Res. 36, 299-310.
7. Falin, G.I. (1990). A survey of retrial queues, Queueing Syst. 7, 127-168.
8. Falin, G.I., Templeton, J.G.C. (1997). Retrial Queues, Chapman and Hall, London.
9. Farahmand, K. (1996). Single line queue with recurrent repeated demands, Queueing Systems 22, 425–435.
10. Fayolle, G. (1986). A simple Telephone Exchange with Delayed Feedbacks, Teletraffic Analysis and Computer Performance Evaluation.
11. Gomez-Corral, A. (1999). Stochastic analysis of a single server retrial queue with general retrial times, Nav. Res. Logis. 46, 561-581.
12. Grey, W.J., Wang, P.P., Scott, M.K. (2000). A vacation queueing model with service breakdowns, Appl. Math. Model. 24 391400.
13. Kulkarni, V. G., Liang, H. M. (1997). Retrial queues revisited. In J. H. Dshalalow (Ed.), Frontiers in queueing: Models and applications in science and engineering (pp. 1934). Boca Raton: CRC Press.
14. Saggou, H., Sadeg, I., Ourbih-Tari, M ,Bourennane,E.B.(2018). Performance measures of *M/G/1* retrial queues with recurrent customers, breakdowns and general delays, Communications in Statistics-Theory and Methods,1-16.
15. Lee, H.S. (1995). Optimal control of the *M/G/1/K* queue with multiple server vacations, Comput. Oper. Res. 22 (5) 543552.

Improving energy access for Africa through regional integration

Xuan Che
United Nations Statistics Division

Abstract

Access to energy is one of the fundamental rights for all citizens. It plays a central role in Africa's development. Regional integration is a fundamental and irreplaceable factor in Africa's energy sector, as it helps reduce the risks for natural and man-made disasters, provides diversity on the types and geography of energy sources, and boosts national economy and fosters future growth. Using quantitative evidences and historical data, this study attempted to explain the effects of regional integration on energy production and access. It looked at how regional integration improves electricity production of the countries, investigated the progress made and challenges facing Africa's regional power pool initiatives, and provided policy advises on enhancing the regional market. Then, the effects of regional integration on electricity and clean cooking fuel accesses were studied. The progress made on the accesses to these energy sources were identified, and options were given on how to overcome the existing challenges and shortfalls. The concluding findings and discussions were provided in the end.

Keywords

Africa, energy statistics, energy access, regional integration, official statistics

1. Introduction

The Sustainable Development Goals (SDG) stipulate that it is one of the fundamental rights for all citizens to have access to clean, reliable, affordable and decentralized energy. The access to energy plays a crucial role from poverty reduction, food security, to health, education, equality, gender and climate change issues.

Africa has adequate energy resources for all her citizens. It has, however, remained a big challenge in many African countries to provide universal energy access. Regional integration among the countries can help streamlining each country's energy needs, provides alternatives and diversity of energy sources, and fuels the economy with future growth. Since oil and natural gas are concentrated mostly in the North and West Africa, with hydropower being abundant in East, Central and Southern Africa, creating a regional network of distribution and trade of energy can benefit all parties, both the providers and users, among the regions. We believe that regional

integration can help improve one of the most fundamental and basic measurements of energy access: the access to electricity.

Among all the development efforts involved in improving energy access, one question always surfaces at the centre: Can the energy access improvement keep up with the population growth in Africa? Africa is not only one of the fastest growing continents for its economy development and energy generation, it is also the region that experiencing the fasting growth of population. It is imperative that we look at both the energy access and the population growth at the same time in order to obtain a correct understanding of the country's situation of energy supply.

This paper first looks at how regional integration may improve electricity production of the countries, and provides policy advises on enhancing the regional market. Then, the effect of regional integration on electricity access will be studied, before concluding findings and discussions are provided in the end.

2. Methodology

Energy access is defined by the International Energy Agency (IEA) as "a household having reliable and affordable access to both clean cooking facilities and to electricity, which is enough to supply a basic bundle of energy services initially, and then an increasing level of electricity over time to reach the regional average" [1]. Energy serves as the backbone and driving force for economy development, industrialization, and urbanization. The production of energy, including electricity, is one of the key mandates to ensure sustainable development and national security.

We looked at historical and internationally comparable data from four main themes: the total electricity generation capacity, the growth of electricity generation compared to population growth, historical progression of the percentage of population with electricity access, and the urban-rural disparity of electricity access. Data were collected from the African Energy Commission (AFREC), International Energy Agency (IEA), U.S. Energy Information Administration; and the UN Population Division. Data from 1980 to 2016 were studied in order to create a temporal profile of the countries' energy access developments.

We recognized that no national electricity network will be complete without a sufficient network to distribute the electricity it stores. In order to maximize its positive impacts, regional mechanisms must be established to make sure households and other end-users can access to this resource. In this regard, we paid special attention to the effect of the five major regional power pools in Africa on the integration of data generation and transformation. The progress, challenges and solutions of improving access to electricity through unconstructive regional integration were studied.

3. Result

Africa's total electricity generation capacity in 2015 is estimated to be more than 125,000 MW [2]. There are still great disparities of this capacity between regions, with Southern Africa produces 45% of the total capacity, North Africa generates 39%, and the other three regions produce significantly less electricity power (8% for West, 5% for East, and 3% for Central Africa) [3]. With respect to the types of power, the vast majority (80%) are produced by fossil fuel, 16% by hydropower, and a mere 4% by solar, nuclear, and biofuels [3].

Table 1. Top 10 countries of electricity generation growth, 1980 to 2015.

Ranking	Country	Growth in electricity generation, 1980-2015
1	Burundi	7,566%
2	Benin	6,132%
3	Cape Verde	2,980%
4	Equatorial Guinea	2,400%
5	Mali	2,011%
6	Sudan	1,725%
7	Ethiopia	1,488%
8	Mauritania	1,340%
9	Congo	1,010%
10	Angola	942%

Source: U.S. Energy Information Administration

Fifty-two out of fifty-four African countries have witnessed increased electricity generation from the last 35 years (from 1980 to 2015). The only exceptions are Liberia and Sierra Leone. Among all, ten countries have gained electricity generation of more than ten folds (Table 1). The countries with smaller size, such as Burundi, Benin, Cape Verde, Equatorial Guinea, occupied the top 4 spots by achieving the speediest growth rate of electricity generation, all of them in 2015 produces more than 25 times of what they did in 1980. The larger economies, on the other hands, also fare pretty well, with Sudan, Ethiopia, and Angola all retain growth of more than 9 folds in the last 35 years.

Since most of the African economies are emerging markets that are developing and growing at phenomenal speeds, the energy pressure imposed upon will be ever challenging and pressing. With a young, rapidly growing population hungry for energy and trying to push into the middle-income level, the demands for electricity will also grow fast at the same time. Regional planning of electricity generation and power trading therefore become a crucial safety net to support the countries growth and meeting their energy demands. When the population growth outpaces electricity generation

capacity, countries need to look beyond borders to search for viable, affordable alternatives.

We investigated the per capita electricity production, observing that 10 countries have experience production grown more than 4 folds (Table 2): leading the way is Burundi (almost 30 times of growth of electricity production per capital in 2015 compared to 1980), followed by Benin (21 times), Cape Verde (16 times), Mali (8 times), Sudan (6 times), Ethiopia (5 times), Equatorial Guinea, Mauritania, Seychelles, and Mauritius (4 times).

Table 2. Growth of electricity generation per capita, 1980 to 2015.

Country	Electricity generation per capita, 1980 (kWh)	Electricity generation per capita, 2015 (kWh)	Growth, 1980-2015
Burundi	0.73	22.55	2,994%
Benin	1.34	29.46	2,090%
Cape Verde	52.33	866.94	1,557%
Mali	14.53	124.51	757%
Sudan	47.91	328.23	585%
Ethiopia	18.46	103.54	460%
Equatorial Guinea	66.58	361.58	443%
Mauritania	58.02	306.53	428%
Seychelles	709.01	3701.73	422%
Mauritius	456.51	2268.46	397%
...	...	...	...
Mozambique	1144.47	698.97	- 39%
Togo	18.38	10.62	- 42%
Zambia	1557.42	825.11	- 47%
Sierra Leone	55.86	24.18	- 57%
Liberia	451.20	66.67	- 85%

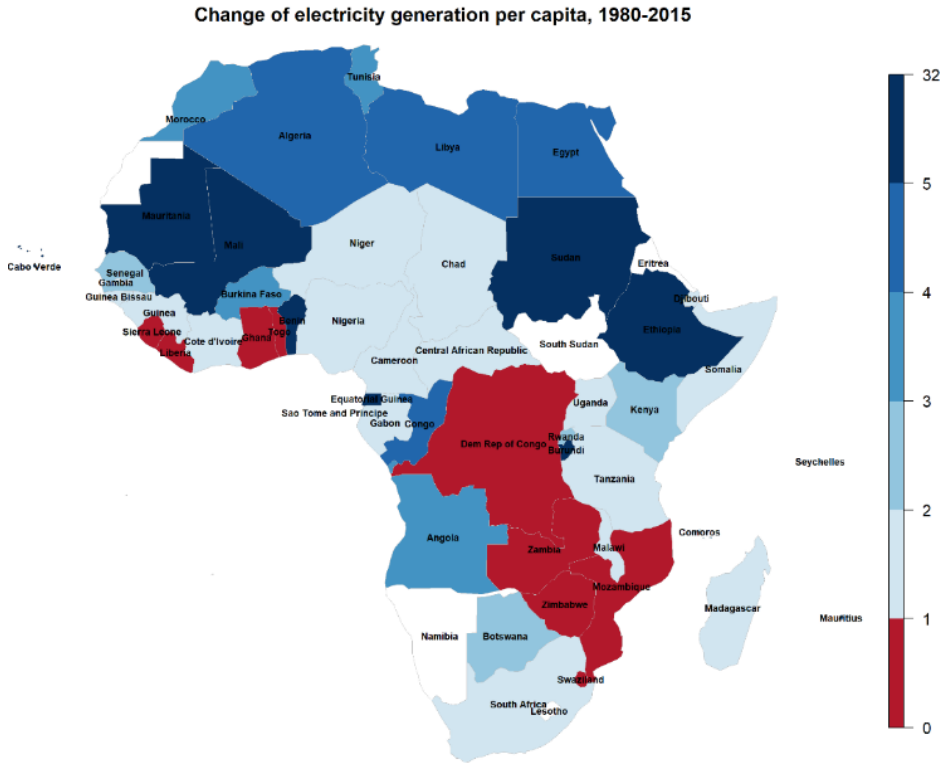
Source: U.S. Energy Information Administration; United Nations Population Division.

From Figure 1 we can see contrasting geographical trends in different regions: countries with smaller land cover areas, as well as North African and the Sahel countries (shown in darker blues), are generally faring well to elevate their production intensity. On the other hands, some notable West and Southern-Central African countries (shown in red) have experienced major difficulty in meeting the energy demands of their respective populations. Many of these slowing growing countries are neighbouring energy rich, fast growers. Regional electricity trade is not only possible but should be put as a priority. The enhanced collaboration and sharing through regional power pools might provide a timely and long-lasting solution.

Since 1950, major efforts have been put into developing regional power pools through bilateral and multilateral agreements. There are five major power pools in Africa now: Central African Power Pool (CAPP), Eastern Africa Power Pool (EAPP), Comité Maghrébin de l'Electricité (COMELEC), Southern

Africa Power Pool (SAPP), and West Africa Power Pool (WAPP). They are specialized agencies in their Regional Economic Communities (REC), and are tasked to facilitate and provide electricity trade within each pool. The power pools serve as an energy optimization and safety mechanism for their members. Among the five pools, CAPP is generating mostly from hydropower (77%), while the other four pools mostly rely on fossil fuels. Much infrastructural and political efforts have been put into these initiatives, and it is expected that they will play a greater role in creating uniformed energy markets and accelerated power trade among regional partners. For example, in SAPP alone, member countries have signed 28 bilateral agreements for their Short Term Electricity Market and Day Ahead Market. In recent years, the power pools not only work to enhance their internal structure, but have also started to collaborate on intercontinental levels. CAPP is also working with SAPP in creating 3,800 kilometres of electricity transmission lines, linking the two pools via Angola. COMELEC also works with Middle East and Europe, through Egypt and Morocco. It is estimated that, when the full energy integration scenario is achieved in 2040, the power pools would save Africa US$ 43 billion annually on their energy bills [4]. This is correct approach to take to improve Africa's electricity generation and distribution capacities.

Figure 1. Change of electricity generation growth per capita, 1980 - 2015

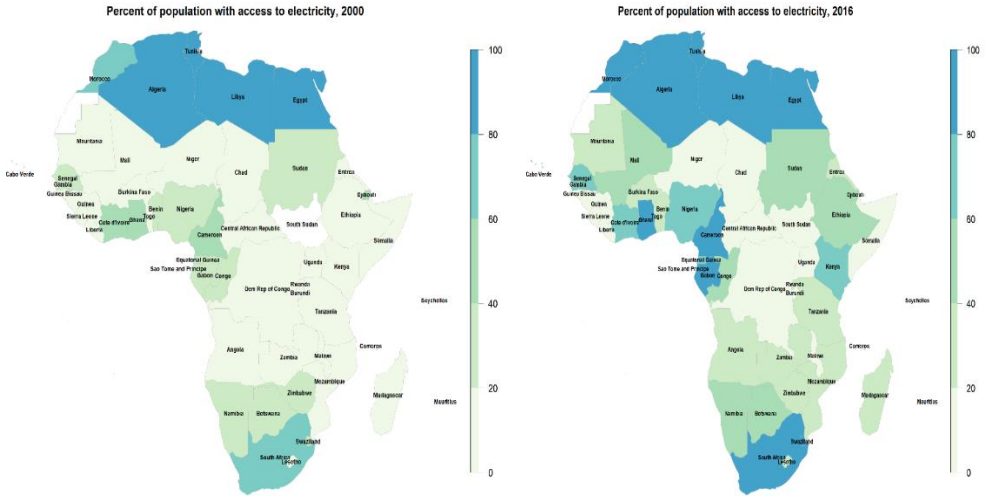


Source: U.S. Energy Information Administration; United Nations Population Division.

Africa is rich in energy resources, but lacks infrastructures for its access. According to the World Bank [5], over 1.1 billion people now lack reliable access to electricity. Although much of this population is living in Africa, over the last two decades, Africa has made tremendous progress on improving access to electricity for all. This includes major political commitment, investment and infrastructural projects which resulted in connecting many millions of households into the grids in some of the regions and areas. From Figure 2, it is clear to see that a lot of effort has been made when it comes to access to electricity between 2000 and 2016, resulting in encouraging progress. In many countries where access rates are under 40% in 2000, the rates have been consistently improved in the last two decades and they are enjoying 60% towards 100% access in the whole countries. Much of this progress can be attributed to improved infrastructure, careful planning, investment from the governments and private sectors, and an increase in oil production, and an ease on the trade for fuel products including national gas and petroleum over the years. The African energy market is benefiting from a stronger and tighter regional integration in the continent.

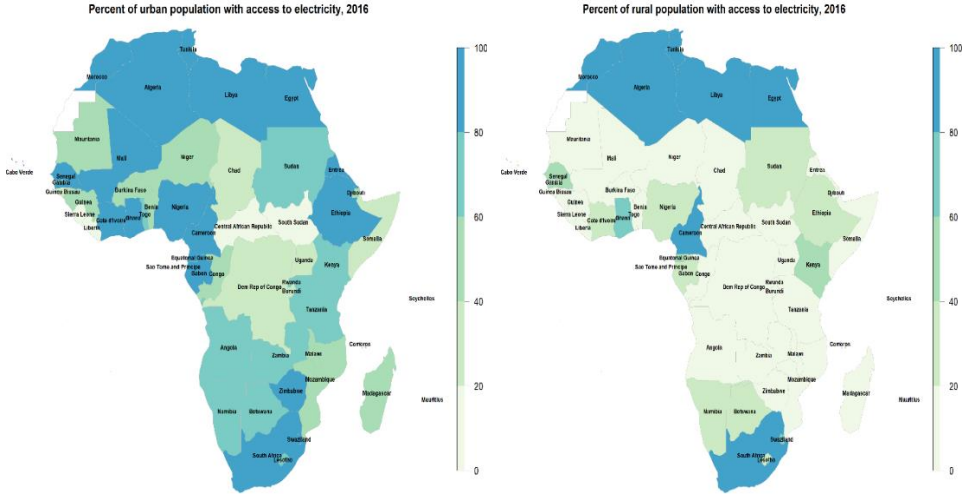
There are still gaps to fill in the African countries when it comes to energy access. Many countries have reached a threshold where their national energy structures are facing uphill difficulties and increased risks and costs to grow further to provide electricity access to all of their citizens. [6] Due to many geographical restraints, countries have realized that a national strategy of energy cannot simply overlook the energy sources from beyond their borders. In order to obtain the most affordable, sustainable and reliable energy, countries need to work together to create a regional network of the resources and facilitate the trade of energy and energy products. The national markets in Africa are of considerable different sizes – think South Africa and Lesotho, for instance. Collaborations between countries and integrated energy networks are in many times the best and most cost-effective method for countries to grow their energy access rates.

Figure 2. Percentage of population with access to electricity, 2000 and 2016



Source: http://www.iea.org/energyaccess/database/

Figure 3. Urban and rural access of electricity, 2016



Source: http://www.iea.org/energyaccess/database/

The need of energy integration is probably best illustrated by the electricity access challenge in urban areas. The gap of access to electricity between urban and rural populations is stark in almost all countries. From Figure 3, it can be seen that in almost all African countries, urban coverage for electricity has been improved and well. Many countries have achieved at least 60% access rate within urban settlements. In the meantime, disproportional populations who have no access to electricity reside in rural area. Across the vast countryside, many have only managed to provide access to less than 20% for the rural population. Many agricultural and rural areas are more reliant on energy provided to them to support all aspects of their livelihood, from living,

to irrigation, agricultural production, and they are also more fragile in withstanding natural or man-made disasters. This is one of the greatest challenges facing energy access for African countries, and where effort, resources and priority should be given.

The Africa Energy Sector Outlook 2040 projected that the average Africa GDP growth will be 6.2%, while the UNPD projected that Africa will grow its population as a rate of 2.3% per annum for the next 30 years [7]. Access to electricity is expected to increase by at least 60% for all African countries, which will benefit 800 million people who currently are out of access, with an estimated cost of US$ 3.5 billion per year [3]. With an average price tag of only US$ 4.3 per person a year, it is an investment well worth its value.

4. Discussion and Conclusion

A regionally planned energy network has a many significant benefits over national ones. It diversifies the energy resources, which make the network more secure and reliable for all its users. It can attract more private investors, which reduces cost for establishing these infrastructures by their hosting countries. It can adapt to economies of both small and large sizes, and provides affordable energy in a greater geographical coverage.

Regional energy infrastructure and integration can also help spearheading the development of renewable energy, which in many areas are abundant in rural areas. Perhaps one of the best examples of renewable energy regional collaboration can be found in Ethiopia. Ethiopia is rich in hydropower, with an estimate of 60,000 megawatts (MW) of potentials for electricity generation [8]. The majority of this huge potential of hydropower, where hydraulic dams will be built, is also located very close to the border of Sudan, South Sudan and Kenya. At this moment, the biggest project in the country, the Grand Ethiopian Renaissance Dam (GERD), with a 6,000 MW generation capacity, is more than 60% completed. Ethiopia has already started to export electricity to Sudan (up to 100 MW), Djibouti (up to 60 MW), and with deals constructed with Kenya, Tanzania and South Sudan as well [9]. These distribution systems can greatly improve the rural access rate of not only one country, but the whole East Africa region. Due to the temporal cyclical patterns of hydropower generation, a regional effort to coordinate and distribute the power is the best practice for improving access and utilization for all the regional partners involved.

It has become an African consensus that progressive planning and implementation of regional energy strategies can help boost electricity generation and provide an overall universal access network for electricity. By forming larger energy markets, the regional power pools reduce market fragmentation, attracts private and public investors, diversify the sources of electricity, and enable countries to collaborate more closely on other energy-related and economic activities as well. Regional integration is the key to

improving access of electricity, especially in rural areas where availability is still critically low. A regional collaboration can help secure the provision of energy supply to rural areas by allocating and delivering electricity from its nearest generation points. Such collaboration can also improve the environmental and health quality of people, especially women and children, promote gender equality, increase the sizes of labour forces, and foster a stronger and sustainably growing economy.

References

1. International Energy Agency (IEA), 2017, World Energy Outlook 2017 Special Report: Energy Access Outlook.
2. African Development Bank (AfDB), 2014, Africa Energy Sector: Outlook 2040.
3. African Energy Commission (AFREC), 2015, Africa Energy Database 2015.
4. United Nations Environment Programme (UNEP), 2017, Atlas of Africa Energy Resources.
5. World Bank Group, 2017, Sustainable Energy for All Database.
6. Bhagavan, M R, 1999, Reforming the Power Sector in Africa, AFREPREN, ZED Books Ltd., London, UK.
7. United Nations Population Division (UNPD), 2017, World Population Prospects.
8. United States Agency for International Development (USAID), Power Africa Ethiopia.
9. Ethiopian Electric Power (EEP). http://www.eep.gov.et/

The relationship between Income Inequality and Disparity Education and effect to achieve SDGs in Egypt

Ali Hebishi Kamel Abdelhamid, Rawia Wagih Abd ElMagid ElSayed Ragab
Central Agency for Public Mobilization and Statistics CAPMAS

Abstract

Egyptian society suffers from the gap in income distribution where the rich categories represent a small ratio of the population but have the large amount of income and the opposite for the poorest categories. It's the main obstacle for achieving inclusive development policies. On the other hand, the countries put a large amount of its budget to improve education. It's considered the base stone in inclusive development for countries. Policymakers usually justify higher educational spending as a highly effective tool for reducing income inequality. Theoretical studies suggest that the relation between education and income inequality is not always clear (Jong, 2002). The purpose of this paper is to investigate the relationship between education disparities and income distribution in Egypt. The study provide information by statistical analysis on household survey data (HICES) through applying Multiple regression and Binary logistic regression to study the effect of income and education on each other and using results to know if there have relation between inequality income and education disparities. It also concluded the gap in income leads us to disparities in completed education (secondary education in Egypt is main education) and the study finds that education disparities has positive impact on income inequality. Also provides useful insights on how the impact of education has changed over the years for a developing country like Egypt to achieve targets 4.1 of SDGs 2030.

Keywords

HICES (Household Income, Expenditure and Consumption Survey); Income Inequality; Disparities in Education

1. Introduction

Egyptian society suffers from the gap in income distribution which indicates that the rich categories that represent a small ratio of the population always have the large amount of income, while the rest of the population are distributed as the poorest categories always get the least ratio of income. This gab addressed the main obstacle for achieving inclusive development policies and social justice to help in developing and improving the standards of living for the society. On the other hand, Education contributes effectively in success of economic and social development. It also helps in skills improvement and

capacity building for individuals what leads to society's progress and prosperity. That's why countries put a large part of its budget to improve education. It's considered the base stone in inclusive and sustainable development for countries.

Expenditure on education is one of the most issues that countries is seek to enlarge its benefits where education process is not only to teach people how to read and write it seeks to increases individual's knowledge and experiences. Developing countries; Egypt one of them are suffering from decrease of ratios of adults who can read and write with low quality of education as a result low benefits of expenditure on education. May this occurs because the relation between the level of education and distribution of income. So; Education as one of the major factors affecting the degree of income inequality that Policymakers usually justify higher educational spending as a highly effective tool for reducing income inequality. Theoretical studies suggest that the relation between education and income inequality is not always clear (Jong – Wha Lee, 2002). The effect of education will be significant if the initial level of education attainment is lower and the expansion of education is relatively faster. Therefore, the countries that have higher initial education attainment levels tend to produce unexpected or insignificant results then income inequality increases with educational inequality.

While the study predicts a positive association between educational disparities, as measured by the variance of completing secondary education or not and income inequality, the effect of completed secondary education (the main education in Egypt) on income inequality may be either positive or negative, depending on the evolution rates of completed education.

The purpose of this paper is to investigate the relationship between education and income distribution, Also the level of education of household's head that have strong relation with his/her family which effects on income inequality, on the other hand the level of income that associated with education disparities (completed or not). Considering the ambiguous theoretical predictions about the relation between education and income distribution, and the paper investigates that income inequality increases with educational disparities according to subpopulation, that by study the relation between that level of education disparities has statistically significant effects on income inequality and also income inequality has related to education disparities by completed basic education and get good education or not, so in this paper will discuss the determination of education by get (good level education that mean completed fundamental education) according to education system in Egypt for achieving the fourth goal of the sustainable development goals by answer these questions:

- Level of education of households head is associated with increased income of household or not in Egypt?
- How does income of household influence on completed education of individual?
- The relation between income inequality and education disparities and effect on achieving SDGs in Egypt?

2. Literature Review

Rahman.T, Taposh.D (2015) paper's discussed the relation between Income Inequality and education inequality using household level for Bangladesh, when Income Inequality has always been at most important issue in economic research and public policy debate. The study finds that education inequality has positive and statistically significant impact on income inequality in Bangladesh. Although the result shows that the effect of education inequality on income disparity is not the biggest one, however it is more realistic approach for Bangladesh to redesign the education system, compared to other available routes, such as, wealth redistribution-which is impractical and also, might create social unrest if tried. These statistical findings can be treated as a starting point in discussing related policies.

Abdul Jabbar Abdullah (2011) this paper provides a comprehensive review of the extant econometrics literature through a meta-regression analysis (MRA) of 64 empirical studies that collectively report 868 estimates of the effects of education on aggregate inequality. The aim MRA is:

Assess the effect of education on inequality. Does education increase, decrease, or have no effect at all on inequality at the national level? Under what conditions does education shape national inequality?

The paper concluded that education appears to have its greatest effect on the two tails of the income distribution, reducing the income share of the rich and increasing the income share of the poor. Hence, we can conclude that education reduces the gap between the rich and the poor. Education appears to have no effect on the share of the middle class. It also concluded that the more unequal is the distribution of education the greater will be income inequality. Hence, it appears that it is important to ensure a fairly equitable access to education. Inequality in education widens income inequality. Education has a larger negative effect on inequality in Africa. The paper indicated that education reduces the gap between the rich and the poor.in addition to the education is an effective tool for reducing income inequality, so the distribution of education is important. Some of the results also indicate that the level of secondary education appears to be more important in reducing inequality than primary schooling does. There are some important regional differences in the effects of education.

A study of Brazil in 1977 revealed that higher income earners enjoyed greater benefit from investment in education since their children had better educational opportunities compared to those from lower income groups (World Bank, 1977).

3. Methodology

Data Sources: This research used data of Household Income, Expenditure and Consumption Survey 2015 and that survey conducted each two years. It collected data about economic, social and demographic characteristics of individual and households; in addition to their consumption, expenditure and income patterns in Egyptian Society and also to identify the size of the gap to help decision and policy makers to consider this information in planning. Sample size is 25000 household distributed as 45% in urban and 55% household in rural, allocated over all governorates (urban/rural) in proportion to the size of each governorate. The data has Decoding, classification, editing, validation and weight.

Software tools: The analysis done by R for statistical analysis program

Research Steps:

- Firstly, test the relation between total household income and education level of household head by using multiple regression models.
- Secondly, test the relation between Completed secondary education of individuals and total household income by using binary logistic regression model.
- Finally, bring the results from step 1 and 2 to know the relation between income inequality and disparities in education and the effects on each other.

Variables used in the study:

Variables	Labels	Codes
Tot_Inc	Total net annually income of household	Continues variable As Numeric
HH_sex	Gander of HH	Male=1,female=0
Hh_edu_lev	Education level of HH	Completed secondary=1 Under secondary=0
Urbrur	Place of residence	Urban=1,rural=0
HH_SIZE	Size of household	Continues variable as Numeric
HH_Disable	Head of house hold have disability or not	Disable=1, Non-disable=0
HH_age	Age of HH	Continues variable As Numeric
Edu_type	Type of education	Public=1, private=0
Gander	Gander of individual	Male=1, female=0
Poor	Poverty status of individual according to Egypt national poverty line	Poor=1, Not poor=0
M_EduSt	Level of education of individual's mother	Completed secondary=1 Under secondary=0

Edust	Indicator variable of education for Logistic Regression	Completed secondary=1 Under secondary=0

Table (1) Variables which used in the study

3.1 Relationship between household income and education level of household head

A simple regression model used to reflect the relationship between income and education; it was positive which means when the education level of household head move from one category to the next (from uncompleted secondary education to completed secondary education) then household income increases. The same result appears when multiple regression used by adding other variables.

Multiple Regression Model

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots\ldots\ldots\ldots + B_kX_k + e$$

Log (Income) = 9.9337475 + 0.1380670*hh_edu_lev + 0.0465058 HH _SIZE + 0.0004238*HH_age + 0.1004928*Urbrur + 0.1017351*HH_sex - 0.4097318*HH_Disable + 0.0131579*hh_edu_lev1:HH_age + 0.0006373*H_SIZE:HH_age + 0.0755987*hh_edu_lev1:Urbrur1 + 0.0034694*HH_age:HH_sex1 + 0.0559655*HH_SIZE:HH_Disable1.

All independent variables in income model are statistically significant relationship with dependent variable log (income) expect age of HH and all interaction between independent variables are statistically significant.

- B_i represents the difference in the predicted value of log income of HH for each one-unit difference in X_i, if the rest of independent variables remain constant.
- B_1 = 0.138067 represents the increase in log income value When education level of HH changes from uncompleted secondary to completed secondary, if the rest of independent variables are constant, the same positive relation of simple regression. And it reflects the most effective variable on income in this model that reflect the strong relationship between income and education.

3.2 Relationship between individuals' education level and household income

How does income of household influence on completed education of individual?

An indicator variable created "Educst" to express education level of individuals because secondary education in Egypt is considered to be the good or enough level required to have a good quality education where value 1 indicate that

individual have completed the secondary education or higher and value 0 indicate that individual is not educated or under secondary education.
Education is the main variable in our study since higher levels of education means higher human capital and thus higher income level.

Binary Logistic Regression

[Z= 10.29154 +0.93101*lnTot_Inc + 3.02868 M_EduSt1 -5.08226 poor1 +1.82298 gander -2.50622 Urbrur1 -0.55383 lnTot_Inc:poor1 +0.37582 poor1:gender1 -0.91931 M_EduSt1:gender1 + 0.28444 lnTot_Inc:Urbrur1 -0.24310 gender1:Urbrur1 + 0.22899 lnTot_Inc:gender1]

Where $Z = \log\left(\frac{\pi}{1-\pi}\right)$ where π is completed secondary education, The Predicator here is log-transformed of income so we can conclude from the model that for one unit income increase is associated with an increase in odds of being completed secondary education than being under secondary completed. All independent variables have significance level then their parameters are different from 0. The parameters with significant negative coefficients decrease the likelihood of that response category (completed secondary education) with respect to the reference category. Parameters with positive coefficients increase the likelihood of that response category.

B_0 can be interpreted as the change in probability of being completed secondary education is equal Z, if all independent variables = 0 in the model. All independent variables in Education level model have statistically significant relationship with the odds of dependent variable (Education level). Also all interactions between independent variables are statistically significant.

B_i represents the difference in the probability of predicted variable (completed secondary education) for each one-unit difference in X_i, if the rest of independent variables remain constant.
B_1 = 0.93101 represents that for one unit increase in log income value will implies an increase of the log odds ratio of education status completed secondary education versus under secondary education by 0.93101, if the rest of independent variables are constant. Also this model reflects the strong relationship between income and education.

3.3 The relation between ineqality income and disparities education

Is there a relation between inequality income and disparities education? Disparities are differences. Because we are statisticians, we are interested in differences that are statistically significant, or statistical disparities. So Disparity analysis will focus on subgroups according to: Gender, Education status, place of residence……etc.

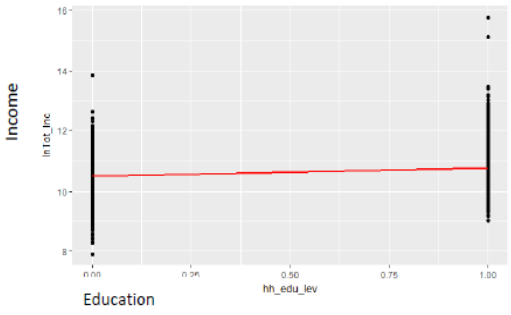
Inequality may refer to either difference within a group (such as income inequality). Disparity specifically refers to differences between groups (or between a group and a reference value). In this study we will focus difference on income and completed education, so it is the answer for the third research question. Also how the disparities in level of education by completed and uncompleted go to disparities in their incomes.

4. Result

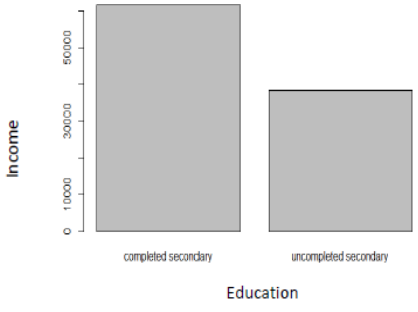
- We used log-transform for income instead of total income to express more variation in models.
- All simple regression models for independent variable included in the income model is statistically significant separately with household income (log income) the same for education model (binary logistic) which reflects the nature of Egyptian population.
- All independent variables in Multipule Regression model of income reflects statistically significant relationship between log(income) and each of education level of HH, household size, place of residence, disability status of HH, sex of HH and Age of HH with different values of level of significant (1% , 5% , 10%) and have positive relationship between income and all independent variables and interactions expect HH disablity status.
- For household income model the best fit model for simple regression when independent variable is education level of HH because it has the lowest AIC (the tool of control the best fit model) and High adjust R^2; Also the best model for Education that has the relation between level of education individual (completed secondary and under secondary) and household income with the lowest AIC and that reflect strong relationship between income and education.
- Higher Level of education of HH increases household income (dependent variable), because if HH completed secondary education that have income higher than under secondary education as (independent variable) alone in simple regression model (Figure(1)). And also when we add some independent variables for the model the level of education of HH have high effect on income from the results of this model (the least AIC in the model), So it answer the first research question "Is THE RELATION BETWEEN INCOME OF HOUSEHOLD AND EDUCATION?".
- In Education model (Binary logistic regression): there is a positive relationship between Completed secondary education and income and all other indepandant variables such as gender, place of residence and mother education except HH poverty status and interaction included that variable.
- Household Income influence on completed education of individual, because when we have increase in income there is an increase in the odd

ratio of completed secondary education of individuals. Also it has high effect of binary logistic regression of completed secondary education with other independent variables such as gender of individuals, HH-size, place of residence and mother education status according to results of the model. That conclusion answers the second research question "Is THE RELATION BETWEEN EDUCATION AND INCOME OF HOUSE HOLD?"

- From first and second analysis there is a relationship between income inequality and education disparities. So when HH moves from under secondary education to completed secondary education the average of income increases to 57305.2 from 38952.48 LE. according to the model. which means the increase in average income according to completed secondary education compared to under secondary education is 18352.73 with 32.03%.
- The inequality in income level related to disparities in education is 9% of low-income levels have individuals completed secondary education compared to 30 % of high-income levels have individuals completed secondary education. On the other hand disparitie in education related to income is 14.7% of completed secondary education individuals have with low household's income compared to 51.2% of completed secondary education individuals with high household's income (Figure 2).



Figure(1) Positive relation between education and income



Figure(2)Disparities between income and education

5. Discussion and Conclusion

This paper examines the relationship between income and education and concludes that income has a significant impact on education disparities where highly income households have better opportunities for its individuals to complete required education (indicated as secondary education in this study), also the greater the inequality in the distribution of education, the greater the inequality of income. So education has a strong impact on income than other variables. We found similar results from previous studies that income and education have importance to each other. This paper support the idea of improving education as the most effective tool for achieving inclusive

development for developing countries especially in Egypt to reduce the gap in income distribution. Any society can achieve sustainable development goals only through knowledge and experiences gained by better and high quality education. Some findings also suggest that the level of secondary education appears to be more important in reducing inequality than primary education. When we gradually remove potential channels that affect the link between income and education. We reveal a significant positive relationship between income and education. As expected, income has a positive relationship with education. Higher per capita income levels may increase, which in turn may lead to higher levels of education by increasing expenditure on education resources. Results of this study can be considered a useful reference point in policy discussions on this issue.

Future Works

The study can provide useful insights into how the impact of education to achieve SDGs (Sustainable Development Goals) 2030. The goal of education is to ensure universal and quality education for all and to promote lifelong learning by providing information to policymakers (IV) Ensure that all girls and boys complete free, fair and good primary and secondary education leading to effective educational outcomes that can be achieved by reduction of education disparities. (4.5.1) Eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for vulnerable groups, including persons with disabilities, indigenous peoples, and children in situations of vulnerability) to achieve this goal its necessary to provide equal opportunities for all to access good and quality education the study indicated the most effective variables on education disparities which is income. Two survey questions are suggested to be added to the HIECS survey to provide indicators for indicator (4.1.1) by adding a subjective question to education section to measure the quality of education as kind of disparities in education opportunities by estimate the least required level of reading and mathematical operations as follow:

Q1: how can you evaluate your level in reading?

1- Can't read at all. 2- Can read with difficulties.
3- Can read correctly. 4- No difficulties in reading.

Q2: According to your last level of education how can you evaluate your ability to do mathematical operations and calculations?

1- Can't do any operations at all. 2- Can with difficulties.
3- Can do with accepted level. 4- No difficulties at all.

References

1. Rahman,T., Taposh, D. (2015). Relationship between income inequality and education inequality: Evidence from Bangladesh: Research study by former students, Department of Economics, Carleton University, Ottawa, Canada.
2. Zuhal Zeynep Yildirim (2013). Assessing the Effect of Education on Income and Fertility in China: Honors thesis, Brown University.
3. José De Gregorio,Jong–Wha Lee (2002). Education and Income Inequality: New Evidence from Cross-Country Data: investigate on how education is related to income distribution in a panel data set covering a broad range of countries for the period between 1960 and 1990. Review of Income and Wealth Series 48, Number 3, September 2002.
4. J. B. Knight and R. H. Sabot (1983). Educational Expansion and the Kuznets Effect, American Economic Review, Vol. 73, No. 5 (Dec., 1983), pp.
5. A. Abdullah, H. Doucouliagos, E. Manning. Education and Income Inequality: A Meta-Regression Analysis. Journal of Economic Surveys 29 (2), 301-316.
6. World Bank, 1977, Basic Education and Income Inequality in Brazil: The Long Term Review, *World Bank Staff Working Paper* No. 268. World Bank: Washington.
7. CAPMAS (Central Agency for Public Mobilization and Statistics): Household Income, Expenditure and Consumption Survey: Volume 1, Survey Methodology. Sep. 2016.

Integration of economic establishments data into a uniquely identified comprehensive frame in Egypt

Rawia Wagih Abd ElMagid ElSayed Ragab, Ali Hebishi Kamel Abdelhamid
Central Agency for Public Mobilization and Statistics CAPMAS

Abstract

Censuses aim to form a statistical frame that includes the necessary requirements of variables for demographic, economic and social surveys and researches. The Economic Census aims to provide complete picture of structure, characteristics of various economic activities. This paper discusses how to overcome data duplication and have only one data source by creating a unique code (ID) for each establishment. Firstly, integrating the frames of Egyptian economic census (8 frames) to form the Comprehensive Economic frame. Secondly, configuring the Main Economic Aggregated Frame (MEAF) from both CAPMAS regular statistics and Ministry of Investment statistics. Thirdly, merging and comparing the comprehensive economic frame with (MEAF) to generate The Master Establishment Frame (MEF) with ID for each establishment. The paper solves the overlapping in frames, create the MEF with simple techniques; overcome the problem where more than one label for single establishment to get a comprehensive frame with the scientific bases.

Keywords

MEF (Master Establishment Frame); Main Economic Aggregated Frame (MEAF); ISIC4; GIS; Economic Census

1. Introduction

Census aim at a statistical frame that includes the necessary requirements of variables for preparation of demographic, economic and social surveys in addition to using GIS in the preparatory work of future censuses which depends on this frame to achieve the integration and quality of data that include the most demographic and economic variables for future work. Surveys are important methods to obtain indicators and characteristics about any specific phenomena, criteria, people or even any region in the world. The Economic Census provides contributions to the private sector in economic and social development. Also, expand of economic database through the creation of a comprehensive economic frame is necessary for the extraction of various economic indicators for CAPMAS periodic bulletins and different economic researches in Egypt. It also indicates how various economic activities contributes to GDP and the limitation of small establishments that are not included in the regular statistics. The overall objectives of the Economic

Census are to provide economic statistics on production input, output, capital formation, fixed assets, value added, employment and wages, and other financial information. This Economic Census covers all economic activities and will be based on a large sample of establishments, given that resources are not available for 100% enumeration. The first economic census conducted in its overall concept in 1992/1991 then the second census was carried out in 1996/1997 then 2000/2001, the last census (the fourth census) in 2012/2013, finally, currently the fifth economic census 2018/2019 is ongoing its field work.

The Economic Census has valuable data that can be used for developing an effective frame, and also as a source of information for the larger long-term establishments. The overall strategy for effectively covering the establishments in all economic activities will be to first identify a list frame of all large and important establishments that are not covered by the regular statistics. The smaller establishments and the larger establishments missing from the list frame and not included in the regular statistics will be covered. Prior to designing the area sample, it is important to understand the coverage of the economic activities in each of these frames during the regular statistics program. CAPMAS obtains financial information for all public sector institutions, and large establishments for various economic activities including, manufacturing, mining and quarrying, production and distribution of electricity and gas, transportation and storage, financial activities and other services. The economic units covered by the regular statistics are generally the large and important establishments, although enterprises are considered the unit of enumeration for some of the economic activities such as financial services. The important economic units included in regular statistics are selected based on criteria such as economic sector (public/private), legal status and size.

The main objective of the Economic Census is to complement the information of the regular statistics with representative sample data for the remaining economic activities of establishments of all size groups, in order to have complete information on the Egyptian economy that are required for national accounts, and analysis for policy and programs development. During the regular statistics program, CAPMAS obtains financial information for all public sector institutions, and large establishments for various economic activities including, manufacturing, mining and quarrying, production and distribution of electricity and gas, transportation and storage, financial activities and other services. The economic units covered by the regular statistics are generally the large and important establishments, although enterprises are considered the units of enumeration for some of the economic activities such as financial services. The important economic units included in regular statistics are selected based on criteria such as economic sector (public/private), legal status and size. It is necessary to expand the economic

database through the creation of a comprehensive economic frame for the extraction of various economic indicators.

Therefore, this paper will discuss how to overcome the duplication of data and having only one data source by creating a unique code (ID) for each establishment with generating the Master Establishment Frame (MEF) because CAPMAS conducts many different regular establishments surveys and each survey has its own frame and the total number of related overlapped frames is 68 frames for 180 bulletin. We found out the problem of data duplication and this will be solved using MEF then integrating the frames of the Egyptian economic census (8 frames), which was made by sampling technique, is the first part and the second part is composed of two parts. Firstly, creating the inner frame from each of CAPMAS regular statistics frames and then merging and comparing with these frames. Secondly, creating the external frame from the Ministry of Investment's statistics, the Federation of Industries and Chambers of Commerce then configuration of the Main Economic Aggregated Frame (MEAF) and doing merge and compare with these frames. Finally, generating the Master Establishment Frame (MEF) with performing a merge between two parts and then eliminating the overlapping between these frames is the best way to solve the problem of data duplication because we have the same establishment with more than one name or label although it has a problem because after generating MFE it will have nearly 100000 establishments that have a lot of cost to validate the frame with fieldwork and must update the MEF permanently because it is the main source of data so that any operating establishment will be registered in the Ministry of Investments. With regard to closed establishments, we do not have any information about that except only from field work that they have a lot of money.

In the future, the most effective sampling frame for the Economic Census and annual economic surveys will be a complete business register, with information on 4 or 6-digit international standard industrial classification (ISIC) codes, number of employees, capital investment and other available size measures, as well as identification details such as name of the establishment, type of ownership or corporation, address, telephone numbers, e-mail address and website (if available). In addition, there should be unique identification codes and information to link enterprises with component establishments.

CAPMAS has already considered developing this type of business register, and has explored different sources of information for compiling this type of sampling frame in the future and to achieve that CAPMAS has generated the MEF. This indeed will have a great effect on the quality of the statistical data and work especially that CAPMAS is the main source of data which affects the economic and financial policies and their future trends.

2. Methodology

The paper will show the duplication of data problem and how to solve it by generating the MEF. Because there are different frames of establishments are overlapped and the same establishment exists in different frames in addition to having more than one source of data. Therefore, CAPMAS seeks unified sources of data to become the main source of data because it is the national statistical office and deletes any duplication in data of establishments and covers all activities (public sector – governmental sector – private/ investment sector); so the steps to achieve that with generating the MEF by creating a unique code for each establishment. This can be done as follows:
Firstly: integrating the frames of the Egyptian economic census (8 frames) which was made by sampling technique because plans for the 2013 Economic Census were based on having a sample size of up to 180000 establishments, in order to provide reliable results for as many four-digit ISIC activities as possible with creating the comprehensive economic frame: using seven frames to ensure a comprehensive coverage of all establishments in the private sector as follows:

1. Area sample

The area sample design is used to cover small-sized establishment, the area sampling frame of EAs that was used for the 2013 Economic Census was based on the data and cartographic information and GIS from the 2006 Census of Establishments and has been selected in two stages:

a. The first stage

- Selecting a sample size of 1,800 counting area (EA) from the 2006 census facilities to be allocated proportionally among the governments according to any number of small establishment in each government taking into account the division of government to urban and rural areas. After that, the sample of small governments is increased with a minimum of 40 counting area and decreasing the sample of big governments. Next, the selection process is done with probability proportional to size (PPS) taking into account the relative importance of economic activity as a measure of the size of the counting area. The field implementation in selected regions of the sample is necessary to prepare a list of all of the establishments in the area in 1800 (sample areas as well as to determine the geographic information and data of economic activity and the number of employment) for each establishment.
- After the completion of the field work and the collected data have been electronically recorded, and after the completion of the full process, a sample of second stage is selected.

b. The second stage

The sample of establishments was selected at the level of economic activity in which each tabular category (ISIC4) from each Counting area has been selected in the first stage with: Selection of all establishments that employed five employees and more for all counting areas of the first stage and would be deleted if found duplicated with regular statistics establishments or duplicates with other frameworks used in the census. The other remaining establishments in areal sample with fewer than 5 employees are class-divided according to category tabular economic activity (ISIC) then a sample of 10 establishments was selected with equal probability, and if the number of establishments in tabular category is less than 10 establishments, then all establishments in the sample would be selected.

2. Rare activities

Comprehensive inventory of economic activities including a number of establishments less than 500 establishments depending on the 2006 Establishment Census data. The purpose of that is to cover all properties of the economic activities at the level of 4 digits ISIC4.

3. Large establishments that are not included in the regular statistics

A comprehensive inventory of establishments which employ 10 employees and more that are not included in the regular statistics depending on the 2006 Establishment Census data.

4. Education Activity

- Comprehensive inventory of schools based on data provided by the Ministry of Education and Al-Azhar.
- Comprehensive inventory of the universities based on data provided by the Ministry of Higher Education.

5. Health Activity

The sample size was selected with 14,000 frames (representing 20% of the clinics and medical centres and labs). The source is the Ministry of Health.

6. Large establishments operating in the wholesale and retail trade

The establishments which includes the following legal entities:

A- Contribution. B- With limited liability. C- Limited by shares. D- Branches of foreign companies.

The source is the Ministry of Social Solidarity database.

7. Activity exploitation of mines and quarries

Comprehensive inventory of establishments that work in the field of mining and quarrying and are not included in regular statistics and the source is the governments and the Ministry of Local Development frame as a supplementary framework.

The regular statistics data are added to the corresponding data in the census to reflect the outcome of the economic situation within establishments in all of the public / business public sector and the private sector and that is the eighth frame and any duplication for each establishment is deleted before creating the Comprehensive Economic Frame.

Secondly: configuration of the Main Economic Aggregated Frame (MEAF) from each of regular statistics and bulletins of CAPMAS (inner Frame) and Ministry of Investment statistics, the Federation of Industries and Chambers of Commerce (external frame)

The Inner frame: It is created from the regular statistics data and from the data of CAPMAS departments and bulletins (130 bulletins) that have 68 frames with overlapping and merging between them and frames of regular statistics and then they are compared to delete any duplicate in the data to set the inner frame.

The external frame: It is created from data provided by the Ministry of Investment (any establishment is operating according to the records of the ministry), The Federation of Industries and Chambers of Commerce then configured to obtain any additional data. Next, MEAF is configured, merged and compared with these frames (inner and external frames) and any duplicate in data is deleted.

Finally, The Master Establishment Frame (MEF) is generated through a merge between the Comprehensive Economic Frame and Main Economic Aggregated Frame (MEAF) and thus solving the overlapping between these frames is the best way to overcome the problem of data duplication.

Having only one frame without duplication of data and creating unique code (ID) for each establishment and having only one source of data which is CAPMAS, then the MEF would be the effective frame with the following structure.

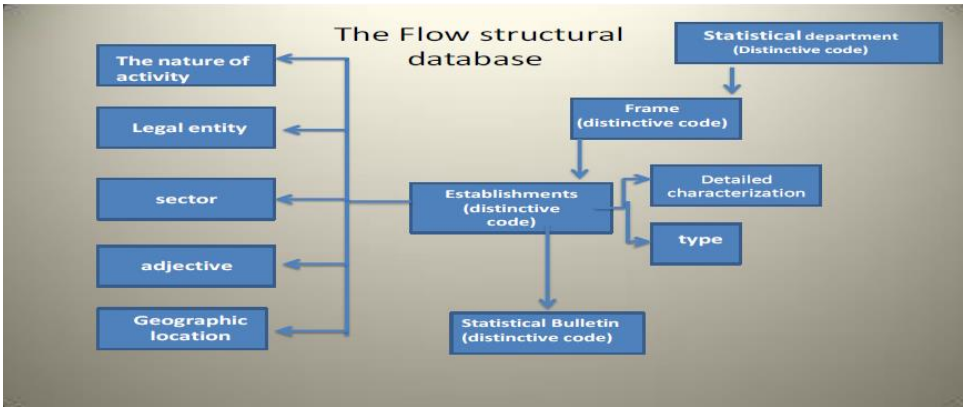


Figure (1) - the flow structural database

3. Result

1. Using Microsoft Excel (2010) to delete any duplication of data with sorting data according to the government code, activity group and establishment name/ label, example is showmen in Figure (2).
2. After deleting any duplicate from the previous step, we create unique code for each establishment using oracle developer 10g.

Address	Name of establishment (label)	economy activity	Act_ code	gov_ name	government code	serial
33 a Ramssus st.	EL Gowhara for international industries	Mobilization of Tea	10	cairo	1	3862
33 a Ramssus st.	EL Gowhara for food industries	Biscuit	10	cairo	1	2918
8 st 6 october in gasr el swez	EL SHARKIA for the confectionery industry Couvrtina	Manufacture of food products	10	cairo	1	9858
8 st 6 october	EL SHARKIA for the confectionery industry Couvrtina	Chocolate	10	cairo	1	3773
46 Abdel Khalak Tharwat st.	EL_Arab for Food - Groppi	Manufacture of food products	10	cairo	1	9863
46 Abdel Khalak Tharwat st.	EL_Arab for Food - Groppi	dessert	10	cairo	1	3826

Figure (2) - database of establishment frame

4. Discussion and Conclusion

The generation of the Master Establishment Frame (MEF) has many effective results on the quality of the statistical data and the field work as follows:

1- This technique reduces time, effort and money along with including the data of the surveys and researches in the same questionnaire for each establishment.
2- The interview is conducted only once for each establishment at a certain time to collect data.
3- Reliance on a unified source of data which is CAPMAS since it is the national statistics bureau.
4- Helps in generating complete business register for establishments in CAPMAS.

Challenges:

The constant updates of the Master Establishment Frame need a lot of cost because the updating process is done with the field work to have one frame, so that any operating establishment is registered in the Ministry of Investment but the challenges are in the cases of closed establishments and how would they be covered during the field work and how much of CAPMAS budget should be allocated for it.

Conclusion

The duplication of data has a lot of problems that affect the quality of the outcomes of the statistical work. It is therefore imperative to solve this problem by using a unified source of data (CAPMAS) and the frame (MAF) for the collected data. Such a mechanism assists in overcoming the overlapping frames problem where data are collected only once from each establishment. As a result, this will provide a lot of time, cost and effort and consequently increases the quality and accuracy of data and improve the impact on the statistical and economic decisions to better create solid policies for future developments.

Future Works

The methodology discussed in this paper will be used to form the MEF which will be updated using data from explained sources also the results of the currently ongoing economic census 2018/2019 with taking into account changes in sampling frames used where only 4 frames are used from General Census of Population, Housing and Establishment 2017:

- Small size establishments with less than 5 employees.
- Middle size establishments with 5-9 employees.

- Inclusive census for large scale enterprises with 10 and more employees and rare activities where the whole activity contain less than 30 establishments.
- Incomplete description establishments with missing data.

References

1. Megill, "Practical Sampling Guidelines for Economic Surveys in Developing Countries" Sampling Consultant, UNIDO, September 2007
2. The Arab institute of training and research, (2005),"Statistical dictionary in sampling".
3. CAPMAS website, available at http://www.capmas.gov.eg.
4. (2009),"Metadata Common Vocabulary".

Effect of the ocean heat content on the global sea ice extent using Fuzzy Logic Approach

Pranesh Kumar[1], Jiefei Yang[2]
[1]University of Northern British Columbia, Prince George, BC, Canada
[2]Dalian University of Technology, Dalian, P. R. China

Abstract

Statistical linear regression has been used in almost every field of science. The purpose of regression analysis is to explain the variation of a dependent variable in terms of the variation of explanatory variables. The classical linear regression has crisp coefficients and is bounded by some strict assumptions about the observed data, that is, the unobserved error terms are mutually independent and identically distributed. Fuzzy linear regression (FLR) was first introduced by H. Tanaka in 1982, which includes a fuzzy output, fuzzy coefficients and a non-fuzzy input vector. Some strict assumptions of the classical linear regression models are relaxed. This paper describes the fuzzy logic approach to fit the response surface model, analysis and the implementation of chosen method by using the global sea ice extent and ocean heat content data from 1979 to 2015. We have calculated the upper and the lower bounds of sea ice extent and carried error analysis which clearly indicates the comparative performance of fuzzy regression over the ordinary least squares regression. Besides, the width of predicted intervals of fuzzy regression model is much smaller than that of ordinary least squares model, which indicates the superiority of fuzzy regression methodology.

Keywords

Fuzzy logic; Least squares regression; Modelling; Climate change

1. Introduction

Crisp data, also known as precise data, are very common in everyday life. The traditional science and technology pursuit for certainty in all its manifestations and almost all the mathematical theories are developed for handling such kind of data. However, in many cases, data have the characteristic of uncertainty. There are primarily two types of uncertainty. The first is probabilistic uncertainty, which is well developed overtime. The second is what is termed as fuzzy uncertainty. Let us start with fuzzy data, which is a combination of fuzzy variable and random variable and can characterize both fuzziness and randomness. Fuzzy Logic as a superset of conventional (Boolean) logic was first introduced by Zadeh (1965) to handle the concept of partial truth. Fuzzy Logic is considered as the most powerful tool for dealing with imprecision and uncertainties. Zadeh proposed that fuzzy set can be applied

to represent data which is fuzzy and this fuzziness can be represented by the degree of participation to a set called a membership function. Let $\boldsymbol{X}$ be a space of points. A fuzzy set $\mathbf{A}$ in space $\boldsymbol{X}$ is characterized by a membership function, $\boldsymbol{\mu_A(x)}$, and the value of $\boldsymbol{\mu_A(x)}$ at $\boldsymbol{x}$ representing the grade of membership of $\boldsymbol{x}$ in $\boldsymbol{A}$ where $\boldsymbol{\mu_A: X \to [0,1]}$. For traditional bivalent logic, the value of membership function of crisp data can only be 0 or 1, that means, outside the set, or within the set, respectively. However, a fuzzy set allows for its members to have degrees between 0 and 1. Thus, it can explain natural phenomenon more accurately. Further, conventional set theory and binary logic have three elementary binary operations, that is, intersection (and), union (or), and complement set (negation). The rules of binary operations were generalized in order to fitting fuzzy data. The fuzzy logic operations truth table is shown in Table 1. The generalized form of the operators works well for the fuzzy and for the bivalent data as well.

Table 1: The Generalized form of operations; Truth table

$x \; and \; y$	$\min(x, y)$
$x \; or \; y$	$\max(x, y)$
$not \; x$	$1 - x$

2. Methodology

The classical linear regression has crisp coefficients and is bounded by some strict assumptions about the given data, that is, the unobserved error terms are mutually independent and identically distributed. However, if the data set is too small in size, or, if there is difficulty in verifying that the errors are normally distributed, or, if there is vagueness in the relationship between the dependent and independent variables, or, if there is ambiguity associated with the events, it is well known that the classical linear regression may fail to work satisfactorily. In such cases, alternatively, fuzzy linear regression may be more useful. Fuzzy linear regression (FLR) was first introduced by Tanaka (1982) and then further developed in Tanaka (1987). The FLR model includes a fuzzy output and non-fuzzy input variables and fuzzy coefficients. In this paper, however, our focus is on the type of fuzzy regression model considered by Tanaka (1987). The basic model assumes a fuzzy linear functional form

$$\tilde{y} = \tilde{A}_0 + \tilde{A}_1 x_1 + \cdots + \tilde{A}_p x_p, \tag{1}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_p]$ is a vector of input variables, $\tilde{\mathbf{A}} = [\tilde{A}_0, \tilde{A}_1, \ldots, \tilde{A}_p]$ is a vector of fuzzy coefficients presented in the form of symmetric triangular fuzzy data denoted by $\tilde{A}_j = (a_j, c_j)$ with its membership function described as

$$\mu_{A_j}(\alpha) = \begin{cases} 1 - \frac{|a_j - \alpha|}{c_j}, & a_j - c_j \le \alpha \le a_j + c_j; j = 1,2,\dots,p, \\ 0, & otherwise, \end{cases} \quad (2)$$

where a_j is the central value and c_j is the width. The membership function of the fuzzy output can be described as

$$\mu_{Y_i}(y) = \begin{cases} 1 - \frac{|y_i - y|}{e_i}, & y_i - e_i \le y \le y_i + e_i; i = 1,2,\dots,n, \\ 0, & otherwise. \end{cases} \quad (3)$$

The degree of fitting of the fuzzy regression model to the given data $\boldsymbol{Y}_i = (y_i, e_i)$ is measured by an index $\min_j[\bar{h}_j]$, where

$$\bar{h}_i = 1 - \frac{|y_i - \mathbf{x}_i^T \alpha|}{\sum_j c_j |x_{ij}| - e_i}. \quad (4)$$

The vagueness of the fuzzy regression model is defined by $J = \sum_j c_j$. The fuzzy coefficient parameter $\tilde{A}_j$ is obtained so as to minimize J subject to $\bar{h}_i \ge H$, where H is chosen as the degree of fitting the model by the experimenter.

The basic idea is to minimize the fuzziness of the model by minimizing the total support of the fuzzy coefficients subject to including all the given data. As a result, we can obtain the best fitted model for the given data by solving the conventional linear programming problem.

$$\begin{gathered} \min \quad J = mc_0 + \sum_{j=1}^{m} \sum_{i=1}^{n} c_i x_{ij}, \text{ such that} \\ y_j \ge \sum_{i=1}^{n} a_i x_{ij} - (1-H) \sum_{i=1}^{n} c_i x_{ij} + (1-H) e_j, \\ y_j \le \sum_{i=1}^{n} a_i x_{ij} + (1-H) \sum_{i=1}^{n} c_i x_{ij} - (1-H) e_j, \\ c_i \ge 0, i = 0,1,\dots,n. \end{gathered} \quad (5)$$

We have prepared Matlab 2018 programming codes for fitting the model which are not included for saving the space, however, can be requested.

3. Result

For illustration of the fuzzy regression model, we have adapted the data of global sea ice extent and ocean heat content from 1979 to 2015 [Source: National Snow and Ice Data Center (NSIDC)]. Global climate data indicate that in 2018, a new record was set for the total amount of warmth stored in the seas known as the ocean heat content (OHC). Measured OHC was warmer than any other year since observations began in the early 1940s. Sea ice was at record or near-record lows in the Arctic, noted to be only the 6th lowest since records began in the late 1970s. There is also currently a near-record low level of multi-year sea ice in the Arctic, with around 80% of sea ice only one to two

years old. The main purpose of the application in this study is to model and predict the effect of ocean heat content, x ($10^{\wedge 22}$ Joules) on the global sea ice extent, y (in million km^2).

3.1. Method of Ordinary Least Squares

Fitted ordinary least squares (OLS) model to the global sea ice extent, y and ocean heat content, x is $y = A_0 + A_1 x$, where the fitted coefficients A_0 and A_1 are shown in Table 2.

Table 2: Crisp Coefficients of Global Sea Ice Extent and Ocean Heat Content

Coefficients	A_0	A_1
Estimate	23.5114	-0.0614
Standard error	0.8360	0.0370

3.2. Method of Fuzzy Regression

This section aims to determine the best fuzzy regression model to explain the relationship between ocean heat content, x and global sea ice extent, y and compare with the classical regression. We have fitted the fuzzy regression model:

$$y = \widetilde{A_0} + \widetilde{A_1} x, \tag{6}$$

where $\widetilde{A_0}$ and $\widetilde{A_1}$ are fuzzy coefficients describing the centre and the width, respectively. To obtain fuzzy coefficients, we have assumed 1% spread of the global sea ice extent which could be due to the measurement errors or due to other unknown sources. The upper bounds and the lower bounds of the global sea ice extent are obtained. The observed values are expected to be in the interval of the computed bounds. For the algorithm, computer codes are prepared and the statistical software Matlab 2018 is used to analyze the data. We obtain the fitted model by solving the linear programming problem and the results are in Table 3.

Table 3: Fuzzy Coefficients of Global Sea Ice Extent and Ocean Heat Content

Fuzzy coefficients	$\widetilde{A_0}$	$\widetilde{A_1}$
Center a_j	23.5713	-0.0729
Width c_j	0.4950	0.0325

In order to compare ordinary least squares methodology and fuzzy linear regression methodology, width of predicted intervals in terms of each observed value of explanatory variable is computed. Fuzzy output lower bound y_l and upper bound y_c are shown in Figure 1. As for the ordinary least

squares model, upper limits or lower limits of prediction interval are computed by taking the coefficients as their corresponding estimated value plus or minus standard error, i.e. using the equation,

$$y = (23.5114 \pm 0.8360) + (-0.0614 \pm 0.0370)x. \quad (7)$$

Similarly, the limits of fuzzy regression model are computed by the equation

$$y = (23.5713 \pm 0.4950) + (-0.0729 \pm 0.0325)x. \quad (8)$$

The width of predicted intervals of fuzzy regression model is much smaller than that of ordinary least squares model, which indicates the superiority of fuzzy regression methodology.

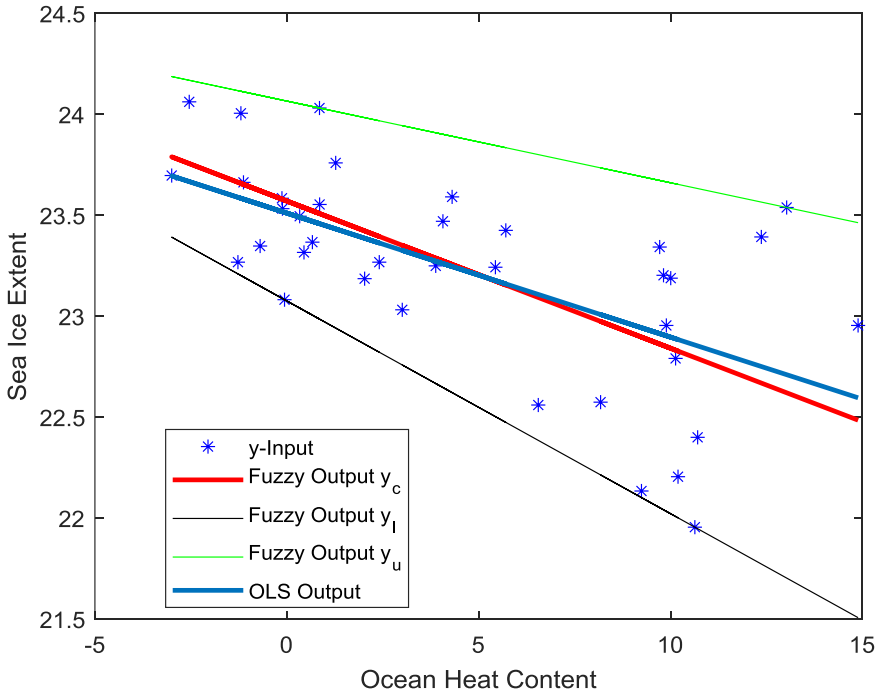


Figure 1: Ocean Heat content (in million km^2) and Global Sea Ice Extent (in million km^2).

3.3. Error Analysis

To make performance comparisons, we have calculated prediction errors denoted by $e_f = y_{observed} - y_{center}$ and $e_{lse} = y_{observed} - y_{lse}$ and corresponding standardized errors $e_{fstandardized} = \frac{e_f}{\sigma_{e_f}}$, and $e_{lstandardized} = \frac{e_{lse}}{\sigma_{else}}$, where σ_{e_f} and σ_{else} denote the standard errors. Further, root mean

squared errors are calculated as $RMSE = \sqrt{\frac{\sum e^2}{n-1}}$ and given in Table 4. Calculated RMSE for FLR and OLS fitted models are 0.366195 and 0.379641, respectively.

Table 4: Prediction Error for the Fuzzy and Ordinary Least Squares Regressions.

		FLR		OLS	
Obs.	Year	e_f	$e_{standardized}$	e_o	$e_{standardized}$
1	1979	0.3473	0.8717	0.4209	1.067654
2	1980	0.0453	0.1137	0.0954	0.241991
3	1981	-0.0463	-0.11621	0.0149	0.037795
4	1982	0.3058	0.767538	0.3949	1.001702
5	1983	-0.0925	-0.23217	0.0019	0.00482
6	1984	-0.2737	-0.68697	-0.2058	-0.52203
...	...	...	...	...	...
35	2013	0.7237	1.816439	0.6417	1.627734
36	2014	0.9177	2.303366	0.8281	2.100556
37	2015	0.4692	1.177661	0.3583	0.908863
R^2		0.366195		0.379641	

We have plotted the prediction standardized errors due to fitted fuzzy and least squares regressions in Figure 2. In the Figure, the blue line denotes the standard errors of FLR and the red line is the standard errors of OLS.

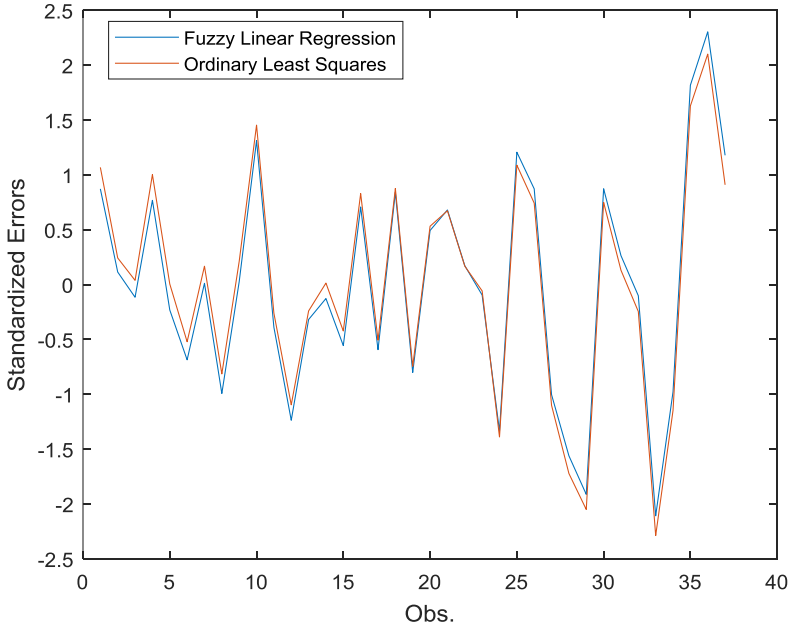


Figure 2: Standardized Prediction Errors

4. Discussion and Conclusion

In this paper, we have compared the fuzzy regression model with ordinary least squares method. To illustrate the fuzzy regression model, we have adapted the data of global sea ice extent and ocean heat content from 1979 to 2015. We have calculated the upper and the lower bounds of global sea ice extent and carried error analysis which clearly indicates the comparative performance of two fitting methods with possible edge of the fuzzy regression over the ordinary least squares regression. Besides, the width of predicted intervals of fuzzy regression model is much smaller than that of ordinary least squares model, which indicates the superiority of fuzzy regression methodology.

References

1. Tanaka, H., Uejima, S. and Asai, K., Linear regression analysis with fuzzy model, *Systems Man & Cybernetics, IEEE Transactions*, 12(6):903-907, 1982.
2. Tanka, H, Fuzzy data analysis by possibilistic linear models, *Fuzzy Sets & Systems*, 24 (3):363-375, 1987.
3. Zadeh, L. A., Fuzzy sets, Information and Control, 8 (3): 338–353, 1965.

The impact of weather risk on the estimation of yield-based agricultural losses and value at risk using Copula Models

Atina Ahdika[1,2], Dedi Rosadi[1], Adhitya Ronnie Effendhie[1], Gunardi[1]
[1]Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia
[2]Department of Statistics, Universitas Islam Indonesia, Yogyakarta, Indonesia

Abstract
Weather risk, such as temperature change, is one of the main factors affecting agricultural products. Temperature change can significantly affect the occurrence of agricultural losses which can be measured from the agricultural yield. An agricultural loss is defined as the difference value between the estimated and the actual yield at some confidence levels. This paper aims to identify the dependency structure between temperature change and agricultural yield using copula functions. The estimation procedure of yield-based agricultural losses is conducted by simulating joint occurrence between the two variables and the selected copula parameters. The result shows that the agricultural losses happened mostly when the temperature is low. Value at risk in the form of yield-based agricultural losses is also estimated based on the distribution of the estimated losses.

Keywords
agricultural losses; agricultural yield; copula; temperature change; value at risk

1. Introduction

Indonesia is one of the developing countries whose main livelihood is farming. Farmers are very susceptible to losses such as crop failure or a decrease in the price index of agricultural production which can be caused by weather risk or disease attacks. Many studies have been conducted to estimate agricultural losses based on the factors that influence them. Vergara *et al.* (2008) modelled the impact of catastrophic weather on crop insurance losses. Dahal & Routray (2011) identified the association between soil variables and agricultural yield using multiple linear regression. Sellam & Poovammal (2016) predicted agricultural yield by analysing the relationship between environmental parameters such as harvest area, annual rainfall, and food price index using linear regression. Luminto & Harlili (2017) built a weather analysis to predict rice cultivation time to increase farmers exchange rate using linear regression.

Along with the advance research in the field of correlation, the relationship between the variables that affect the risk of agricultural losses is assumed to not always be linear so that the prediction model based on the Pearson correlation coefficient, such as a linear regression model, is no longer relevant

to use. One of the association measurements which can measure dependency structure both linear and non-linear is rank correlation consists of Kendall's τ andSpearman's ρ. The two measurements can be expressed as a multivariate distribution function called copula. Copula function has been widely used to identify dependency structures between two variables, including in agriculture sector. Zhu et al. (2008) modelled the dependency structure between corn and soybean yield and price using copula. Xu et al. (2010) measured the spatial dependency of the weather risk in some areas and modelled its impact to the indemnity payment of crop insurance using copula. Xu et al. (2010) modelled the joint loss distribution based on the systemic weather risk using hierarchical copula.

In this study, we aim to identify the dependency structure between weather risk, in this case is temperature change, and the agricultural yield using copula models. The best copula model is selected to build a simulation to estimate yield-based agricultural losses. Furthermore, we estimate the value at risk of the agricultural losses by using the distribution of the estimated yield-based agricultural losses obtained from the copula modelling.

2. Model Framework

2.1 Copula Modelling

Following Xu et al. (2010), the dependency structure between weather risk and agricultural yield can be modelled using copula function. Suppose that $F_{X,Y}(x,y)$ is a joint distribution function with marginal distribution functions $F_X(x)$ and $F_Y(y)$. There is a copula C such that (Nelsen, 2006)

$$F_{X,Y}(x,y) = C(F_X(x), F_Y(y)) \tag{1}$$

If $F_X(x)$ and $F_Y(y)$ continue, then C is unique. Otherwise, if C is copula, $F_X(x)$ and $F_Y(y)$ are distribution functions, then $F_{X,Y}(x,y)$ is a joint distribution function with marginal distribution function $F_X(x)$ and $F_Y(y)$. Eq. (1) can also be written as

$$F_{X,Y}(x,y) = C(u,v) \tag{2}$$

where $u = F_X(x)$ and $v = F_Y(y)$ which are uniformly distributed at $[0,1]$.

Eq. (1) gives information about marginal distribution and copula function. In copula concept, X and Y can be modelled with any distribution function. Therefore, before selecting the best copula function, distribution fitting for both marginal variables has to be done first.

In this paper, we use Gaussian copula, Student t -copula, Archimedean copula (Clayton, Gumbel, and Frank), and Farlie-Gumbel-Morgenstern (FGM) copula. Each of the copula is defined as follow

a. Gaussian Copula

Gaussian copula is derived from bivariate normal distribution, is defined by

$$C_{Gauss}(u,v)=\int_{-\infty}^{\Phi^{-1}(u)}\int_{-\infty}^{\Phi^{-1}(v)}\frac{1}{2\pi\sqrt{(1-\rho^2)}}\left\{-\frac{y^2-2\rho yz+z^2}{2(1-\rho^2)}\right\}dydz \qquad (3)$$

where ρ is the correlation function and $\rho\in(-1,1)$. Gaussian copula does not have a tail dependence, therefore $\lambda^L=\lambda^U=0$.

b. Student t -Copula

Student t -copula is related with the bivariate student t distribution. The copula is defined by

$$C_{Student-t}(u,v)=\int_{-\infty}^{t_\upsilon^{-1}(u)}\int_{-\infty}^{t_\upsilon^{-1}(v)}\frac{1}{2\pi\sqrt{(1-\rho^2)}}\left\{1+\frac{y^2-2\rho yz+z^2}{\upsilon(1-\rho^2)}\right\}^{\frac{\upsilon+2}{2}}dydz \qquad (4)$$

Similar with Gaussian copula, the range of the parameter ρ is $(-1,1)$. While for the tail dependence, student t -copula has symmetric lower and upper tail dependence, i.e. $\lambda^L=\lambda^U=2T_{\upsilon+1}\left(-\sqrt{\frac{(\upsilon+1)(1+\rho)}{1+\rho}}\right)$, where $T_{\upsilon+1}$ is the cumulative distribution function of student t distribution with degree of freedom $\upsilon+1$.

c. Archimedean Copula

Archimedean copula used in this paper consists of Clayton, Gumbel, and Frank copula defined by

$$C_{Clayton}(u,v)=(u^{-\theta}+v^{-\theta}-1)^{-\frac{1}{\theta}},\ \theta\in(0,+\infty) \qquad (5)$$

$$C_{Gumbel}(u,v)=\exp\left\{-\left[(-\ln u)^\theta+(-\ln v)^\theta\right]^{\frac{1}{\theta}}\right\},\ \theta\in[1,+\infty) \qquad (6)$$

$$C_{Frank}(u,v)=-\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right),\ \theta\in(-\infty,0)\cup(0,+\infty) \qquad (7)$$

where Clayton has lower tail dependence $\lambda^L = 2^{-1/\theta}$, Gumbel has upper tail dependence $\lambda^U = 2 - 2^{-1/\theta}$, and Frank does not have tail dependence $\lambda^L = \lambda^U = 0$.

d. Farlie-Gumbel-Morgenstern (FGM) Copula
The bivariate FGM copula is defined by

$$C_{FGM}(u,v) = uv + \theta u(1-u)v(1-v), \quad \theta \in [-1,1] \tag{8}$$

Parameter estimation of copula function is obtained by maximizing the value of log-likelihood function with respect to the parameter as follows

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \ln[c(F_X(x_i), F_Y(y_i))] \tag{9}$$

The best copula function is obtained by selecting the smallest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) value given by

$$AIC = 2k - 2\ln(\hat{L}) \tag{10}$$

$$BIC = \ln(n)k - 2\ln(\hat{L}) \tag{11}$$

where k be the number of estimated parameters, n be the number of observations, and $\hat{L}$ be the maximum value of the likelihood function.

2.2 Yield-Based Agricultural Losses and Value at Risk Estimation

After fitting the distribution for both marginal variables and selecting the best copula function, the estimation of the yield-based agricultural losses and value at risk in terms of losses can be done. The procedure is as follows. Suppose that $\hat{\Lambda} = \{\hat{\lambda}, \hat{k}, \hat{\theta}\}$ be a set of estimated parameters where $\hat{\lambda}$ and $\hat{k}$ be the estimated parameters of the marginal distributions and $\hat{\theta}$ be the estimated parameter of the best copula function. Suppose that X represents the weather variable and Y represents the yield variable. First, simulate a joint vector $(\hat{u}, \hat{v})$ from the best copula function with its estimated parameter $\hat{\theta}$. Let $\hat{u} = F_X(x)$ and $\hat{v} = F_Y(y)$, then generate the value of $(\hat{x}, \hat{y})$ by inverse the cumulative distribution function to be $\hat{x} = F_X^{-1}(\hat{u})$ and $\hat{y} = F_Y^{-1}(\hat{v})$ using the estimated parameters $\hat{\lambda}$ and $\hat{k}$ of each marginal distribution. Then, the yield-based agricultural losses is defined as the difference value between the estimated and the actual yield at some confidence levels p. The formula is given by

$$\hat{L} = \max\{0, \hat{y} \cdot p - y\} \tag{12}$$

where p is the confidence level, $\hat{y}$ is the estimated yield, and y is the actual yield.

Losses are said to occur if the actual yield is smaller than the estimated yield at a certain confidence level. Otherwise, if the actual yield is greater than the estimated yield at a certain confidence level, then farmers experience profits.

Furthermore, value at risk in terms of losses is estimated as a reference for farmers, on how much yield the farmers are prepared to experience losses at certain probability level. The value at risk is calculated based on the marginal distribution of the $\hat{Y}$ obtained from the simulation of the copula models, where the basic formula is

$$P(Y > \pi_p) = 1 - p \tag{13}$$

where π_p is the value at risk at the $100p\%$ level.

3. Empirical Result

The procedure explained in the previous section is applied to the rice production (hg/ha) and temperature change (^{o}C) data of Indonesia from 1961 to 2016. The data obtained from the official website of BPS-Statistics Indonesia. The rice production and temperature change data are presented in Fig. 1.

Based on Fig. 1, rice production data has an increase trend with a little fluctuation. While temperature changes data is quite fluctuating. To determine whether the two variables have relationship or not, the value of Kendall's τ and Spearman's ρ are calculated. The results show that the value of each association measure is 0.670997 and 0.852851, respectively. It means that the two variables have quite strong dependency.

The next step is to fit the distribution of each variable. Descriptive statistics of rice production and temperature change along with the normality test using Shapiro-Wilk test is presented in Table 1.

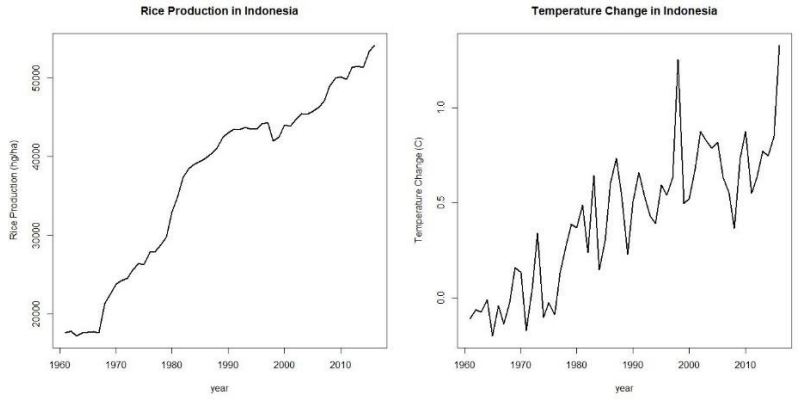


Figure 1. Rice Production and Temperature Change in Indonesia from 1961 to 2016

Table 1. Descriptive Statistics and Shapiro-Wilk Statistics

Variable	Mean	*Sd*	Median	Min	Max	Skew	Kurtosis	W	$p-value$
Rice Prod.	37057.54	11374.89	41541	17226	54148	-0.45	-1.19	0.90343	0.000289
Temperature Change	0.42	0.36	0.49	-0.2	1.33	0.12	-0.56	0.96164	0.07225

Table 1 shows the descriptive statistics of the marginal distributions. From its skewness, initial identification of the marginal distributions can be done. Skewness of rice production data shows that the data is negatively skewed where the left tail is longer. Otherwise, temperature change data is positively skewed where the right tail is longer. To identify whether the two variables are normally distributed or not, the normality test using Shapiro-Wilk test is held. The null hypotheses expressed that the population is normally distributed. If $p-value$ is less than α, then H_0 is rejected which mean that the population is not normally distributed. Based on the results show in Table 1, rice production data is not normally distributed while temperature change data is normally distributed. Table 2 and Fig. 2 present the result of the distribution fitting for both variables.

Table 2. Distribution Fitting

Variable	Distribution	Log-Likelihood	AIC	BIC
Rice Production	Normal	-601.949	1207.898	1211.949
	Log-Normal	-607.261	1218.522	1222.573
	Weibull*	-600.389	1204.777	1208.828
Temperature Change	Normal*	-21.871	47.7419	51.79261

**sign indicate the fitted distribution*

Table 2 and Fig. 2 show that rice production data follows Weibull distribution and temperature change data follows Normal distribution based on the smallest AIC and BIC value. The estimated parameters of rice production variable are $\hat{\lambda} = 41096.88$ and $\hat{k} = 3.956537$. While for temperature change data are $\hat{\mu} = 0.415589$ and $\hat{\sigma} = 0.357584$.

After getting the fitted marginal distribution, the next step is to estimate and select the best copula model which can describe the dependency structure between rice production and temperature change data. The estimated parameters of copula models are presented in Table 3.

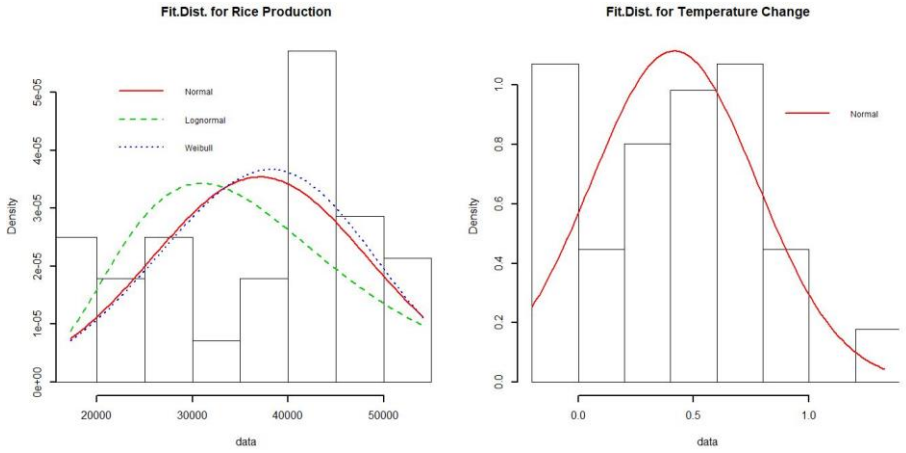


Figure 2. Distribution Fitting for Both Variables

Table 3. Estimated Parameters of Copula Models

Copula	$\hat{\theta}$	df	$\hat{\tau}$	$\hat{\tau}^L$	$\hat{\tau}^U$	AIC	BIC
Gaussian*	0.846589	0	0.642696	0	0	-70.9426	-68.9172
Student-t	0.849506	17.61235	0.6462	0.233783	0.233783	-69.4323	-65.3816
Clayton	2.276665	0	0.532346	0.7375235	0	-65.3608	-63.3355
Gumbel	2.51256	0	0.602	0	0.68232	-60.4348	-58.4095
Frank	9.406259	0	0.649079	0	0	-70.415	-68.3897
FGM	1	0	0.22222	0	0	-10.8329	-8.80753

**selected copula*

Table 3 shows the estimated parameter of each copula along with the AIC and BIC value. From the table, we can see that Gaussian copula is the best copula function to model rice production and temperature change data because it has the smallest AIC and BIC values. By applying the procedure in the previous section, the estimation of yield-based agricultural losses can be done. The result is presented in Fig. 3.

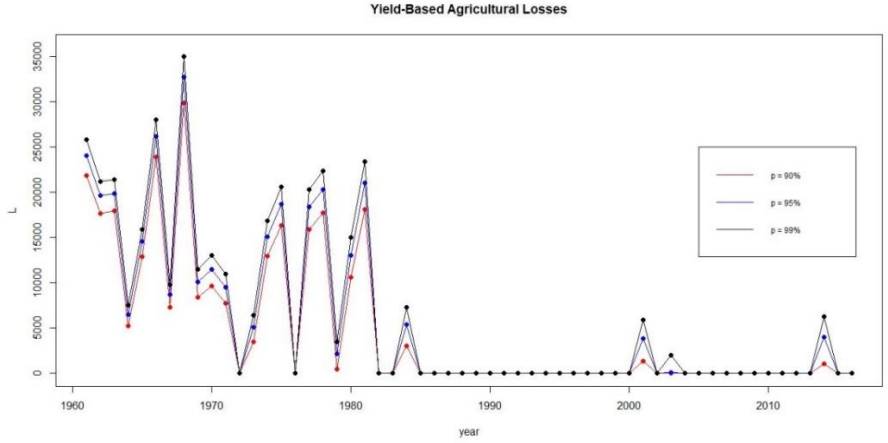


Figure 3. Yield-Based Agricultural Losses Estimation

Fig. 3 shows the estimation of yield-based agricultural losses. Based on the result, the losses have occurred mostly in the early years. It might happen because in the early years there was a lot of temperature decline which was indicated by the temperature change graph that often declined (see Fig. 1). The losses also occurred around the beginning of year 2000 and close to 2016. The reason is similar with the previous indication. Therefore, it can be concluded that the losses happen mostly when the temperature is low. Furthermore, the estimation of value at risk is calculated based on the marginal distribution of the rice production. Because it follows Weibull distribution, then the value at risk is calculated by

$$VaR_p(Y) = \pi_p = \hat{\lambda}\left(-\log(1-p)\right)^{\frac{1}{\hat{k}}} \tag{14}$$

The estimation of value at risk at confidence levels of 90%, 95%, and 99% are 50740.83, 54230.45, and 60456.33 (in hg/ha), respectively. It means that the farmers prepare to suffer losses at the level of $100p > 90\%$ at the rice production value greater than 50000 hg/ha.

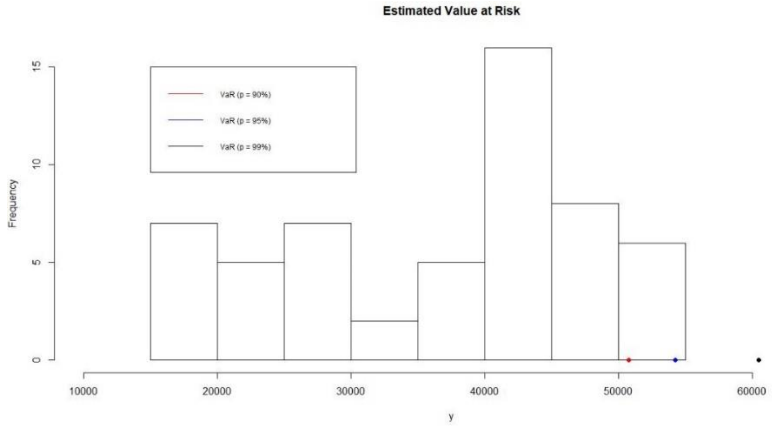


Figure 4. Value at Risk Estimation

4. Conclusion

The estimation of yield-based agricultural losses and value at risk in the form of losses has been conducted using copula models. Overall, the result shows that the losses might occurred mostly when the temperature decline. It indicates that rice production has quite strong dependency with the weather risk, in this case is temperature change. Furthermore, by using value at risk, the farmers can estimate when to prepare for losses.

References

1. Dahal, H., & Routray, J. K. (2011). Identifying Associations Between Soil and Production Variables Using Linear Multiple Regression Models. *The Journal of Agriculture and Environment*, *12*, 27–37.
2. Luminto, & Harlili. (2017). Weather analysis to predict rice cultivation time using multiple linear regression to escalate farmer's exchange rate. *Proceedings - 2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017*, 0–3.
3. Nelsen, R. B. (2006). *An Introduction to Copulas*.
4. Sellam, V., & Poovammal, E. (2016). Prediction of Crop Yield Using Regression Analysis. *Indian Journal of Science and Technology*, *9*(38), 1–5.
5. Vergara, O., Zuba, G., Doggett, T., & Seaquist, J. (2008). Modeling the Potential Impact of Catastrophic Weather on Crop Insurance Industry Portfolio Losses. *American Journal of Agricultural Economics*, *90*(5), 1256–1262.
6. Xu, W., Filler, G., Odening, M., & Okhrin, O. (2010). On The Systemic Nature of Weather Risk. *Agricultural Finance Review*, *70*(2), 267–284.
7. Xu, W., Odening, M., Ji, C., & Okhrin, O. (2010). *Systemic Weather Risk and Crop Insurance: The Case of China*. *SFB 649 Discussion Paper 2010-053*.
8. Zhu, Y., Ghosh, S. K., & Goodwin, B. K. (2008). Modeling Dependence in the Design of Whole Farm - A Copula-Based Model Approach. *Selected Paper Prepared for Presentation at the American Agricultural Economics Association Annual Meeting, July 27 - 29*, 27–29.

Structured additive Regression Modeling of pulmonary tuberculosis infection

Bruno de Sousa[1], Carlos Pires[1], Dulce Gomes[2], Patrícia Filipe[2], Ana Costa-Veiga[3,4], Carla Nunes[3]

[1]Faculty of Psychology and Education Sciences, University of Coimbra, Portugal
[2]Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora, Departamento de Matemática, Escola de Ciências e Tecnologia, Portugal
[3]CISP - Centro de Investigação em Saúde Pública, National School of Public Health, Universidade Nova de Lisboa, Portugal
[4]H&TRC - Health & Technology Research Center, ESTeSL, Lisbon School of Health Technology, Instituto Politécnico de Lisboa, Av. D. João II, Portugal

Abstract

Tuberculosis (TB) is one of the top 10 causes of death and the leading cause from a single infectious agent (above HIV/AIDS). In 2017, the World Health Organization (WHO) estimated 10.0 million people developed TB and 1.3 million deaths (range, 1.2–1.4 million) among HIV-negative people with an additional 300 000 deaths from TB (range, 266 000–335 000) among HIV-positive people. Studies that understand the socio-demographic characteristics, time and spatial distribution of the disease are vital to allocating resources in order to improve National TB Programs. The database includes information from all confirmed Pulmonary TB (PTB) cases notified in Continental Portugal between 2000 and 2010. Following a descriptive analysis of the main risk factors of the disease, a Structured Additive Regression (STAR) model is presented exploring possible spatial and temporal correlations in PTB incidence rates in order to identify the regions of increased incidence rates. Three main regions are identified as statistically significant areas of increased PTB incidence rates in Continental Portugal. STAR models proved to be a valuable and effective approach in identifying PTB incidence rates and will be used in future research to identify the associated risk factors in Continental Portugal, yielding high-level information for decision-making in TB control.

Keywords

Structured Additive Regression Models; Pulmonary Tuberculosis; Spatial-Temporal Epidemiology; Full Bayesian; Empirical Bayesian

1. Introduction

Pulmonary Tuberculosis (PTB) is an infectious disease which affects millions of people every year, being the second most deadly infectious disease worldwide after the human immunodeficiency virus (HIV) [1]. The disease is caused by the bacillus Mycobacterium tuberculosis that affects mainly the

lungs, and can be transmitted through the air when the bacteria is expelled by coughing, sneezing or speaking.

From all notified cases in the WHO European Region in 2017, about 80% had pulmonary localization (PTB) [1,2], a fact also verified in Portugal, with 73.5% of the cases in our database being PTB. An earlier study conducted in Portugal in 2011 aimed at identifying critical areas for the joint occurrence of PTB and HIV/AIDS (Acquired Immune Deficiency Syndrome). The study, based on spatiotemporal clustering analyses, identified the Oporto and Lisbon Metropolitan Areas as critical areas for both diseases, either independently or jointly occurring [3].

Research on spatial and temporal correlations among PTB incidence rates together with disease factors are of the utmost importance from a Public Health perspective. This study will focus on analyzing through STAR (Structured Additive Regression) modeling temporal trends and geographic patterns of PTB incidence rates associated with notified PTB cases in Continental Portugal (278 municipalities) from 2000 to 2010.

2. Methodology

2.1 The data

This study was entirely based on data from registers with the permission of the National Program for Tuberculosis Control. The data was extracted from SVIG-TB (*Sistema de Vigilância da TB em Portugal*) database of the National Program for Tuberculosis Control and included information from all confirmed TB cases, whose notification is mandatory in Continental Portugal (henceforth referred to as Portugal) between 2000 and 2010. Ethics committee approval and informed consent were not required, as data was based on an Official National Surveillance System, provided by the General Directorate of Health, and was previously anonymized.

A total of 25,279 new cases with PTB were used, together with the information regarding municipality of residence, age, sex and disease risk factors, such as alcohol dependence, intravenous drug dependence (IV Drugs), other drug dependence, being an inmate, homeless, an immigrant and co-infected with HIV. This study considered a new case as one defined by WHO [1], that is, a patient with PTB disease involving lung parenchyma who has never received a treatment or who has been taking anti-TB drugs for less than one month. Yearly population data (global and per municipality) were taken from Statistics Portugal.

2.2 The model

Structured Additive Regression Models (STAR) enable the placement within the same framework of nonlinear effects of continuous covariates, spatial effects, time trends and the usual linear or fixed effects in regression

models with non-Gaussian responses [4]. A suitable STAR model for spatiotemporal data is given by

$$\eta_{it} = f_1(x_{it1}) + \ldots + fk\,(x_{itk}) + f_{trend}(t) + f_{spat}\,(s_{it}) + u'_{it}\,\gamma, \quad (1)$$

where η_{it} is the additive predictor for observation i at time t, $f_1(x_{it1})$, ..., $f_k(x_{itk})$ are smooth functions of k continuous covariates x_{it1},... x_{itk}, $f_{trend}(t)$ is a temporal trend, $u'_{it}\,\gamma$ represents the parametric component with γ being the parameter vector of the fixed effects, and $f_{spat}\,(s_{it})$ is a spatially correlated effect of the location (s) where the observation belongs. The spatial effect can furthermore be split into a spatially correlated part and a spatially uncorrelated part: $f_{spat}\,(.) = f_{srt}\,(.) + f_{unstr}\,(.)$, allowing for a distinction to be made between the unobserved influential factors which obey a global spatial structure and those which may be present only locally [5]

For smooth non-linear effects of continuous covariates and time trends Bayesian penalized splines are used [6, 7]. Correlated and uncorrelated spatial effects follow a Gaussian Markov random field and an independent identically distributed (iid) Gaussian random effects priors, respectively [8].

Inference in the above STAR model can be made through a full (FB) or empirical Bayesian (EB) approach. In a FB approach the unknown variance or smoothing parameters are considered as random variables with suitable hyperpriors and are estimated together with the unknown functions and covariate effects, using MCMC (Markov chain Monte Carlo) simulation techniques [9]. EB approach is based on penalized likelihood inference for the regression coefficients and restricted maximum likelihood estimation (REML) for the variance components [4, 5, 9].

The model here presented analyzes the temporal trend and the spatial distribution of PTB incidence rates in Portugal between 2000 and 2010. The main goal was to identify areas with different risk levels in terms of PTB incidence rates, if they exist.

For this model, municipality was considered as the statistical unit and Y_{it}, the number of new PTB cases in the i^{th} municipality at year t, as the response variable. To be able to model PTB incidence rates, an offset term with regression coefficient fixed to 1 is included in the model and is defined as $log(P_{it}/100{,}000)$, where P_{it} represents the number of habitants in the municipality i at the year t. The final model can then be specified as:

$$\eta_{it} = log(P_{it}/100\,000)\,(offset) + f_{year}\,(t) + f_{srt}\,(s_{it}) + f_{unstr}\,(s_{it}), \quad (2)$$

where $\eta_{it} = log(E(Y_{it}))$ represents the additive predictor for the $i = 1, \ldots, 278^{th}$ municipally at year $t = 2000, \ldots, 2010$. The function f_{year} is a smooth function estimated using a Bayesian cubic P-spline [6, 7] with second

order random walk penalty with 20 inner knots. For the spatial components, a Gaussian Markov random field is used for the structured effects, $f_{srt}(.)$, and an iid Gaussian random effects for the unstructured effects, $f_{unstr}(.)$ [10]. To take into account the excess of zeros and possible overdispersion of the data, a zero-inflated negative binomial distribution for the response variable was assumed [11]. Inference results are obtained considering a FB approach.

3. Results and Discussion

3.1 Descriptive analysis

Portugal shows a decrease of 42.3% in PTB incidence rates from 28.6 cases per 100 000 population in 2000 to 16.5 cases in 2010 (Figure 1a). When looking at sex differences (Figure 1b), the ratio man to woman was 2.4 in the period 2000-2010, being stable over this time period. Regarding the ratio by age group, Figure 1b, there is almost the same number of new cases for men and women before the age of 25, with over 3 times more new cases of men between the ages of 35 and 64. It is also worth noting that, although there is a decrease in the sex ratio for the age group greater than 64 years of age, this ratio is still equal to 2 for this class.

With respect to changes in age over time, the consistent decrease in incidence is followed by a consistent increase of the median age, Figure 1a, suggesting a decrease in PTB endemic in Portugal.

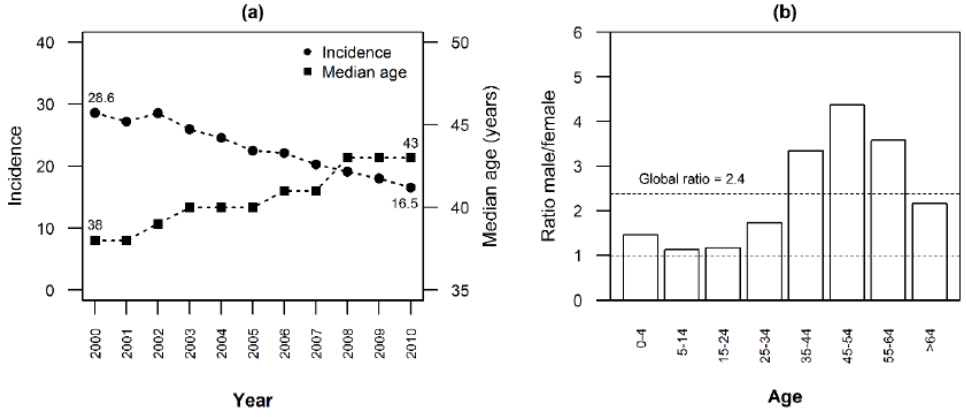


Figure 1: (a) Incidence (new cases per 100 000 population) versus median age, by year; (b) Sex ratio (men to women) of new PTB cases by age group for the period 2000-2010

Factors such as alcohol or drug dependency, HIV co-infection, being an inmate, homeless or an immigrant could contribute to the increased risk of infection with TB, as well as of disseminating it if already ill. Figure 2 shows the yearly evolution of these factors in our database.

Worth note is the steady decrease in the proportion of HIV diagnosed individuals, from 22.3% in 2000 to 10.7% in 2010. A similar trend was observed in IV drugs (Intravenous drugs) dependents that decreased from 12.8% to 7.1% in the same period. Although more moderate, the proportion of new PTB cases being alcohol dependent is also decreasing over time. Notice the increase of the proportion of immigrants after the year 2005.

When looking at the risk factors by sex (Figure 3), it is very clear that the percentage of men with a certain risk factor is always higher when compared to women, except when an immigrant. This difference is quite remarkable when looking at alcohol, where almost 25% of the men in the database are alcohol dependent.

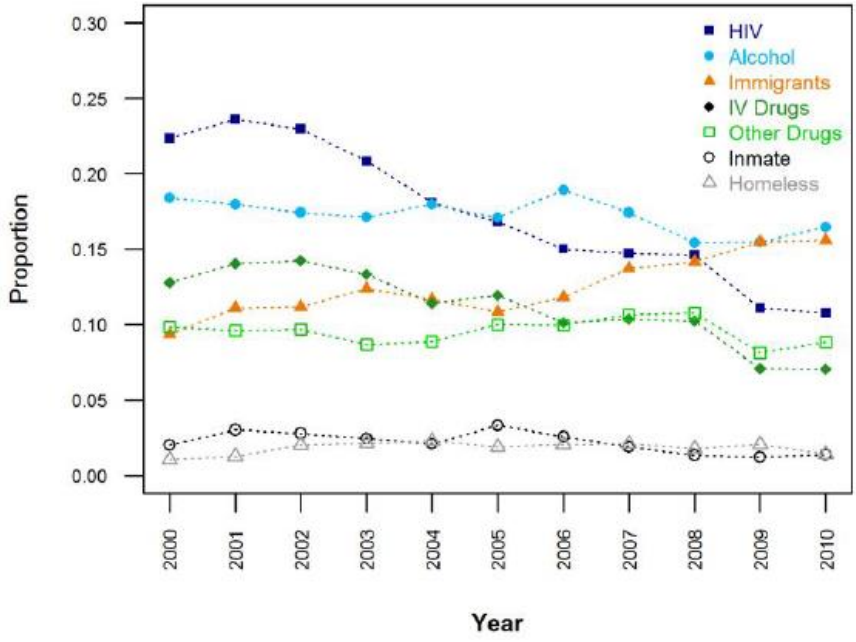


Figure 2: Proportion of risk factors per year in PTB new cases, 2000-2010

Although the total number of men and women are quite different (numbers in brackets in Figure 3), these differences are indeed statistically significant with a $p < 0.001$ for all the comparisons, with the exception of the proportion of being an immigrant, which showed statistically significant differences between sexes with a $p = 0.040$.

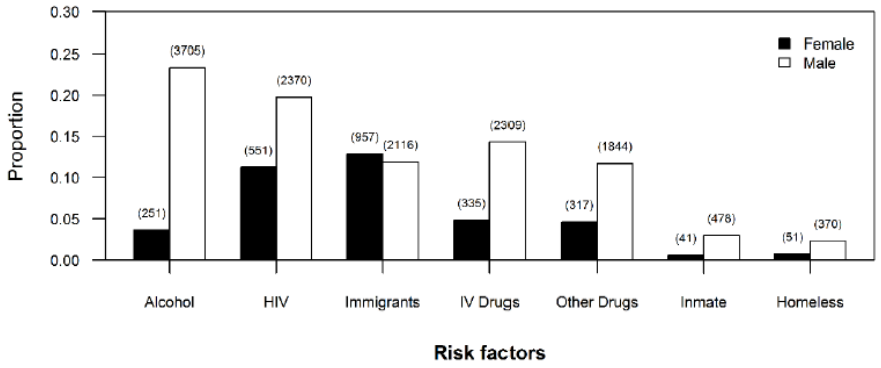


Figure 3: Proportions of risk factors per sex in PTB cases, 2000-2010. In brackets are presented the absolute number of cases.

3.2 Spatial and temporal analysis

Figure 4 shows a clear spatial pattern, with the Metropolitan Area of Porto/Upper North (Region I - MAP), Metropolitan Area of Lisbon (Region II - MAL) and Algarve/Lower Alentejo (Region III) areas (red/darker and black areas in Figure 4) being the higher risk regions that significantly contribute to an increase of the PTB incidence rates. On the contrary, it shows some regions with lower risk in the interior north, center, and Alentejo (lighter areas in Figure 4), that are significantly decreasing PTB incidence rates. The model did not show any significant unstructured (local) spatial effects (not shown here).

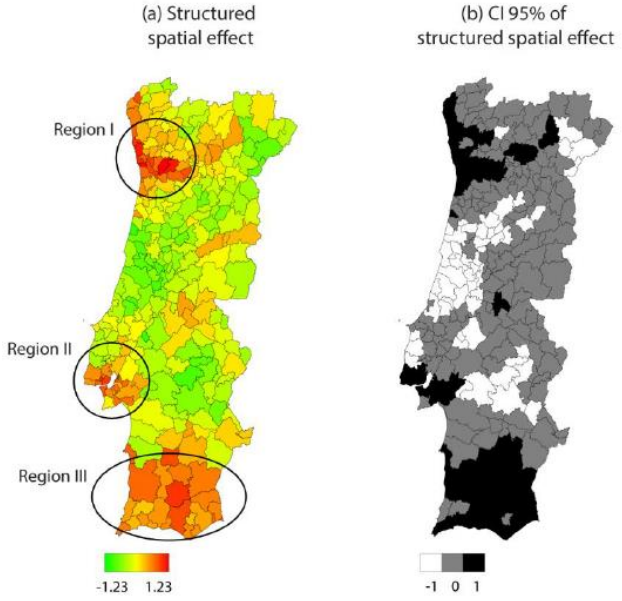


Figure 4: For the period 2000 -2010, (a) Spatial distribution of the posterior means of the global spatial effect; (b) 95% posterior probabilities. Black areas on (b) denote municipalities with strictly positive credible intervals; white areas representing municipalities with strictly negative credible intervals; and

grey areas represent municipalities of non-significant effects for PTB incidence rates (credible intervals containing zero)

Regarding time, Figure 5 shows a slightly non-linear decreasing effect between 2000 and 2010, confirming the capacity of the model to pick up the decreasing effect of PTB new cases shown in the previous descriptive analysis (Figure 1a).

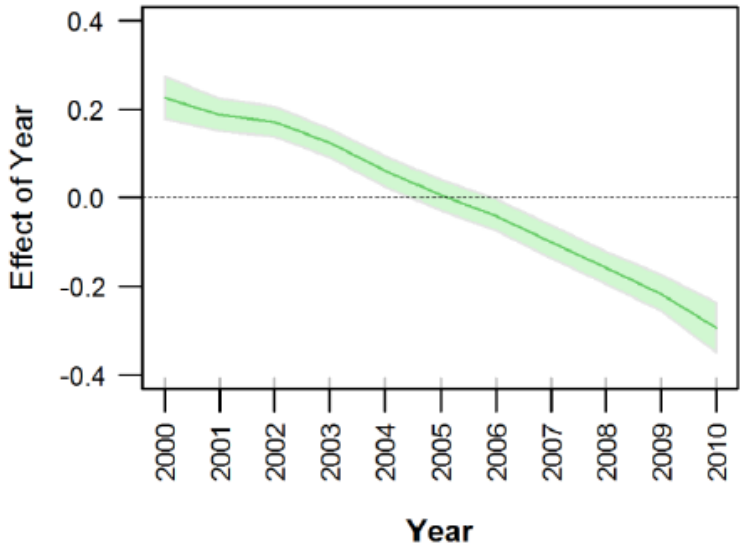


Figure 5: Estimated nonlinear effect of year in PTB incidence rates, together with 95% credible intervals.

4. Conclusion

Nunes et al. [3] identified two main regions, MAL and MAP, as being high risk areas for contracting PTB in Portugal in 2001. The results of our study also suggest a clear urban problem, with MAL (Region II) and the MAP (Region I) being two of the main areas identified as statistically significant areas of increased PTB incidence rates (Figure 4). Although with smaller numbers of new cases of PTB, Algarve and Lower Alentejo (Region III) also emerge as a region within this category. The metropolitan areas of Lisbon (Region II) and Oporto (Region I) correspond to two regions with high population density, resulting immediately in an agglomeration of the main risk groups associated with high incidence of tuberculosis (e.g. homeless, unemployed, IV drug addicts and other drugs). On the other hand, Region III which includes Algarve, not corresponding to an area of high population density throughout the year, it is associated with seasonal tourism and workers particularly through the months of April to September, when it also becomes a high density populated region. It is worth noticing that, after Lisbon with 52% of the total of foreigners living in Portugal, Algarve, North and Center of Portugal, are the three regions with the highest percentage of foreigners (13%, each). In addition, 12% of

Algarve's population is foreigner, making it the region with the greatest representativeness of foreigners' residents (Census 2011, Statistics Portugal).

Future research will focus on the risk factors associated with the identified four regions, namely Region I – Metropolitan Area of Porto and Upper North (34 municipalities), Region II –Metropolitan Area of Lisbon (20 municipalities), Region III – Algarve and Lower Alentejo (17 municipalities), and the Low Risk region with the remaining municipalities (207 municipalities).

As a final note, it is essential to emphasize how Structured Additive Regression (STAR) models offer a rich framework that allows the presence of a wide range of covariates while simultaneously exploring possible spatial and temporal correlations within a very diverse type of response variables.

Acknowledgments

This work was supported by the Portuguese National Funding Agency for Science, Research and Technology, *Fundação para a Ciência e Tecnologia – Ministério da Educação e Ciência*, through the research project [PTDC/SAU-SAP/116950/2010]. The third and fourth authors are also supported through the project [UID/MAT/04674/2019].

References

1. World Health Organization. Global Tuberculosis Report 2018. Geneva: World Health Organization, 2018.
2. European Centre for Disease Prevention and Control/WHO Regional Office for Europe. Tuberculosis surveillance and monitoring in Europe 2014. Stockholm: European Centre for Disease Prevention and Control, 2014.
3. Nunes C, Briz T, Gomes D, & Filipe PA (2011). Pulmonary Tuberculosis and HIV/AIDS: joint space-time clustering under an epidemiological perspective. In: Cafarelli B, editors. Proceedings of the Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition; Foggia e Gargano, p. 1-4.
4. Kneib T (2006). *Mixed model based inference in structured additive regression*. PhD Thesis. Munchen: Universität Munchen, Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians.
5. Fahrmeir L, Kneib T, & Lang, S (2004). Penalized structured additive regression for space-time data - a bayesian perspective. *Stat Sinica*, 14:731-761.
6. Lang S, & Brezger A (2004). Bayesian P-splines. *J Comput Graph Stat.*, 13:183-212.
7. Brezger A, & Lang S (2006). Generalized additive regression based on Bayesian P-splines*. Comput Stat Data An.*, 50:967-991.
8. Rue H, & Held L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: Chapman & Hall/CRC.
9. Fahrmeir L, & Lang S (2001). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Fields Priors. *J R Stat Soc C-Appl.*, 50(2):201-220.
10. Osei FB, Duker AA, & Stein A (2012). Bayesian structured additive regression modeling of epidemic data: application to cholera. *BMC Med Res Methodol.*, 12(118).
11. Klein N, Kneib T, & Lang S (2015). Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110:509, 405-419, DOI: 10.1080/01621459.2014.912955.

Spatial and temporal trends in non-monetary wealth in Latin America (1990-2010)

Rodrigo Lovatón, Sula Sarkar
Institute for Social Research and Data Innovation, University of Minnesota
Minneapolis-Minnesota, United States

Abstract

Research on poverty has been more often focused on monetary measures such as household expenditures or income. In contrast, the use of non-monetary indicators of wealth provides different insights about inequality and development over time. In this paper, we examine spatial and temporal trends in household wealth in 10 countries in Latin America through a non-monetary approach. The analysis takes advantage of census micro data samples covering the 1990, 2000, and 2010 rounds. We focus on a set of nine indicators that are common across countries and census years. Preliminary results show progress for the countries examined, but that is uneven and biased towards the population residing in urban areas.

Keywords

poverty; assets; principal component analysis; census; IPUMS

1. Introduction

Poverty reduction remains an important public policy objective in Latin America. In this paper, we analyze spatial and temporal trends in household wealth in 10 countries in the region from an asset-based perspective. This analytical approach has been extensively used when traditional monetary measures such as household expenditures or income are not available, in a similar manner to the early application developed by Filmer and Pritchett (2001). Previous research on related topics for Latin America includes the unsatisfied basic needs framework, development of multidimensional poverty measures, and inequalities in human capital accumulation (e.g. Attanasio and Székely, 1999; Hammill, 2009; Permanyer, 2013). This study intends to specifically exploit the availability of various non-monetary indicators collected in census microdata to assess changes in household wealth over a relatively longer time period.

The analysis of poverty from a non-monetary approach has been implemented in previous studies for Africa using data from the Demographic and Health Surveys (DHS). Stifel, Sahn, and Younger (1999) examine different poverty indicators across nine countries in Africa, finding mixed results in terms of progress depending on the chosen variable. Sahn and Stifel (2000) compare poverty at different points in time for various African countries using

an index based on household characteristics, durables, and household heads' education. Their results show declines in poverty during the nineties decade, mainly due to advance achieved in rural areas. However, Sahn and Stifel (2003) investigate urban-rural differences using several living standards indicators, including an asset index, and find gaps that are not diminishing over time. Finally, Booysen et al (2008) create an asset index to examine changes in seven African countries and conclude that progress is mainly associated to accumulation of private assets, as compared to the setback in the access to public services. This paper will close a gap in the literature by carrying out a similar analysis for the Latin America region using census microdata.

2. Methodology

This study uses census microdata from the IPUMS International project, the largest repository of international census samples. Household wealth is observed across the 1990, 2000, and 2010 census rounds. The data include 24 census samples from 10 countries, which cover at least for two of the referenced rounds. The following censuses are employed for the analysis: Argentina 1991, 2001, and 2010, Bolivia 1992 and 2001, Brazil 1991, 2000, and 2010, Chile 1992 and 2002, Colombia 1993 and 2005, Ecuador 1990, 2001, and 2010, Paraguay 1992 and 2002, Peru 1993 and 2007, Uruguay 1996, 2006, and 2011, and Venezuela 1990 and 2001. The samples for each of these censuses represent 5 to 10% of the country's population. The data provided by IPUMS are harmonized (i.e. it uses a consistent coding structure for the same variables), which facilities the construction of indicators and enhances comparability for the analysis.

Household wealth is defined using dwelling characteristics and human capital. We identified a set of indicators at the household level that are common across all datasets. The choice of these indicators is based on previous literature (Booysen et al, 2008; Lovaton et al, 2014; Sahn & Stifel, 2000) and it also follows some indicators comprised within the Sustainable Development Goals (SDGs) framework. The calculation of indicators not only uses consistent definitions across countries and census rounds to achieve higher comparability, but it is also based on harmonized microdata offered by IPUMS. The set of nine non-monetary wealth indicators analyzed are: 1) whether the dwelling is owned or rented, 2) access to electricity, 3) access to piped water, 4) connection to public sewage, 5) whether the dwelling has a toilet or a bathroom, 6) finished floors, 7) cement, brick or concrete walls or roof, 8) number of persons per room, and 9) years of schooling of the household head. A summary measure is produced to assess overall household wealth through the aggregation of these nine indicators. This measure is defined as a linear combination of the indicator variables by applying

appropriate weights to each of them. If we consider weights w_i and wealth indicators x_i, the index is defined as:

$$WI = W'X = w_1x_1 + w_2x_2 + \cdots + w_9x_9$$

We use two alternative approaches to calculate the weights for this household wealth measure. First, we calculate the weights through principal component analysis (PCA). This data-reduction technique produces weights by identifying the directions of larger data variability. PCA is applied to a pooled dataset including all census samples, so that each indicator receives the same weights across countries and years (similar to Booysen et al, 2008, Sahn and Stifel, 2000). The household wealth index is then created from the first principal component of the data. Given the variance-covariance matrix of the data $\sum$, then PCA derives the weights w_j from the following optimization problem:

$$Max\ VAR(W'X) = W'\sum W$$

$$subject\ to\ W'W = 1$$

Second, weights are also produced by estimating a model for household expenditures, where we use each of the nine indicators as explanatory variables. For this purpose, we rely on household surveys that are contemporary with the most recent census year as an additional data source, given that expenditures or income are rarely included in census microdata. Therefore, in order to carry out this analysis, the supplementary data source must include the same set of variables and household expenditures. Given household expenditures E, the weights w_i are estimated using the regression equation below. The household wealth index corresponds to predicted expenditures using these estimated weights.

$$E = w_0 + w_1x_1 + w_2x_2 + \cdots + w_9x_9 + \varepsilon$$

Based on each of these two alternative wealth indices, we examine changes in poverty. Two definitions of poverty are operationalized with the data. First, households are considered poor if they are at the bottom 40% of the wealth distribution, based on the pooled data for a specific country. Second, we identify a set of minimum household characteristics that would be necessary to achieve a predicted expenditure equivalent to the poverty line used by the country. By using these two definitions, we identify whether poverty increased or decreased over time, and how it was spatially distributed.

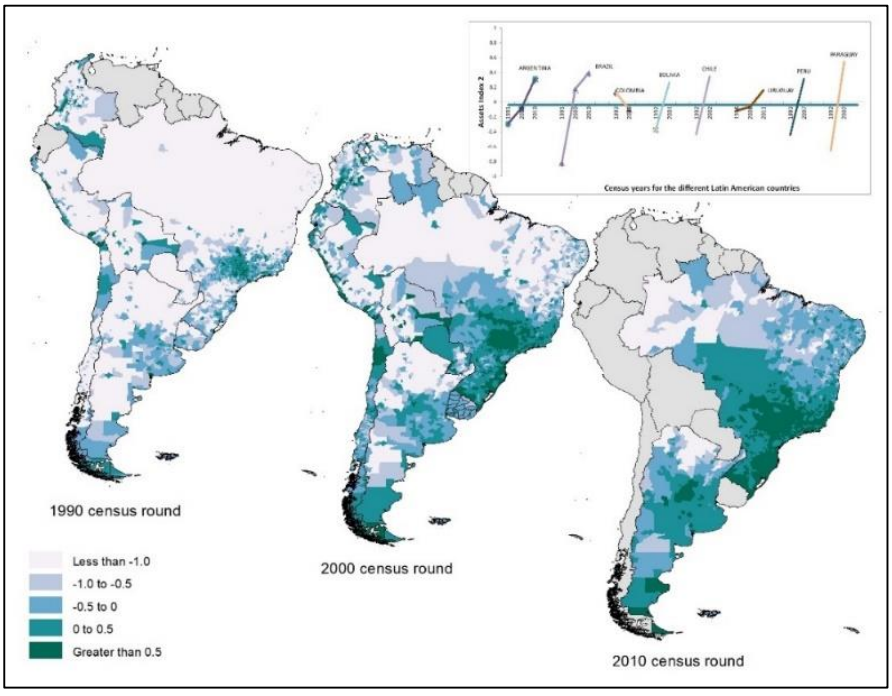
The spatial component of the analysis faces a major challenge posed by changes in administrative boundaries over time. Researchers interested in

analyzing the changes in non-monitory wealth over time and across countries, need to hold space constant. One of the biggest hurdles is the question of whether, and to what extent, geographic boundaries change across census years. Until now, little has been done to verify the spatial areas corresponding to coded units in the census microdata. Even less has been done to research spatial changes across time. Users of census microdata are limited by the timing of censuses (typically every 5 or 10 years) and by the unit levels identified in the data (typically administrative divisions within country). However, given the rise in digital mapping capabilities and spatial analytical technologies, the IPUMS census data collection has created integrated geographical units at the first and second administrative level of geography. The integrated geographical units take into consideration changing boundaries, the temporal aspect of the data from multiple censuses, and the scalar aspect by considering the different administrative levels of geography. The work involves extensive metadata acquisition, research, and verification (acquisition and correspondence); the creation of small-area building blocks that cover consistent spatial extent over time (harmonization); the testing and implementation of techniques to group spatial units to meet the 20,000 persons threshold (regionalization); and the development of GIS shapefiles and variables (map and variable creation) (Sarkar et al, 2015).

3. Result

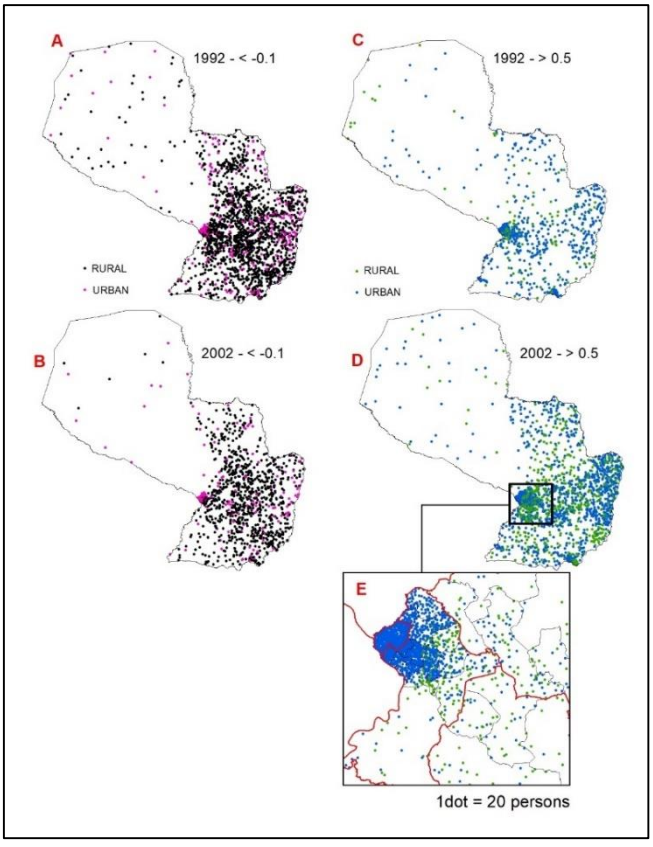
Preliminary results presented here include the calculation of a wealth index using weights produced through PCA. We visualize our results with maps of South America representing the three census rounds covered in the data (1990, 2000, and 2010). We used GIS software Arc View 10.3 to map all our results. Figure 1 displays the mean value for the asset index at the second administrative level of geography; e.g. the maps represent *municipios* for Brazil, Colombia, and Chile and provinces for Peru. The mean values are divided into five groups following the natural breaks in the wealth index distribution. Along with the maps, Figure 1 includes graphs for the national changes in the mean asset index over time. Since space is held constant, it is easy to analyze changes in the standard of living from one census year to another for the different countries. Based on these maps, the asset index shows visually an overall improvement for all countries. Colombia is the only country where the mean asset index actually decreased over time. If we look at the changes at the *municipio* level in Figure 1, the decrease seems to be restricted to the sparsely populated eastern part of the country. The densely populated Bogota area and the western parts of the country show an improvement in both indices over the years. Our final paper will also analyze the median asset index along with the mean values to adjust for such inconsistencies.

Figure 1: South America, mean values of asset index at the second administrative level of geography; Data source: Integrated Public Use Microdata Series (IPUMS) International



Gaps in progress between urban and rural areas are also observed in our results, similarly to previous research (Sahn and Stifel, 2003). Figure 2 represents the number of people with low and high asset index values in Paraguay by urban or rural area of residence, where the cutoff values coincide with the top and bottom groups from the previous figure. Note that rural population in Paraguay is 31% of total population in 1992 and 29% of total population in 2002, such that we do not observe a significant increase in urbanization over a span of 10 years. Figures 2A and 2B represent all rural and urban population with an asset index value lower than -0.1 in the years of 1992 and 2002, respectively. The rural population is represented with black dots and the urban population is represented with pink dots, where each dot corresponds to 20 people. Results show a decrease in the number of both rural and urban population with an asset index lower than -0.1 between 1992 and 2002. Figure 2C and 2D show the higher cutoff for the asset index, where the rural population is represented with green dots and the urban population with blue dots. Figure 2E is a zoomed portion of Figure 2D where we only show the urban area of Asuncion in Paraguay. The figures show an overall increase in population in the high asset group over time, which is significantly higher for people residing in urban areas.

Figure 2: Paraguay, population with low and high values of the asset index, by urban/rural location and at the second administrative level of geography; Data source: Integrated Public Use Microdata Series (IPUMS) International.



4. Discussion and Conclusion

Preliminary results for the countries included in the study show overall improvements in wealth from an asset-based perspective for most of the samples analyzed (except for Colombia) between the census rounds of 1990, 2000, and 2010. Even though the national trend for Colombia shows a decline in household wealth, the spatial distribution of the asset index highlights specific areas that did achieve progress over time. Furthermore, improvements based on the set of indicators analyzed are found to be uneven and appear to be biased towards the population residing in urban areas. Thus, the use of a non-monetary approach to examine wealth and poverty provides additional insights with respect to traditionally used measures such as household expenditures or income. The final version of this paper will compare asset-based wealth using principal components analysis against indices predicting household expenditures. In addition, we will measure the gaps between urban and rural areas, the magnitude of spatial inequalities, and decompose the

overall change in wealth to identify which indicators are driving progress for each of the countries analyzed.

References

1. Attanasio, O. & Székely, M. (1999). An asset-based approach to the analysis of poverty in Latin America. Washington D.C.: Inter-American Development Bank, working paper R-376.
2. Booysen, F., Van Der Berg, S., Burger, R., Von Maltitz, M. and Du Rand, G. (2008). Using an asset index to assess trends in poverty in seven sub-Saharan African countries. World Development, Vol. 36, No. 6, pp. 1113-1130.
3. Filmer, D. & Pritchett, L. (2001). Estimating wealth effects without expenditure data-or tears: An application to educational enrollments in states of India. Demography, Vol. 38, No. 1, pp. 115-132.
4. Hammill, M. (2009). Income poverty and unsatisfied basic needs. Mexico City: ECLAC, LC/MEX/L.949.
5. Lovaton, R., McCarthy, A., Gondwe, D., Kirdruang, P., & Sharma, U. (2014). Water, walls and bicycles: Wealth index composition using census microdata. Minneapolis: University of Minnesota, Minnesota Population Center, working paper 2014-7.
6. Permanyer, I. (2013). Using census data to explore the spatial distribution of human development. World Development, Vol. 46, pp. 1-13.
7. Sahn, D. & Stifel, D. (2000). Poverty comparisons over time and across countries in Africa. World Development, Vol. 28, No. 12, pp. 2123-2155.
8. Sahn, D. & Stifel, D. (2003). Urban-rural inequality in living standards in Africa. Journal of African Economies, Vol. 12, No. 3, pp. 564-597.
9. Sarkar, S., Cleveland, L., Silisyene, M. & Sobek, M. (2015). Harmonized census geography and spatio-temporal analysis: gender equality and empowerment of women in Africa. Paper presented at the Annual Meeting of the Population Association of America, San Diego, CA in April.
10. Stifel, D., Sahn, D., & Younger, S. (1999). Inter-temporal changes in welfare: Preliminary results from ten African countries. Ithaca: Cornell University, Cornell Food and Nutrition Policy Program working paper 94.

Using SOM-based visualization to analyse the financial performance of consumer discretionary firms

Dominique Haughton[1,2,3], Zefeng Bai[1], Nitin Jain[1], Ying Wang[1]
[1]Bentley University
[2]Université Paris 1 (SAMM)
[3]Université Toulouse 1 (TSE-R)

Abstract
This paper analyzes financial ratios of 27 consumer discretionary firms listed on the S&P 500 over an eleven-year period from 2006-2016. It adopts a two-step approach wherein first a confirmatory factor analysis (CFA) on the financial time-series is conducted and the resulting constructs' scores are then used to perform a cluster analysis using self-organizing maps (SOMs). The consumer discretionary sector is considered an economic and stock market predictor. It consists of non-essential goods and services which in an economic slump are more likely to be foregone. The suggested approach is expected to be a useful reference guide to help understand the past performance of inter- and intra-sector companies. It also enriches the body of literature on the application of machine learning techniques to the analysis of firm- and sectoral-level performance.

Keywords
Consumer Discretionary Sector; Clustering; Financial Ratios; Self-Organizing Maps; Time series

1. Introduction

The advent of machine learning has lent a new dimension to the analysis of financial and accounting ratios. A growing body of research integrates machine learning techniques – both supervised and unsupervised – to bring out useful insights from these ratios, beyond the traditional approach to analysing financial ratios. This paper contributes to this research by proposing a two-step dynamic process to facilitate understanding firms from the consumer discretionary sector in the US.

This paper considers the consumer discretionary firms due to the inherent nature of this sector. It consists of goods and services that are not essential but are desirable if income is sufficient to purchase them. Therefore, lower stock values in this sector, which includes durable goods, apparel, entertainment and leisure, etc., can be considered as a signal to an economic slump. Such stock tend to outperform other sectors' stock during strong economic times and underperform them during an economic slump. It is

therefore particularly interesting to explore the dynamics of firms from this sector.

Most of the literature focuses on the financial services or the information technology sectors. Hence, this paper adds value to the sector analysis also. Further, the paper also considers a longer time-window of 11 years from 2006-16. This serves two purposes: it encompasses the sub-prime period and at least one complete economic cycle.

The findings from the analyses bring out interesting perspectives. For example, they show how Amazon differs from its peers in the sector, how Macy's can be differentiated from others on certain financial aspects, etc. Such insights can help investors understand these firms better and make their investment decisions accordingly.

2. Methodology

The two-step approach adopted in this paper includes a dimension reduction of the indicators, i.e., the financial ratios, and using the dimension scores as input variables for non-linear clustering analyses using SOM.

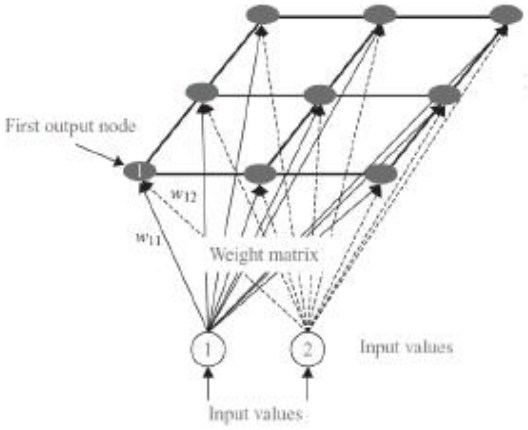
SOM is a type of competitive neural network that projects a high-dimensional input space on prototypes of a low-dimensional regular grid. It accomplishes two goals: to reduce dimensions and to display similarities among prototypes. The algorithm aims at clustering together similar observations while preserving the original topology of the data (i.e., similar observations in the input space are clustered together into the same unit or into neighbouring units on the map) (Olteanu, M., & Villa-Vialaneix, N, 2015). Specifically, in contrast with other artificial neural networks which apply error correction learning, SOM is based on competitive learning, where the output nodes compete among themselves to produce the winning node (or neuron). Only the winning node and its neighbourhoods are activated by a particular input observation. This architecture allows SOM to preserve the topological properties of the input space. Therefore, SOM can be effectively utilized to visualize and explore the properties of the data.

SOM is a two-layer feed forward and completely connected network as shown in Figure 1. The data from the input layer are passed along directly to the output layer. The output layer is represented in the form of a grid, usually in one or two dimensions, and typically in the shape of a rectangle or hexagon. The algorithm can be summarized in the following 3 steps:

1. Initialization: initialize the neurons' weights;
2. Competing: ①compute the scoring function (such as Euclidean distance) for each output node; ② locate the winning node (the closest match with the input)
3. Learning: update the weight vectors of the winning node and its neighbours using a linear combination of the input vector $x(t)$ and the

current weight vector $w_i(t)$: $w_i\ (t\ +\ 1)\ =\ w_i(t)\ +\ h_i\ (t)[x(t)\ -\ w_i\ (t)]$, where $h_i(t)$ is a neighbourhood function, i represents the winning node. The simplest neighbourhood function is a monotonically decreasing Gaussian function: $h_i(t) = \alpha\ *\ \exp(\frac{-d(i,w)}{2\sigma^2(t)})$, where α represents the learning rate, and $\sigma(t)$ is the width of the kernel.

Figure 1: SOM with 2 - D Input, 3 x 3 Output, obtamed from "DATA MINING Concepts, Models, Methods, and Algorithms"



3. Results

The data were sourced from Bloomberg for the 67 consumer discretionary firms listed on the S&P 500 for the years 2006-2016. During the data pre-processing stage, it was decided to consider 27 firms with 2006-16 as the time-window and 19 financial ratios since it is prudent to not impute missing values in a financial time-series, which can display unusual jumps. The 11-year period covers at least one full economic cycle and encompasses the sub-prime period.

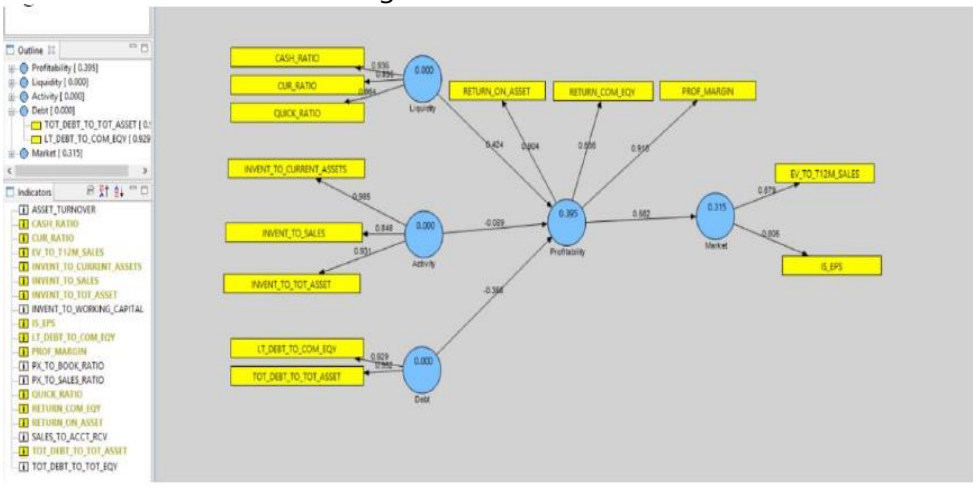
The first-step involved building a confirmatory factor analysis (CFA) model. Nineteen indicators were used initially and were pruned to thirteen listed in Table 1 after removing variables during an iterative procedure. The final CFA model is displayed in Figure 2. The SmartPLS 2.0 package was used to carry out this analysis.

As shown in Figure 2, we observe a positive association between liquidity and profitability, and we see that profitability decreases as activity or debt increases. Eventually, profitability positively contributes to the market factor.

Table 1: List of variables for the CFA analysis

S.no.	Financial Ratio
1	Return on Assets
2	Return on Common Equity
3	Profit Margin
4	Current Ratio
5	Quick Ratio
6	Cash Ratio
7	Total Debt to Total Asset
8	Long-term Debt to Common Equity
9	Inventory to Sales
10	Inventory to Total Assets
11	Inventory to Current Assets
12	Basic Earnings Per Share
13	Enterprise Value to 12M Sales

Figure 2: CFA model



The second step involves constructing a SOM on the 5 indicators extracted from the CFA analysis. In the application of self-organizing maps, we rely on the R based SOMbrero package. The map size is decided after testing several sizes (3*3, 4*4, 5*5) of the SOM to check that the cluster structures are shown with sufficient resolution and an acceptable number of empty nodes. We adopt a 4 by 4 grid in which less than 50% of the nodes are empty. The algorithm used is described in the previous sub-section, with the Euclidean distance and Gaussian neighbourhood function. Before implementing, each of the five indicators is centred and rescaled. To visualize the dynamics over years, we treat each year as one input dimension. Therefore, we have five maps corresponding to the five indicators. For each indicator, we have 11

dimensions representing the data from year 2006 to 2016. For each dimension, there are 27 observations from the 27 companies.

4. Discussion

The debt ratio construct provides a quick measure of the amount of debt that the company has on its balance sheets compared to its assets. It shows how much the company relies on debt to finance assets. Usually, the higher the ratio, the greater the risk associated with the firm's operation. A low debt ratio indicates conservative financing with an opportunity to borrow in the future at no significant risk. For the sake of brevity we present results related to the debt ratio construct; similar analyses have been conducted for the other constructs.

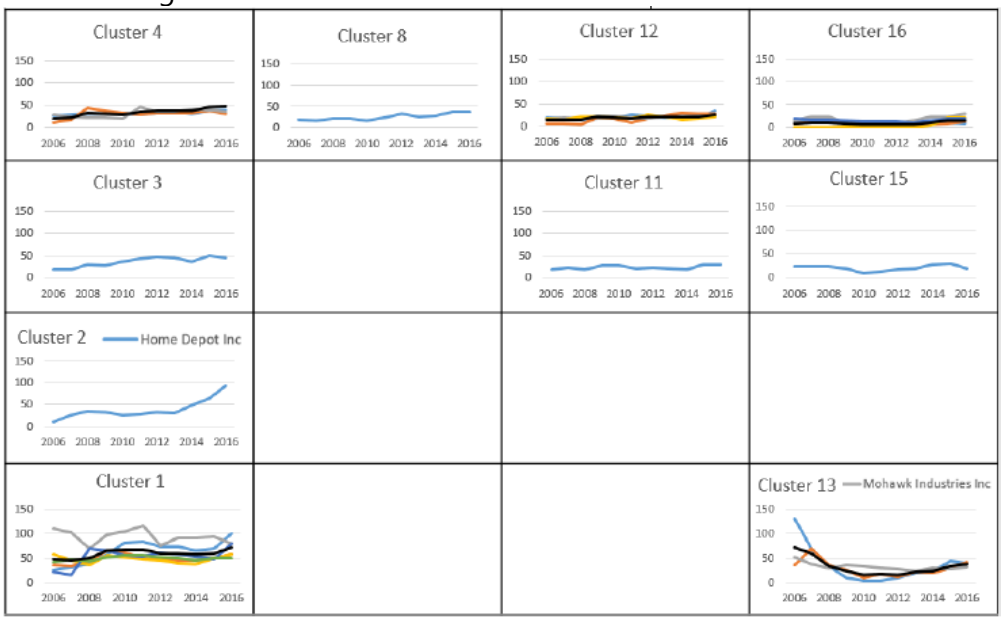
Figure 3: Debt ratio SOM clusters

Cluster 4 LKQ Corp Leggett & Platt Inc PVH Corp	**Cluster 8** Mattel Inc	**Cluster 12** BorgWarner Inc Tiffany & Co O'Reilly Automotive Inc VF Corp	**Cluster 16** Starbucks Corp Garmin Ltd Genuine Parts Co TJX Cos Inc Ralph Lauren Corp Tapestry Inc
Cluster 3 Hasbro Inc	**Cluster 7**	**Cluster 11** Whirlpool Corp	**Cluster 15** Advance Auto Parts Inc
Cluster 2 Home Depot Inc	**Cluster 6**	**Cluster 10**	**Cluster 14**
Cluster 1 Newell Brands Inc Macy's Inc MGM Resorts International Royal Caribbean Cruises Ltd Harley-Davidson Inc Nordstrom Inc	**Cluster 5**	**Cluster 9**	**Cluster 13** Booking Holdings Inc Mohawk Industries Inc Amazon.com Inc

From Figure 3, we see that cluster 1 and cluster 16, at the two opposite corners of the map, possess the largest number of observations; and 6 out of 14 clusters in between are empty. This indicates that the debt ratio in the consumer discretionary sector presents a polarized situation. Both

conservative and bold financing strategies are adopted. It is interesting to note that the companies within a cluster present a heterogeneous mix of business content and industry. This means that the temporal dynamics of companies' financial performance seems not determined by traditional classification indicators, such as by industry or line of business, and that these traditional classification indicators are insufficient to underpin an efficient comparison among companies.

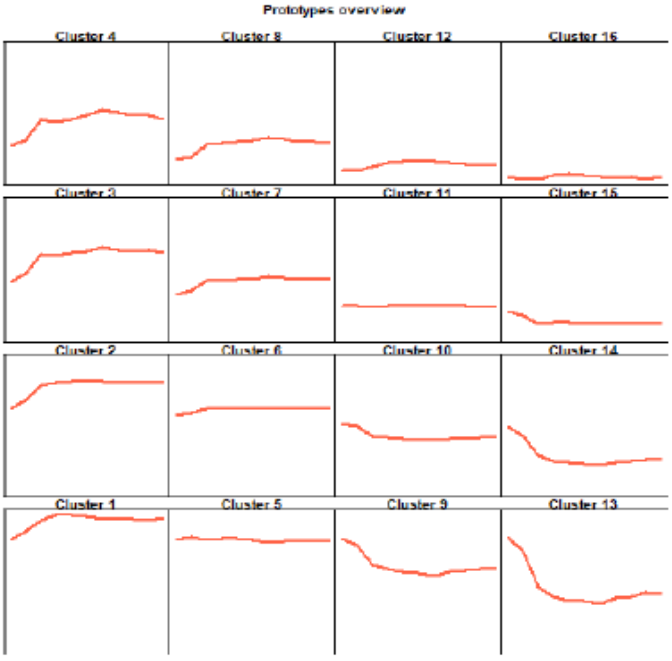
Figure 4: Debt ratio constructs time series in each cluster



In terms of temporal dynamics, it seems that the majority of companies seek stable financial leverage in the long-run (see Figure 4); the fluctuation in the short run may reflect the inefficiency in the process of debt issuance. However, we identify two clusters (cluster 2 and cluster 13) that act differently. Cluster 2 shows a strong upward momentum at the beginning and end stages, while flat in the middle stage. Cluster 13 shows a downward trend during the first half and gradually becomes flat during the second half. Further research regarding the particular companies in these two clusters indicates that these unique behaviours of financial leverage signal important changes within the firms. For example, the excessive leverage showed in cluster 2 is likely associated with Home Depot's aggressive engagement in buybacks in the stock market; the debt behaviour in cluster 13 may be highly associated with MHK's acquisition in 2005 which is financed with debt.

Under the framework of SOM, the comparison can be done by treating prototypes (Figure 5) as benchmarks.

Figure 5: Debt ratio SOM prototypes



Finally, the distance relationship among prototypes is revealed by Figure 6. If we consider a long distance with neighbourhood clusters as a signal of uniqueness, we find that cluster 1 is the most unique cluster.

Figure 6: Debt ratio SOM distance polygons

5. Conclusion

In this paper, we have adopted an innovated technique, self-organizing map, to capture and visualize the temporal dynamics of financial performance of companies in the consumer discretionary sector in the United States. Strong temporal differences among clusters are revealed by SOM. This analysis allows for identifying the signature behaviour in clusters of companies, recalling the unique debt behaviour (excessive leverage) of Home Depot and the strong profitability secured by TJMaxx during the financial crisis

References

1. Back, B., K. Sere, and H. Vanharanta (1998). Analyzing financial performance with self-organizing maps. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, Volume 1, pp. 266–270. IEEE.
2. Barreto, G. A. (2007). Time series prediction with the self-organizing map: A review. In *Perspectives of neural-symbolic integration*, pp. 135–158. Springer.
3. Blazejewski, A. and R. Coggins (2004). Application of self-organizing maps to clustering of high-frequency financial data. In *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalization,* Volume 32, pp. 85–90. Australian Computer Society, Inc.
4. Deboeck, G. J. (1998). Financial applications of self-organizing maps. *Neural Network World 8*(2), 213–241.
5. Edler, L. (2007). Analysing economic data with self-organizing maps.
6. Eklund, T., B. Back, H. Vanharanta, and A. Visa (2002). Assessing the feasibility of self-organizing maps for data mining financial information. *ECIS 2002 Proceedings*, 140.
7. Fu, T.-c., F.-l. Chung, V. Ng, and R. Luk (2001). Pattern discovery from stock time series using self-organizing maps. In *Workshop Notes of KDD2001 Workshop on Temporal Data Mining*, pp. 26–29.
8. Gafiychuk, V., B. Y. Datsko, and J. Izmaylova (2004). Analysis of data clusters obtained by self-organizing methods. *Physica A: Statistical Mechanics and its Applications 341*, 547–555.
9. Guo, C., H. Jia, and N. Zhang (2008). Time series clustering based on ICA for stock data analysis. In *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08. 4th International Conference on*, pp. 1–4. IEEE.
10. Kiviluoto, K. and P. Bergius (1997). Analyzing financial statements with the self-organizing map. In *Proceedings of WSOM*, Volume 97, pp. 4–6. Citeseer.
11. Kiviluoto, K. and P. Bergius (1998). Two-level self-organizing maps for analysis of financial statements. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, Volume 1, pp. 189–192. IEEE.
12. Kohonen, T. (2013). Essentials of the self-organizing map. *Neural networks 37*, 52–65.
13. Lev, B. and S. Sunder (1979). Methodological issues in the use of financial ratios. *Journal of Accounting and Economics 1*(3), 187–210.

14. Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition 38*(11), 1857–1874.
15. Merkevi˘cius, E., G. Gar˘sva, and R. Simutis (2004). Forecasting of credit classes with the self-organizing maps. *Information technology and control 33*(4).
16. Moreno, D., P. Marco, and I. Olmeda (2006). Self-organizing maps could improve the classification of Spanish mutual funds. *European Journal of Operational Research 174*(2), 1039–1054.
17. Olteanu, M., & Villa-Vialaneix, N. (2015). Using SOMbrero for clustering and visualizing graphs. *Journal de la Société Française de Statistique*, *156*(3), 95-119.
18. Rani, S. and G. Sikka (2012). Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications 52*(15).
19. Sarlin, P. (2015). Data and dimension reduction for visual financial performance analysis. *Information Visualization 14*(2), 148–167.
20. Sarlin, P. and T. Eklund (2011). Fuzzy clustering of the self-organizing map: some applications on financial time series. In *International Workshop on Self-Organizing Maps*, pp. 40–50. Springer.
21. Sarlin, P., Z. Yao, and T. Eklund (2012). A framework for state transitions on the self-organizing map: Some temporal financial applications. *Intelligent Systems in Accounting, Finance and Management 19*(3), 189–203.
22. Shih, J.-Y. (2011). Using self-organizing maps for analyzing credit rating and financial ratio data. In *Business Innovation and Technology Management (APBITM), 2011 IEEE International Summer Conference of Asia Pacific*, pp. 109–112. IEEE.
23. Silva, B. and N. C. Marques (2010). Feature clustering with self-organizing maps and an application to financial time-series for portfolio selection. In *IJCCI (ICFC-ICNC)*, pp. 301–309.
24. Simon, G., A. Lendasse, M. Cottrell, J.-C. Fort, and M. Verleysen (2005). Time series forecasting: Obtaining long term trends with self-organizing maps. *Pattern Recognition Letters 26*(12), 1795–1808.
25. Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on neural networks*, *11*(3), 586-600.
26. Wang, Y.-J. and H.-S. Lee (2008). A clustering method to identify representative financial ratios. *Information Sciences 178*(4), 1087–1097.

Forecasting conditional covariance matrices in high-dimensional data: A general dynamic factor approach

Carlos Trucíos[1], Pedro Valls[1], João Mazzeu[2], Maurício Zevallos[2], Luiz Hotta[2], Marc Hallin[3]

[1]Sao Paulo School of Economics, FGV, Sao Paulo, Brazil
[2]Department of Statistics, University of Campinas, Campinas, Brazil
[3]ECARES, Universite libre de Bruxelles, Brussels, Belgium

Abstract

In this paper, we use the General Dynamic Factor Model with infinite-dimensional factor space to develop new estimation and forecasting procedures for conditional covariance matrices in high-dimensional data. The performance of our approach is evaluated via Monte Carlo experiments and yield excellent finite-sample properties. The new procedure is used to construct minimum variance portfolios in a high-dimensional real dataset. The results are shown to achieve better out-of-sample portfolio performance than alternative existing procedures.

Keywords

Dimension reduction; Dynamic factor models; Large panels; High-dimensional time series

1. Introduction

Forecasting volatilities plays an important role in a variety of economic and financial applications, such as portfolio allocation, risk management, option pricing, hedging strategies, etc. Several multivariate methods have been proposed to model and forecast the conditional covariance matrix of a collection of assets. Unfortunately, most of those methods suffer from the so-called as the number of assets grows, and cannot be implemented in a high-dimensional context. Therefore, alternative procedures have been proposed. One of them is factor models with high-dimensional asymptotics, which offer a promising alternative in that context. Factor models are based on the assumption that the prices and volatilities of different assets are driven by a small number of latent factors, which account for their co-movements. They have been used by several authors to model and forecast conditional covariance matrices. These works are based on a *static* factor-loading scheme, the main advantage of which is to allow for estimation methods based on traditional principal components. This approach is easy to implement and widely used in empirical application. However, as pointed out, for instance, in Section 1.1 of Forni et al. (2015), the assumption of a static factor-loading scheme considered in that literature is quite restrictive and rules out some very

simple and plausible cross-correlation patterns leading to infinite-dimensional factor spaces. To overcome this issue, Forni et al. (2000) introduced the so-called *generalized or general dynamic factor model* (GDFM), in which factors (equivalently, common shocks) are loaded through filters rather than matrices. As shown in Hallin and Lippi (2013), the GDFM actually follows from a representation result which holds, essentially, without placing any restrictions-beyond second-order stationarity and the existence of a spectrum-on the data-generating process. Nevertheless, as far as we know, the procedure has never been used to estimate conditional covariances, which is the main goal of this paper. We propose a new procedure for modelling and forecasting the conditional covariance matrix in a high-dimensional context, based on Forni et al. (2015, 2017). The rest of the paper is organised as follows. Section 2 describes the GDFM and our forecasting method. Section 3 reports a Monte Carlo study of the finite-sample properties of the proposed procedure. In Section 4, we apply the new procedure in a large collection of assets. Finally, Section 5 presents the main conclusions.

2. The general dynamic factor model

In this section, we describe the GDFM to be used throughout, which essentially contains all other factor models considered in the econometric and time series literature as particular cases. Let $\{X_t = (X_{1t}\, X_{2t} \ldots)', t \in \mathbb{Z}\}$, be a double-indexed zero-mean second-order stationary stochastic process, where the first index refers to assets and t stands for time. The GDFM is defined as

$$X_{it} = \chi_{it} + \xi_{it}, \qquad \text{with} \tag{1}$$

$$\chi_{it} = \sum_{j=1}^{q}\sum_{k=0}^{\infty} b_{ijk}u_{jt-k} = \mathbf{b}_i'(L)\mathbf{u}_t \text{ and } \xi_{it} = \sum_{k=0}^{\infty} d_{ik}v_{it-k} = d_i(L)v_{it}, \tag{2}$$

where χ_{it}, the common component, ξ_{it}, the *idiosyncratic component*, $\mathbf{u}_t = (u_{1t}u_{2t}\ldots u_{qt})'$ the process of common shocks driving the common components, and $\mathbf{v}_t = (v_{1t}v_{2t}\ldots v_{qt})'$, the process of idiosyncratic shocks driving the idiosyncratic components, are all non-observable. Then, letting $\mathbf{X}_n := \{X_{it} | i = 1,\ldots,n, t \in \mathbb{Z}\}$ and $\boldsymbol{\xi}_n := \{\xi_{it} | i = 1,\ldots,n, t \in \mathbb{Z}\}$, equation 2 in vector notation takes the form

$$\mathbf{X}_{nt} = \mathbf{B}_n(L)\mathbf{u}_t, \boldsymbol{\xi}_{nt} = \mathbf{D}_n(L)\mathbf{v}_{nt}, \tag{3}$$

with $\mathbf{B}_n(L) := (\mathbf{b}_1(L)\ldots\mathbf{b}_n(L))'$, and $\mathbf{D}_n(L) := (\mathbf{d}_1(L)\ldots d_n(L))'$. Additional assumptions that guarantee the existence of the spectral densities of $\mathbf{X}_t$, χ_t and $\boldsymbol{\xi}_n$, denoted by $\sum_n^X(\theta)$, $\sum_n^\chi(\theta)$ and $\sum_n^\xi(\theta)$, $\theta \in [-\pi, \pi]$, respectively, and others can be found in Trucíos et al. (2019).

Forni et al. (2015), under slightly different assumptions, obtain one-sided filters and construct estimators for (2); slightly reinforcing the same assumptions (e.g., assuming that $\mathbf{u}_t$ is i.i.d.), Forni et al. (2017) derive a complete asymptotic analysis for the same estimators. On the other hand, Barigozzi and Hallin (2018) do not rely on the i.i.d assumption and, under (i)-(ix) and additional assumptions, provide consistency and consistency rates of those estimators.

The main theoretical result behind the one-sided approach of Forni et al. (2015) is the existence of a block-diagonal VAR filtering of the observations turning the GDFM representation (1) into a static one as:

$$\mathbf{A}(L) := \begin{bmatrix} A^1(L) & 0 & \cdots & 0 & \cdots \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & A^m(L) & \cdots \\ \vdots & \vdots & \cdots & \cdots & \ddots \end{bmatrix} \quad and\ \boldsymbol{R} := \begin{bmatrix} \mathbf{R}^1 \\ \mathbf{R}^2 \\ \vdots \\ \mathbf{R}^m \\ \vdots \end{bmatrix}, \tag{4}$$

i.e., the filtered process $\mathbf{Y}_t := \mathbf{A}(L)\mathbf{X}_t$ admits a *static* factor model representation

$$\mathbf{Y}_t \ = \mathbf{R}\mathbf{u}_t \ + \boldsymbol{\epsilon}_t\,, t \in \mathbb{Z} \tag{5}$$

with q-dimensional factor space spanned by $\mathbf{u}_t$. While $\mathbf{R}$ and $\mathbf{u}_t$ are not individually identified, the product $\mathbf{R}\mathbf{u}_t$ is.

2.1 Predicting the conditional covariance matrix

Although the GDFM, with the one-sided estimation method proposed in Forni et al. (2015, 2017), was successfully applied to forecast inflation and financial returns, so far, it has not been used to estimate one-step ahead conditional covariance matrices, i.e, to estimate $\mathrm{V}(\mathbf{X}_t|\mathcal{F}_t - 1)$, where $\mathcal{F}_t - 1$ is the σ-field generated by $\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots$. The prediction is given by the following proposition and theorem. The proof can be found in Trucíos et al. (2019)

Proposition 1. Consider that assumptions that ensure the existence of the static representation holds. Assume moreover that $\boldsymbol{u}_t$ and $\boldsymbol{\xi}_t$ conditional on $\mathcal{F}_t - 1$, are uncorrelated at all leads and lags. Then, the covariance matrix of $\boldsymbol{X}_t$ conditional on $\mathcal{F}_t - 1$is

$$\mathrm{V}(\boldsymbol{X}_t|\mathcal{F}_{t-1}) \ = \ \boldsymbol{R}\,\mathrm{V}(\boldsymbol{u}_t|\mathcal{F}_{t-1})\,\boldsymbol{R}^{\prime} \ + \ \mathrm{V}(\boldsymbol{\xi}_t|\mathcal{F}_{t-1}) \tag{6}$$

Theorem 2. Under certain assumptions (see Proposition 1 in Barigozzi and Hallin (2018)) and considering $\rho_{nT} = max\left(B_T/\sqrt{T},\ 1/B_T\,, 1/\sqrt{n}\right)$ and $\eta \ = \ O(T^c)$ for some finite $c \ > \ 0$ as n and $T \ \to \ \infty$, where $B_T \ = \ o\left(\sqrt{T}\right)$, we can show that

$$\overline{\boldsymbol{R}}\widehat{\mathbb{V}}(\overline{\boldsymbol{u}}_t/\mathcal{F}_{t-1})\overline{\boldsymbol{R}}' \ + \ \hat{V}\left(\hat{\boldsymbol{\xi}}_t/\mathcal{F}_{t-1}\right) \xrightarrow{p} \mathrm{V}(\mathrm{X_t}|\mathcal{F}_{t-1}) \tag{7}$$

The assumptions, estimation of the models and all the estimators which appear in Eq. (6) and (7) can be found in Trucíos et al. (2019).

3. Monte Carlo experiments

We assess the performance of the proposed procedure through Monte Carlo simulations. The evaluation of conditional variance-covariance forecasts is done through four measures of distance. Truc´ıos et al. (2019) considers four data generating processes (DGPs): the first two DGPs follow static models with one and two common factors, respectively, with factors and idiosyncratic components being heteroscedastic; the third and fourth DGPs are dynamic factor models with finite and infinitedimensional factor spaces, respectively, where the common shocks and the idiosyncratic components are heteroscedastic. The first three DGPs are all particular cases of the GDFM with static representation and can be consistently estimated by the procedure of Alessi et al. (2009). On the other hand, the last DGP cannot be consistently estimated by the procedure of Alessi et al. (2009) since the assumption of finite-dimensional factor space does not hold. In all DGPs the idiosyncratic components satisfy $\boldsymbol{\xi}_t|\mathbf{P}_t \sim N(0, \mathbf{P}_t)$, where $\mathbf{P}_t$ is $N \times N$ N diagonal matrix containing the conditional variances of each idiosyncratic component evolving to a GARCH(1,1) with $V(\xi_{it}) = 1$

DGP3: GDFM with finite-dimensional factor space with $\boldsymbol{u}_t$ following a bivariate DCC(1,1) model as in DGP2. This model is given by

$$\begin{aligned} \mathbf{X}_t &= \Lambda \mathbf{F}_t + \boldsymbol{\xi}_t \\ \mathbf{F}_t &= \Phi\, \mathbf{F}_{t-1} + \mathbf{K}\mathbf{u}_t,\ \ \mathbf{u}_t|\Omega_{t-1} \sim N\,(0, \mathbf{Q}_t), \end{aligned} \tag{8}$$

where $Ft = (F1t\ F2t\ F3t\ F4t)\ 0$, Λ is $n \times 4, \Phi$ is 4×4 and $\mathbf{K}$ is $4 \times q$. This DGP is similar to the one used in Alessi et al. (2009).

DGP4: The GDFM with infinite-dimensional factor space, given by

$$X_{it} = a_{i1}(1 - \alpha_{i1})^{-1}u_{1t} + a_{i2}(1 - \alpha_{i2})^{-1}u_{2t} + \xi_{it},\ \mathbf{u}_t|\mathcal{F}_{t-1} \sim N(0, \mathbf{Q}_t),$$

where $\mathbf{u}_t = (u_{1t}u_{2t})^{-1}$ follows a DCC(1,1) model. More details about the DGP and the simulation can be found in Truc´ıos et al. (2019). The distance between the simulated one-step-ahead conditional covariance matrix $H_{T+1|T}$ and the estimated one $\widehat{\sum}_{T+1|T}$, is given by:

$$\boldsymbol{D}\big(H_{T+1|T}, \widehat{\sum}_{T+1|T}\big) = \sum_{i=1}^{N}\sum_{j=i}^{N} w(i,j)\big(h_{i,j} - \hat{\sigma}_{i,j}\big)^2 \tag{9}$$

where $h_{i,j}$ and $\hat{\sigma}_{i,j}$ are the (i,j) elements of $H_{T+1|T}$ and $\widehat{\sum}_{T+1|T}$, , respectively and $\omega(i,j)$ are weights. We use four distances denoted by D_1, D_2, D_3 and D_4, according with the weights. The weights for D_1 are $w(i,j) = 1\ \forall i$ and j; for D_2, $w(i,j) = 1$ when $i = j$ and 0 otherwise; for D_3, $w(i,j) = 2$ when $\hat{\sigma}_{i,j} > h_{i,j}$ and 1 otherwise; for D_4, $w(i,j) = 2$ when $\hat{\sigma}_{i,j} < h_{i,j}$ and 1 otherwise.

We simulate 500 panels of dimension N=60 and T=1000 for each DGP. Figure 1 reports the results when the DGPs are the GDFMs with finite

(DGP3) and infinitedimensional factor space (DGP4). It reveals that GDFMcc reports a good performance in both cases while ABC only presents a good performance only for the DGP3. In both cases, PCA-(M)GARCH and DCC perform poorly.

4. Empirical application

We one-step-ahead conditional covariance matrix forecast are used to construct the minimum variance portfolio (MVP) with short-sale constraints in a dataset with 656 assets. A rolling window scheme with daily rebalancing have been used. The dataset comprises stocks used in the composition of the indexes S&P 500, National Association of Securities Dealers Automated Quotations (NASDAQ-100) and NYSE Amex Composite Index (AMEX) on July 27, 2018 and traded from January 2, 2011 to June 29, 2018, with a total of N = 656 assets. A window size of T= 750 days is used for estimation and 1134 days used as the out-of-sample period. More details about the simulation can be found in Trucíos et al. (2019).

We use four annualised performance measures to evaluate the out-of-sample portfolio performance, defined as follow: (i) Annualised average portfolio (AV): average of the out-of-sample portfolio returns multiplied by 252; (ii) Annualised standard deviation (SD): standard deviation of the out-of-sample portfolio return multiplied by $\sqrt{252}$; (iii) Annualised information ratio (IR): AV/SD; and (iv) Annualised Sortino ra tio (SR). Because we are selecting the MVP, the performance of the methods should be analysed mainly according to the SD criterion. We compare our proposal with the naive equal-weighted portfolio strategy, the RiskMetrics 2006 methodology, the OGARCH, ABC, generalised principal volatility components (GPVC) of and the procedure called PCA4TS, which extends the principal component analysis to second-order stationary vector time series. The procedures used in our empirical application were chosen due their feasibility in high-dimensional data. The GDFMcc is applied using no permutation (using the order the series appear in the dataset) and also considering 30 permutations. The results, with the rank according to each criterion, are reported in Table 1. They reveal that the best performance is reached by the GDFMcc with 30 permutations according to SD, IR and SR criteria. Note also that, even with no permutation, our proposal reports good results, achieving the second best performance according to all criteria. The OGARCH model has the third best performance, according to the SD criterion, followed by the ABC. The GPVC and the OGARCH procedures exhibit the worst performance according to the AV criterion. ABC has the best performance according to the AV criterion followed by our proposal with none permutation. The worst out-of-sample performance is obtained by the equal-weighted portfolio strategy according to all criteria, except for the AV criterion. Taking into account all criteria, our proposal with permutation exhibits the

best performance followed by our proposal with none permutation and by the ABC procedure.

Table 1: Annualised performance measures: AV, SD, IR and SR. The rank according to each criterion is given in brackets

	AV	SD	IR	SR
1/N	5.7708 (4)	11.5067 (8)	0.5015 (8)	0.6834 (8)
RM2006	5.5983 (5)	4.5447 (5)	1.2318 (6)	1.7229 (4)
OGARCH	4.9227 (7)	4.4551 (3)	1.1050 (4)	1.5614 (6)
ABC	6.5267 (1)	4.5313 (4)	1.4404 (3)	1.9677 (3)
GPVC	4.5989 (8)	4.5889 (6)	1.0022 (7)	1.4077 (7)
PCA4TS	5.3677 (6)	4.7255 (7)	1.1359 (5)	1.6024 (5)
GDFMcc (no permutation)	6.2135 (2)	4.1583 (2)	1.4942 (2)	2.1300 (2)
GDFMcc (30 permutations)	6.2369 (2)	4.0209 (1)	1.5511 (1)	2.2137 (1)

5. Conclusions

Based on the procedure of Forni et al. (2015, 2017), we propose a new procedure to forecast the conditional covariance matrix in high-dimensional data. The main conclusions of the Monte Carlo simulations are that, in all DGPs our proposal has a good performance even in cases where the static representation holds. Specifically, when the DGP is a static factor model, ABC and GDFMcc perform equally well. Considering a GDFM with static representation the performance of our procedure is as good as the obtained with ABC and in some cases our procedure is even better. Similar results are reported in Forni et al. (2017) and Forni et al. (2018) regarding the forecasting of returns. Besides, when the DGP is a GDFM with no static representation (infinite-dimensional factor space), the performance of our proposal is much better than the procedure of Alessi et al. (2009). In the empirical application, our proposal reported the best out-of-sample performance according to SD, IR and SR criteria while ABC reported the best performance according to the AV criterion. In conclusion, we can say that our proposal is a sound alternative to forecast the one-step-ahead conditional covariance matrix in high-dimensional data.

References

1. Alessi, L., Barigozzi, M., and Capasso, M. (2009). Estimation and forecasting in large datasets with conditionally heteroskedastic dynamic common factors. Working paper series 1115, European Central Bank, Frankfurt am Main, Germany.
2. Barigozzi, M. and Hallin, M. (2018). Generalized dynamic factor models and volatilities: Consistency, rates, and prediction intervals. *arXiv preprint:1811.10045.*

3. Forni, M., Giovannelli, A., Lippi, M., and Soccorsi, S. (2018). Dynamic factor model with infinite-dimensional factor space: forecasting. *Journal of Applied Econometrics*, 33(5):625{642.
4. Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82(4):540{554.
5. Forni, M., Hallin, M., Lippi, M., and Zaaroni, P. (2015). Dynamic factor models with infinite-dimensional factor spaces: One-sided representations. *Journal of Econometrics*, 185(2):359{371.
6. Forni, M., Hallin, M., Lippi, M., and Zaaroni, P. (2017). Dynamic factor models with infinite-dimensional factor space: asymptotic analysis. *Journal of Econometrics*, 199(1):74{92.
7. Hallin, M. and Lippi, M. (2013). Factor models in high-dimensional time series|a timedomain approach. *Stochastic Processes and Their Applications*, 123(7):2678{2695.
8. Trucios, C., Mazzeu, J., Zevallos, M., Hotta, L., Valls Pereira, P.,, and Hallin., M. (2019). Forecasting conditional covariance matrices in high-dimensional data: a general dynamic factor approach. *Working Paper, University of Campinas*.

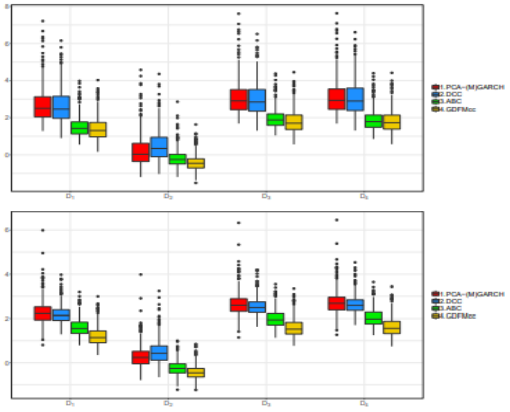


Figure 1: Box plot of the logarithm of the distances D_1, D_2, D_3 and D_4 for DGP3 (top) and DGP4 (bottom) across 500 Monte Carlo replications. PCA-(M)GARCH, DCC, ABC and GDFMcc stand for the DCC model applied on the first four principal components and univariate GARCH models on the idiosyncratic components, the DCC with composite likelihood, the procedure of Alessi et al. (2009) and our proposal, respectively.

Spatial multivariate outlier detection in the water quality of Klang River basin, Malaysia

Nur Fatihah Mohd Ali, Rossita Mohamad Yunus, Ibrahim Mohamed
Institute of Mathematical Sciences, Faculty of Science, University of Malaya, Malaysia

Abstract

A classical outlier detection method ignores the influence of neighbourhood effects and it may cause a misclassification of atypical observation. Thus, a spatial outlier or local outlier is referred as the observation whose non-spatial attribute value is significantly different from those of its neighbour. The detection of spatial outliers seems relatively difficult to carry out for multivariate non-spatial attributes. Thus, this study intends to identify the global and local outliers of the water quality data across 16 selected sites in the Klang River basin, Malaysia in 2016. The dataset consists of seven nonspatial attributes which are dissolve oxygen, biochemical oxygen demand, chemical oxygen demand, suspended solids, ammoniacal nitrogen, temperature and pH. The statistical method used to detect these outliers is based on the Mahalanobis Distance proposed by Filzmoser et al. (2014). The data were first classified as either regular observations or outlying observations. 25% of the water quality sites have been identified as global outliers. Site 11 has been identified as local outlier while Site 14 as global and local outlier. All statistical data analysis was analyzed using R package 'mvoutlier'.

Keywords

Global outlier; local outlier; multivariate outliers; spatial; Mahalanobis Distance

1. Introduction

The occurrence of outlying observations in spatial data may trigger the discovery of unexpected, interesting and implicit knowledge. A spatial or local outlier is a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood (Shekhar et al., 2003). In literature, an increasing number of studies have found that the local outlier in spatial datasets is more meaningful because detecting local outliers need to be work on locally, on neighborhoods to consider the spatial dependence between the spatial objects as "Everything is related to everything else, but nearby things are more related than distance things". The problem of handling the occurrence of spatial outliers has been distinctly addressed for univariate cases (Shekhar, 2003; Lu et al., 2003; Kou, 2006). However, the problem becomes more

complicated when spatial outliers occur in multivariate data. A number of detection methods have been fully described in Schubert et al. (2014).

Recently, the detection techniques applicable for multivariate non-spatial attributes have also been discussed in Ernst & Haesbroeck, (2015). Recently, attention has been focused on managing the influence of spatial outlying observations in environmental data analysis (Bobbia, 2015; Xin, et al., 2015; Harris et al., 2014). In the study of water quality in Malaysia, the impact of rapid industrialization, infrastructures and urban-expansions on biological, chemical, ecological and physical parameters (Chowdhury et al., 2018; Mohamed et al., 2015; Othman et al., 2012) may lead to spatial outlying measurements results. Some literature studies on environmental data analysis focus more on spatial outlier detection methods. Bobbia (2015) uses a jacknife type approach and is based on the comparison of the actual measurements with some robust prediction for PM10 concentrations in Normandy (France). Xin et al. (2015) has identified spatial outliers using a spatial local outlier factor (SLOF) algorithm approach on CO2 monitoring databased. In addition, Harris et al., (2014) has improved the multivariate spatial outlier detection using robust geographically weighted methods and it has been applied to freshwater chemistry data for Great Britain. For our purposes, the outlier detection method proposed by Filzmoser et al. (2014) is considered, which has been shown to be at least as advanced as other alternative methods for spatial outlier multivariate data. The original water quality data from the Klang River basin were generated for the year 2016.

In this study, the objective is to classify the global and local outliers of the water quality data across 16 selected sites in the Klang River basin, Malaysia. The sites with detected outliers are grouped as the sites with outlying observations and their profiles are then discussed further.

2. Methodology

The water quality data are collected from 16 sites in the Klang River basin, Malaysia for year 2016. Parameters measured at each sites include dissolve oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solids (TSS), ammoniacal nitrogen (NH3), temperature and pH. Site 1-7 are located along the main Klang River while Site 8 and Site 9 are located at Damansara River and Penchala River respectively. Site10-11 are located along Kerayong River and Site 12-14 are located along Gombak River.

The local outlier detection method in Filzmoser et al.. (2014) 1s based on pairwise robust Mahalanobis distances between the observations. The robust covariance estimation proposed by Rousseeuw and Driessen (1999): Minimum Covariance Determinant (Fast-MCD) is plugged in for computing the Mahalanobis distance. In addition. the detection method requires a definition of local neighbourhood. For this purpose. neighbourhood is defined by the

nearest k observations. For finding the k nearest neighbours (KNN) of an observation, z_i, the sorted distances to all other observations need to be considered. The local outlier is measured by the degree of isolation of an observation from a fraction of its neighbours and denoted as $\alpha(i)$-quantile such that:

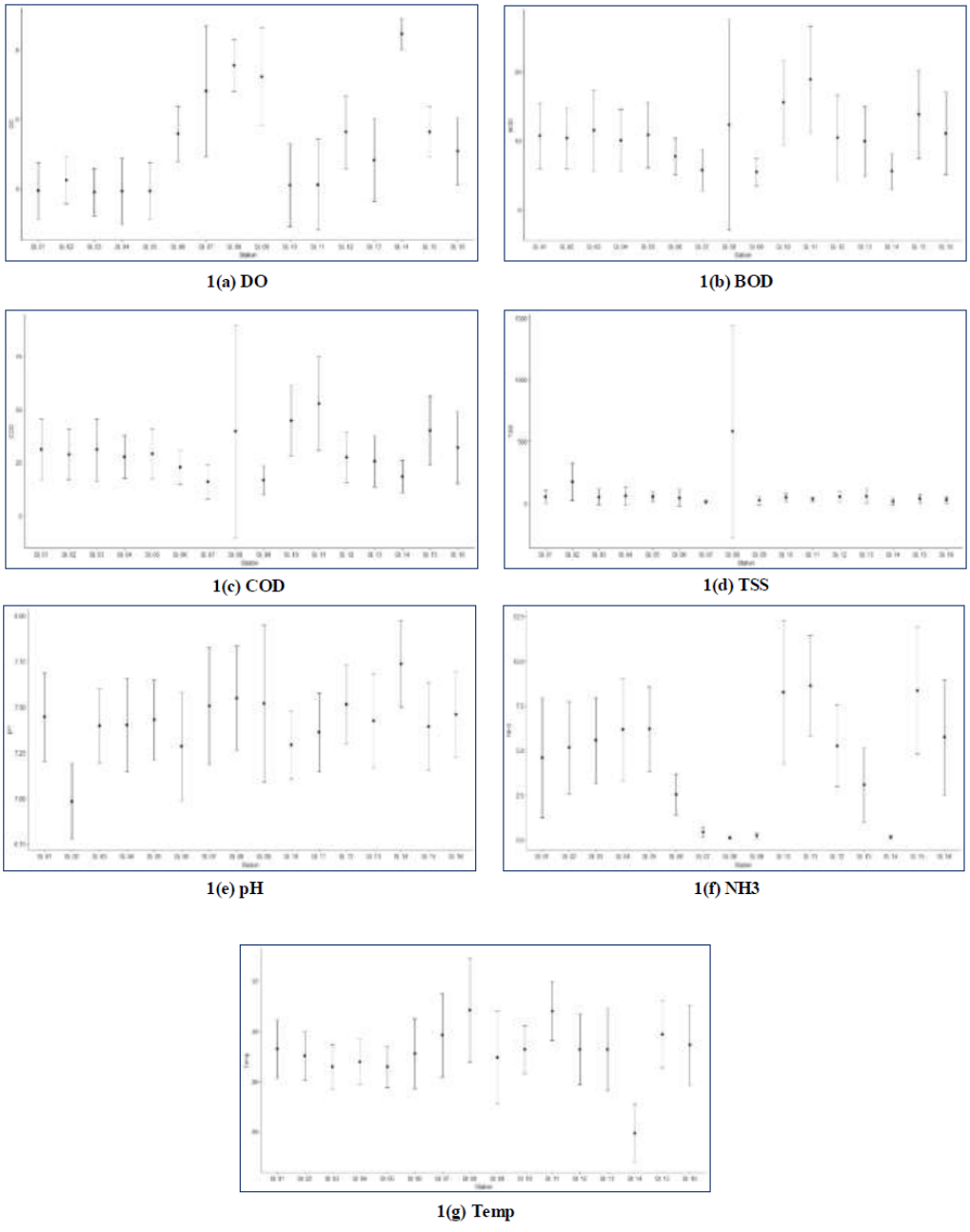
$$x^2_{p;\alpha(i)}(MD^2(z_i)) = MD^2(z_i, z_{([n(i).\beta])}) \quad for\ i = 1,\ldots,n. \tag{1}$$

The pairwise squared Mahalanobis Distance which is the right hand side in Equation (1) is a non-central chi-square distribution with p degree of freedom and the non-centrality parameter of the squared Mahalanobis distance on the left hand side in Equation (1). The neighbours of observation z_i are denoted as $z_{([n(i).\beta])}$ where $n(i)$ is the number of neighbours, k: while β denote a fraction of neighbours. All outlier detection tools proposed in Filzmoser et al.. (2014) are implemented in R package $mvoutlier$.

3. Result

In this section. the descriptive statistics of the river water quality is presented and the spatial outlier is then identified. The basic statistics of the year 2016 data set on the Klang River basin water quality were plotted in the error bar plot in Fig. 1a-g. Fig. 1a shows that the DO levels for Site 8 and Site 14 are among the highest with means are greater than $7\mathrm{mgL}^{-1}$. In contrast. Fig. 1b displays that the mean of BOD readings at Site 14 is among the lowest mean. However. Site 11 shows the highest mean of BOD value. From Fig. 1c and 1f. it can be observed that the COD and NH_3 readings for Site 11 are also the highest among all the other sites. The mean readings for TSS are generally similar for all sites except for Site 08 with the highest mean and large variation of values. In general the pH level in Fig. 1e for all rivers are in the range between 7 and 8 and mean temperature of the surface of river shown in Fig. 1g is about 29 °C with the lowest mean temperature of 26 °C is observed at Site 14.

Spatial outlier detection by Filzmoser et al. (2014) categorized the sites into two groups. First is regular observations and second is outlying observation. The regular observations are the observations which are inside the ellipse of Mahalanobis Distance while the outlying observations are the opposite of it. Fig. 2 and Fig. 3 show the results for the regular observations and global outlying observations respectively. The regular observations plot indicates the sites (iver) could be the local outliers but are not global outliers. Meanwhile. the outlying observations plot shows the sites (river) are the global outliers and they could also be the local outliers. Each line represents one specific site. Site 01-02. 8 and 13 are grouped in the outlying observation category and the other sites are in the regular observation. Both plots illustrated the degree of isolation from 90% of the neighbours versus the number of neighbours.



1(a) DO

1(b) BOD

1(c) COD

1(d) TSS

1(e) pH

1(f) NH3

1(g) Temp

Figure 1: Distributional plots for each water quality parameter

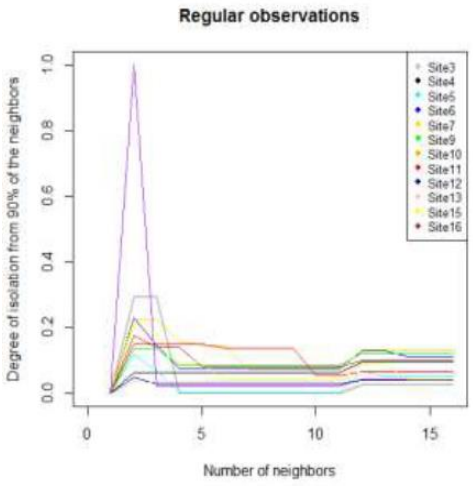


Figure 2: Regular observations

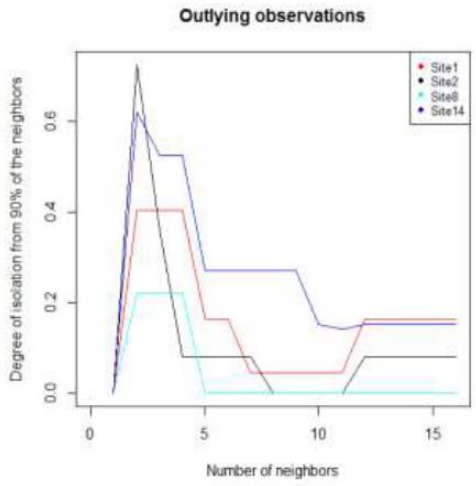


Figure 3: Outlying observations

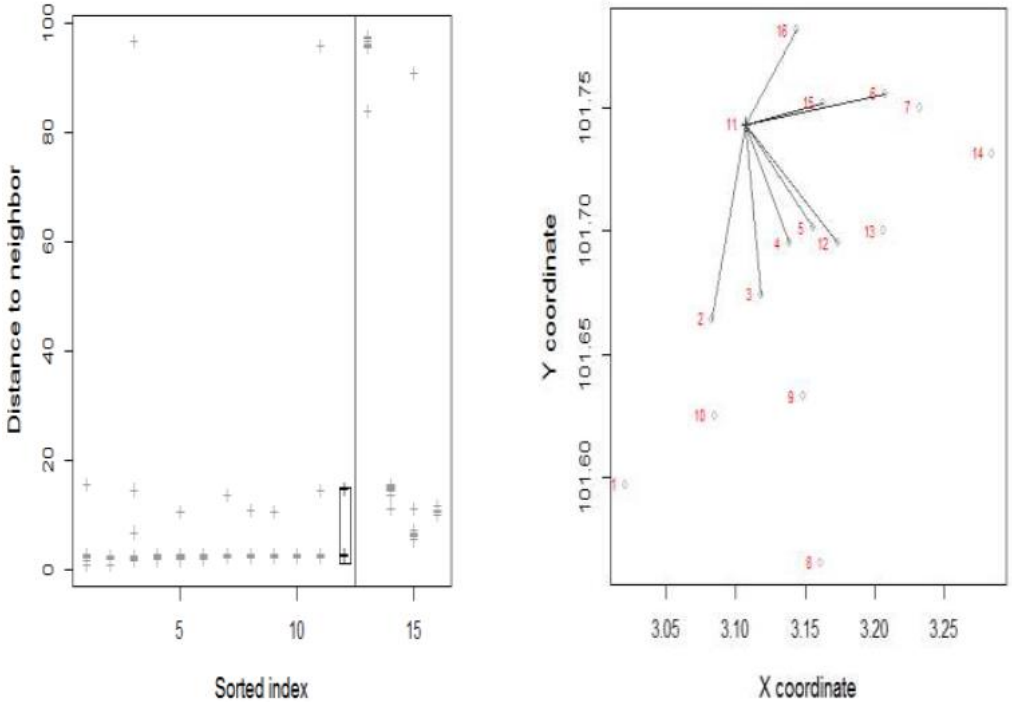


Figure 4: (Left) Plot of the observation index sorted by the degree of isolation versus the distance to all neighbors. (Right) Spatial coordinates of the sites. The marked points in the left plot are connected by lines in the right plot

The plots in Fig. 4 shows the sorted index of the sites from smallest to highest degree of isolation versus the pairwise Mahalanobis distance to neighbours. The plot windows are split into two parts; the left part for the regular observations and the right part for the outlying observations. In addition, the right plot in Fig. 4, displays the coordinate of the point in the polygon from the left plot. In this plot, we choose kNN equal to 8. The point

of intersection between the lines is labelled as local outlier in the regular observation and it is Site 11 which is located at Kerayong River.

Fig. 5 displays same plot as in Fig. 4. However, the polygon is sketched on the right part and it consists of sorted index of outlying observations. The point of intersection between the lines is labelled as Site 14 which is located at Gombak River.

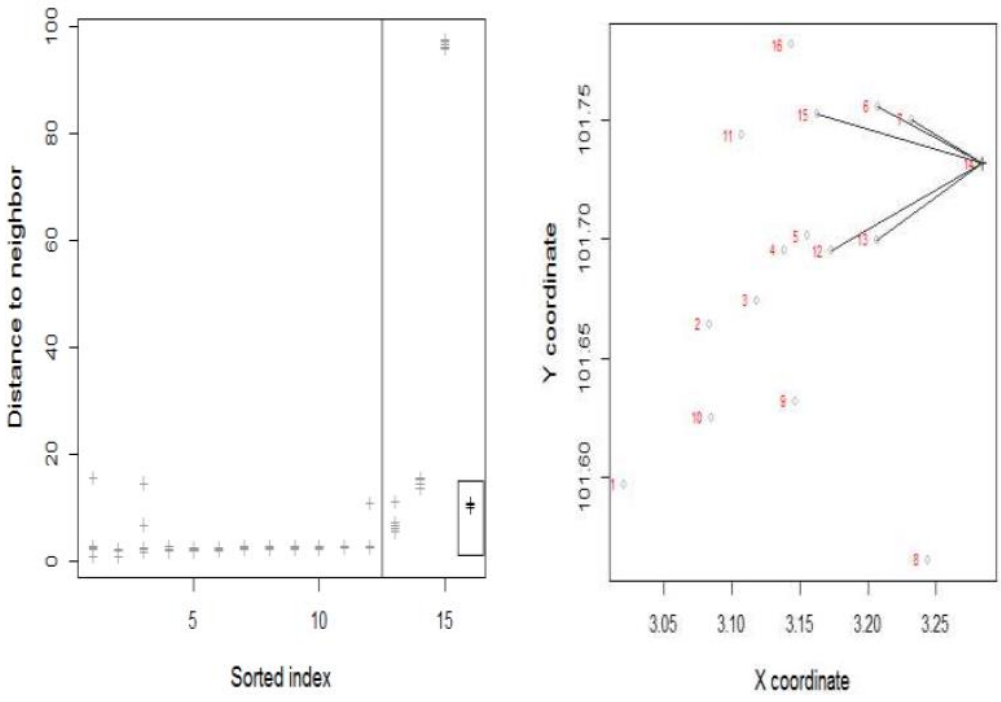


Figure 5: (Left) Plot of the observation index sorted by the degree of isolation versus the distance to all neighbors. (Right) Spatial coordinates of the sites. The marked points in the left plot are connected by lines in the right plot.

4. Discussion and Conclusion

From the analysis of local outlier identification given in result from Fig. 2, we consider the degree of isolation of β = 10 %, and vary the neighbourhood size with kNN. The left plot for the regular observations reveals one observation, Site 13 as clearly deviating for the first two next neighbours but then decreasing for a large range of next neighbours. However, the line of Site 07 in the plot shows high degree of isolation when the number of neighbours are increasing. This indicates that this site may be a local outlier. In the Fig. 3 we can see that four global outliers are identified. Similar behaviour is observed for the outlying observation plot where Site 14 line has the highest number of isolation and Site 08 is the second highest when number of neighbours is more than 5. Therefore, Site 14 is highlighted as global and local outlier and has been confirmed in Fig. 5. On the other hand,

from Fig. 4, Site 11 is chosen as a local outlier because the site is different compared to the 8 neighbours. Indeed, it has low DO levels but high in BOD, COD and NH_3levels. So, this Site 11 may have bad water quality.

However, finding the reason for local outlier of all marked points would require much more detailed study. Reasons for their abnormal behaviour could be a different data structure caused by local rapid industrialization, infrastructures and urban-expansions, or exchanged samples, incorrect sample preparation, wrong laboratory analyses, etc.

Acknowledgement: This work is supported by University Malaya, Research University Grant. No. RF015B-2018.

References

1. Ernst, Marie & Haesbroeck, Gentiane. (2016). Comparison of local outlier detection techniques in spatial multivariate data. *Data Mining and Knowledge Discovery*, 31(2).
2. Chowdhury, U, S., Othman, Jaafar, W., Mood & Adham. (2018). Assessment of Pollution and Improvement Measure of Water Quality Parameters using Scenarios Modeling for Sungai Selangor Basin. *Sains Malaysiana*, 47(3)(2018), p.457–469.
3. Filzmoser, Peter & Ruiz-Gazen, Anne & Thomas-Agnan, Christine. (2014). Identification of local multivariate outliers. *Statistical Papers*, 55(1), pp 29–47.
4. Harris, P., Brunsdon, C., Charlton, M. et al. Math Geosci (2014). Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods. 46(1).
5. Ibrahim Mohamed, Faridah Othman, Adriana I. N. Ibrahim, M. E. Alaa-Eldin & Rossita M. Yunus (2015). *Environ Monit Assess*,187(1), p.4182.
6. Kou, Y. (2006). Abnormal Pattern Recognition in Spatial Data.
7. Lu Chang-tien, Chen De-chang, Kou Yu-feng (2003). Algorithms for spatial outlier detection. *In: Proceedings of the Third IEEE International Conference on Data Mining (ICDM), Melbourne, Florida, US.*, pp. 597-600.
8. Michel, B., Michel, M., Yves, M., Jean-Michel, P., & Bruno, P. (2015). Spatial outlier detection in the PM 10 monitoring network of Normandy (France). *Atmospheric Pollution Research*, 6(3), 476-483.
9. Othman, Faridah & Mohamed Elamin, Alaa Eldin & Mohamed, Ibrahim. (2012). Trend Analysis of a Tropical Urban River Water Quality in Malaysia. *Journal of Environmental Monitoring*,14, p.3164-3173.
10. Rousseeuw PJ, Van Driessen K (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*. 41(3), pp.212–223.
11. Schubert, Erich & Zimek, Arthur & Kriegel, Hans-Peter. (2014). Local outlier detection reconsidered: A generalized view on locality with

applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 28(1)

12. Shekhar S, Lu Chang-tien, Pu-sheng (2003). A Unified Approach to Detecting Spatial Outliers. *Geo-Informatics,An International Journal on Advances of Computer Science for Geographic Information System*, 7(2), pp.139-166.
13. Xin, L., Shaoliang, Z., & Pulin, Z. (2015). Spatial Outlier Detection of CO2 Monitoring Data Based on Spatial Local Outlier Factor. *Journal of Engineering Science and Technology Review*, 8(5), 110-116.

Using Google trend data as an initial signal Indonesia unemployment rate

Rani Nooraeni[1], Nurmitra Sari Purba[2], Nugroho Puspito Yudho[3]
[1,2]STIS, Polytechnic of Statistic, Jakarta, Indonesia
[3]BPS, Statistic Indonesia, Jakarta

Abstract
Macroeconomic data are generally published with a time lag. Nowcasting uses the latest available data to provide estimates of macroeconomic variables in the short term. Surely this prediction applies until the actual estimation is generated. The ability to search data through the internet has provided a new data source for researchers who are interested in predicting macroeconomic variables in the short term. This paper trying to learn the behavior intensity of keyword from google search in order to predict unemployment growth in Indonesia. Some modeling is done to get the results of nowcasting which is quite ideal using ARIMA, BSTS without and with Google Data as an additional variable. We also conduct several treatments related to variations of keywords obtained. selection of important keywords with stepwise regression and keyword grouping using the principal component analysis method to obtain a composite keyword that can produce a better model for predicting unemployment growth rates.

Keywords
unemployment; Google trend; nowcasting; BSTS; composite

1. Introduction

The government needs various macroeconomic indicators as the basis of determination national policy strategies. They use Macroeconomic indicators that resulted by national institutions. Some indicators that are often used as a basis for policy include GDP, inflation, unemployment, exchange rates, poverty, and others. The overall indicators are obtained from the process data collection such as census/surveys. The indicator of the unemployment was conducted either by BPS Statistics of Indonesia, the Ministry of Manpower, or other agencies that conduct separate surveys regarding the labor market. Even though many parties have collected data related to employment, each of the data collection processes has limitations, especially the period of dissemination. For example, the unemployment rate produced by BPS can only be disseminated twice in a year. This certainly influences policymakers because the basis of the policy is only able based on available data which does not necessarily match with the condition at the time the policy will be made. Because often policies need to be made frequently in line with changing times.

On the other hand, the development of the era has made information technology more sophisticated, from time to time the percentage of households that have computers and cell phones has increased. In sequence, they growth from 19.88 percent to 88.04 percent while households that have mobile phones have grown from 3.65 percent to 18,71 percent, during the period of 2005-2015. In line with the rapid increase in ownership and mastery of these technologies. Development of internet users is also growing rapidly. At the same period time, in Indonesia, the percentage of internet users increased from 7.4 percent to 34.9 percent. With this phenomenon, the impact of that depelovment is that people become easier to get information and also easy to give information. This has an impact on the use of search engines on the internet, more and more people are looking for information about anything through search engines that can cross between countries.

An example of a phenomenal search engine is Google Search. Google Search is commonly used to find relevant information using various keywords. The intensity of search uses various kinds of keywords that produce data that is so large (Big Data) recorded by Google. Further, the data can show the latest socio-economic conditions. Then Google began to disseminate the intensity of the search in 2009 through the Google Trend interface. Of course, this is good news for researchers. They use this data to improve the accuracy of the latest predictions. Mitra, Sanyal, and Choudhury (2017), they were nowcasting Real Estate growth in India using Google Trend Data. The availability of digital data whose time references are more up-to-date, weekly, monthly or annual, this can be supporting data for data produced by official institutions through the census/survey data collection process.

Google provides three data sources that can be useful for social science (i.e. google trend, google correlate, and google Consumer Surveys (Stephens-Davidowitz & Varian, 2015), but they claim that besides social science there are many things that can be found from Google. Koop (2013) used google data to predict macroeconomic indicators, with Google Data we can make predictions about things in the most recent period, this is called Nowcasting, so we can use Google data to support indicators related to economic or social indicators of a country. In 2009, Google began to disseminate data from Google Search, including data from 2004 to the present, covering all countries in the world.

Based on the foregoing, this study will explore Google Data in describing the conditions of economic indicators in the current period. The focus of the macroeconomic indicators that we observe is the growth rate of unemployment in Indonesia. However, in the use of Google Search Data, caution is needed in determining keywords related to the problem to be studied, the behavior of Google Search data must also be identified so that the right method can be used to produce a good prediction model. One of

the projection methods that can be used for this kind of data, internet data, is the Bayesian Structural Time Series. This paper utilizes data from the internet, Google trend (GT) data, to know the recent conditions of macroeconomic indicators in Indonesia. Then combinate the GT data with another macroeconomic variable to improve short-term predictive/nowcasting capabilities model.

Google Trend Data

When we are going to use GT data, there are some things that need to be underlined. First, data of GT has the potential to predict short-term but not for the long term. Second, it is rarely used for broadening macroeconomic variables such as inflation, industrial production, etc. It is more useful for predicting certain variables related to consumption, housing or the labor market. For example, Choi and Varian (2011) succeeded in predicting motor vehicle parts and cars, initial claims for unemployment benefits and tourist arrivals in Hong Kong

In line with the development of research that utilizes GT data, we found some literature/references that conducted research using GT Data to predict macroeconomic indicators. The Google variable can be used as additional information and check whether GT involvement can improve nowcasting capabilities.

2. Methodology

2.1 Data used

The data used in this study consisted of two sources, BPS-Statistics Indonesia and Google. Macroeconomic variables, namely unemployment rate, Consumen Price Index (CPI) and inflation are obtained from BPS-Statistics Indonesia. Google trend data collected based on several keywords that are relevant to the unemployment rate of Indonesia. Unemployment rate official data of Indonesia result from survey national of labour (SAKERNAS). BPS conduct this survey twice a year, February and August. A time reference that we use in build model prediction was from February 2005 to August 2018.

2.2 Selection of Keywords

The selection of keywords has a crucial role in producing precision prediction results. It takes several stages to finally produce keywords that are closely related to the subject we tell. In the use of keywords in the early stages can refer to keywords that have been used by previous researchers who have the same concern. In addition, we can find the relevant keywords from google correlate. However, not all of the keywords produced will be relevant to the subject we are observing. Thus, we have to combine all references, all procedures, and statistic method to get the most representative keyword.

Finally, based on relevancy, the following keywords are suggested to our study (in Bahasa):

bisnis	contoh cv	karir	lowongan kerja
bisnis indonesia	contoh lamaran pekerjaan	kerja	lowongan kerja di
bisnis online	employment	kerja di	lowongan kerja teknik
bursa kerja	job	kerja part time	lowongan pekerjaan
bursa lowongan	job application	lamaran kerja	lowongan teknik
career	jobsdb	lamaran pekerjaan	pekerjaan
cari duit	jobsdb.com	loker	peluang bisnis
cari kerja	jobstreet	lowongan	surat lamaran
cari uang	job vacancy	lowongan di	

The prediction of condition Unemployment rate in Indonesia will execute with only several important keywords. The selection will apply stepwise regression method to determine what keywords are most relevant to the research subject and Principal Component Analyse to combine number of variables.

2.3 ARIMA Model

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average. It has very good accuracy for short-term forecasting. Standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used. Parameter p is the number of lag observations included in the model, also called the lag order. Parameter d is the number of times that the raw observations are differenced, also called the degree of differencing. Parameter q is the size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model. Differencing, autoregressive, and moving average components make up a non-seasonal ARIMA. ARIMA (p,0,q) model can be written as a linear equation:

$$Y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + e_t + \cdots + \theta_q e_{t-k}$$

where φ_1, φ_2 are parameters for Auto Regressive and θ_q is parameter of Moving Average. When we include either explanatory variable in to ARIMA model it is named ARIMAX model.

2.4 Bayesian Structural Time Series (BSTS) Model

One of the advantages of Bayesian modeling is to account for the uncertainty associated with parameter estimates and provide exact measures of uncertainty on the posterior distributions of these parameters, which is traditionally ignored in classical estimation models. In addition, Bayesian

estimation and inference provide confidence intervals on parameters and probability values on hypotheses that are more in line with commonsense interpretations (Congdon, 2003).

BSTS combines the two approaches of Bayesian inference and structural time series. The second component, time series in a state space form, can be described with an observation and a transition equation given by

$$\boldsymbol{y_t} = Z_t^T \alpha_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{1}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \qquad \eta_t \sim N(0, Q_t) \tag{2}$$

With random Gaussian noise ε_t and η_t with variance σ_ε^2 2 and Q_t respectively. Equation (1) provides information regarding the relation between the observed variables $\boldsymbol{y_t}$ and the latent state variables α_t . Equation (2) describes the transition behavior of the latent state variables over time. The model matrix Z_t , T_t, R_t typically contain a mix of known values (often 0 and 1), and unknown parameters. Notation in the form of a state space model is convenient because one can effortlessly add further components to the state vector, such as seasonality and trend. A Model can be obtained by adding a regression component to the popular "basic structural model." This model can be written

$$y_t = \mu_t + \tau_t \boldsymbol{\beta}^T \boldsymbol{X}_t + \varepsilon_t \tag{3}$$

Where y_t , μt, τ_t , ω_t , ξ_t and ε_t representing target time series, local linier trend component, seasonal component, regression component and observation error terms respectively.

$$\mu_t = \mu_{t-1} + \delta_{t-1} + \mu_t \qquad u_t \sim N(0, \sigma_u^2)$$

$$\tau_t = \textstyle\sum_{s=1}^{s-1} \tau_{t-s} + w_t \qquad w_t \sim N(0, \sigma_\tau^2)$$

$$\delta_t = \delta_{t-1} + v_t \qquad v_t \sim N(0, \sigma_v^2)$$

S represent the number of season for y and τ_t denotes their join contribution to the observed target time series y_t . As is typical in Bayesian data analysis, forecasts from our model are based on the posterior predictive distribution. It is trivial to simulate from the posterior predictive distribution given draws of model parameter and state from their posterior distribution. Let ỹ denote the set of values to be forecast. The posterior predictive distribution of ỹ is

$$p(\tilde{y}|y) = \int p(\tilde{y}|\phi) p(\phi|y) d\phi$$

3. Result

The unemployment rate in Indonesia, based on SAKERNAS data, continued to decline from 10.26 in February 2005 to 5.3 in August 2018. This make conclusion that open employment opportunities has been increased. But in particularly point of period, the unemployment rate does not always decrease,

at certain times the unemployment rate higher than the previous time. For example, the unemployment rate on February 2018 higher than rate on August 2018. It was increased from 5.1 to 5.3. This happened because of the influence of seasonal factors and other influences such as depreciation of the rupiah

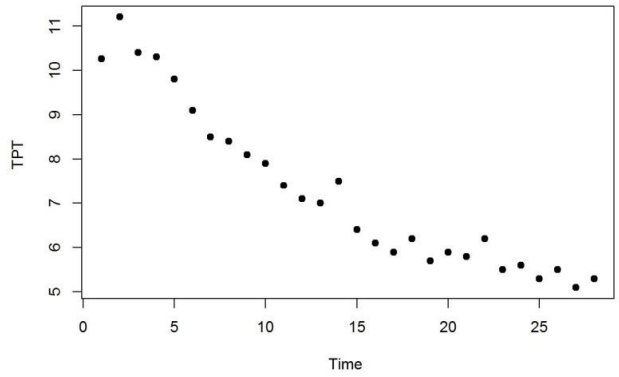


Figure 1. Unemployment Rate in Indonesia for the period 2005-2018

To predict the pattern of unemployment rates for the next recent period, GT data will be used based on 35 keywords that are relevant to Indonesia's unemployment rate. Based on 35 keywords, the most important keywords were selected using the stepwise regression method. Finally, four words are obtained, namely "bisnis indonesia", "loker", "bursa kerja", and "job". The macroeconomic variables that used in predicting unemployment rate are CPI and inflation rate. The interesting one that we can see from figure 2 is the correlation between Unemployment with "loker". They have high negative correlation. The more people who search for this keyword then the unemployment rate will decrease.

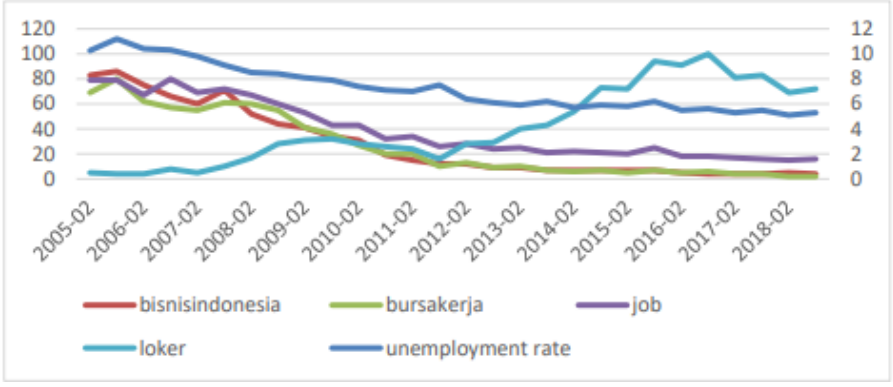


Figure 2. Inter-sample variation in keywords intensity

Based on plotting cumulative absolute error (fig 3), The performance of the prediction model with the ARIMA and Bayesian Time Series (BSTS) methods shows that the BSTS is able to produce better projection models than ARIMA. So the projection of the unemployment rate is continued with the BSTS method in several modeling scenarios. The first model, predicting

unemployment rate using official data only. the second model, adding the 5 most important keywords into the first model. The other model, adding the other macroeconomic variables such as CPI or inflation into model 1 and model 2. The complete scenario model can be seen in figure 3.

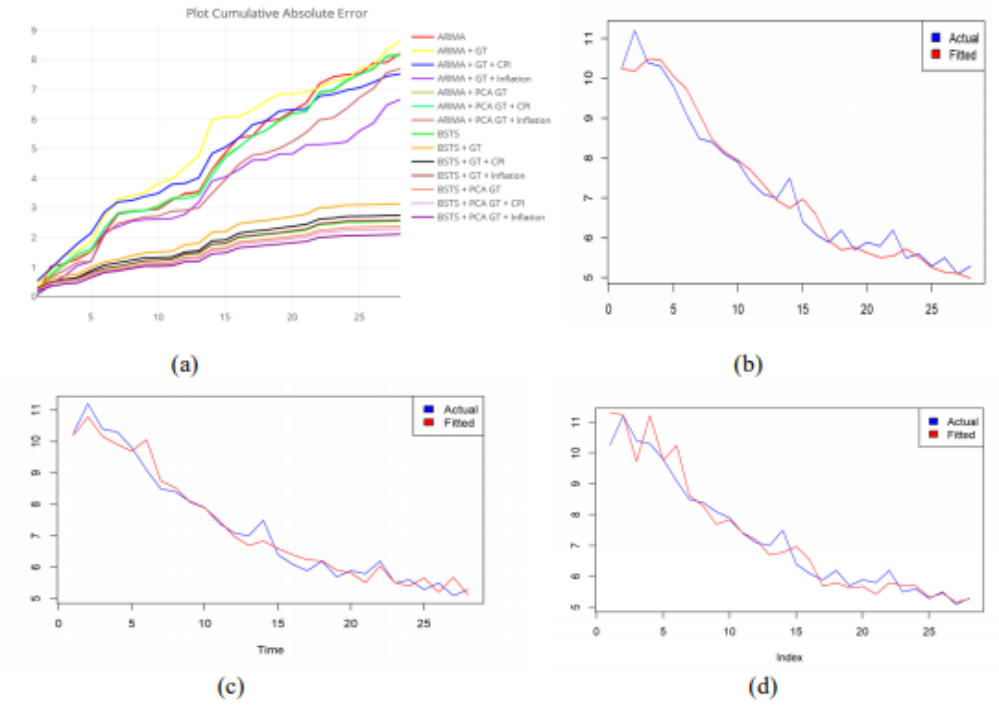


Figure 3. Plot (a) Cumulative Absolute Error for each scenario model, (b) ARIMA, (c) ARIMA+GT+Inflasi, (d)BSTS+PCA GT+Inflasi

The results of the unemployment projection using the ARIMA, ARIMAX and BSTS methods can be seen in figure 4. Projections with the BSTS method are more volatile and more relevant to official data than the ARIMA/ARIMAX method (figure 3). Based on the model BSTS and ARIMA, conditions of Unemployment Rate in February 2019 will be dcreased than August 2018.

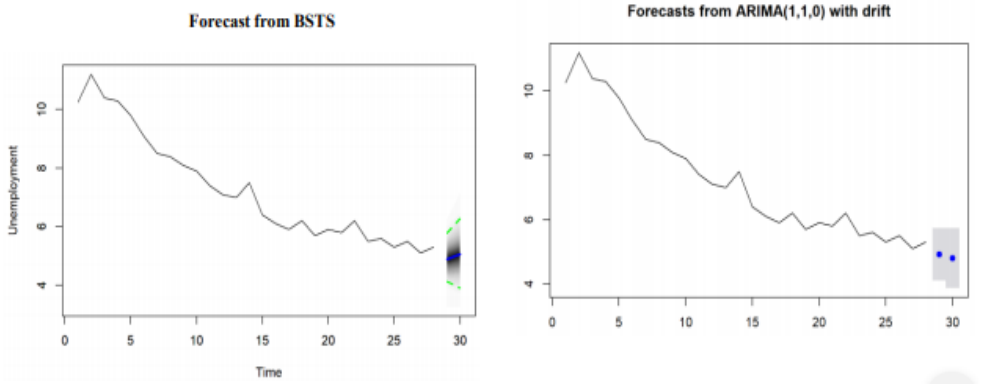


Figure 4. The Forecast plot with BSTS and ARIMA method

4. Discussion and Conclusion

The unemployment rate in Indonesia is predicted to fluctuate, this is due to seasonal influences. With Indonesia's background as an agrarian country, the dominance of sectors that absorb the most labor force in agricultural sector so that the harvest or non-harvest periods greatly affect the amount of unemployment. In addition, the stability of the rupiah value also affected fluctuations of the unemployment rate. Thus government efforts are needed to further open and enlarge employment opportunities for the Indonesian workforce as they reduce the effects of seasonal factors.

Based on the absolute error value, nowcasting the unemployment rate is better produced by BSTS than ARIMA method. the projection of BSTS can show seasonal effects such as the actual phenomenon of unemployment condition in Indonesia. While ARIMA was projecting the unemployment rate will decrease in time by the time. The use of Google Trend data as an additional information in projecting unemployment can improve nowcasting performance. Likewise, when combined with other macroeconomic variables, the involvement of Google trend data can improve the performance of the unemployment nowcasting model. Thus the role of Google trend data can increase the performance of the projection macroeconomic indicators.

References

1. Irfan, A. S. (2017). Tracking Labour Market Condition using Google Search Data-Indian Case Study.
2. Koop, G. (2013). Macroeconomic Nowcasting Using Google Probabilities.
3. Mitra, P., Sanyal, A., & Choudhury, S. (2017). Nowcasting Real Estate sector growth using Google Trend Data: An empirical retrospect for India. *Reserve Bank of India*, 1-32.
4. Stephens-Davidowitz, S., & Varian, H. (2015). A Hands-on Guide to Google Data. 1-25.
5. Congdon, Peter (2007). Introduction: The Bayesian Method, its Benefits and Implementation

Comparing rainfall curves between climatological regions using Functional Analysis of Variance

Jamaludin Suhaila[1,2], Muhammad Fauzee Hamdan[2]
[1]UTM Centre of Applied and Industrial Mathematics
[2]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, Johor, Malaysia

Abstract
This study is concerned on building up a functional data object from rainfall observations which could be used to represent the continuous rainfall process at a station. The typical functional data analysis approach is to fit each curve individually using expansion in basis functions and the choice of a basis implies the type of features of the data series. This study is interested in looking on how rainfall fluctuates at each station over time. Smoothing technique such as Fourier series is used to capture these variations. The main interest in this study is describing on how the pattern of daily variation of rainfall can be explained by its geographical locations via functional analysis of variance. Results indicated that different numbers of basis functions are needed to describe the rainfall curves for different climatological regions. Using the concept of functional analysis of variance, the test results indicated that there exist significance differences in the functional means of rainfall between the climatological regions. These differences may probably due to the effect of topography and geographical location and monsoons influence. The outcome of this study is expected to be useful to policy makers, hydrologist, and water resource planners dealing with climate change for the sustainable development and planning of water resources.

Keywords
Functional data; Functional Analysis of Variance; Fourier basis function; Rainfall curves; Smoothing technique.

1. Introduction

Statisticians have discovered functional data, which they observe data which can naturally be considered as an observation of a function rather than univariate or multivariate data. Functional data analysis (FDA) is the analysis of curves or functions where no parametric assumptions are needed to be made about the function or data. The basic idea of FDA is to express discrete observations arising from time series into a functional data that represents the entire measured function as a single observation, and to draw modeling and make inference based on the collection of a functional data by applying statistical concept from multivariate data analysis. FDA could give information

on the pattern and variation of the data and make use of the information in the slopes and curvatures of curves that are reflected in their derivatives that is not normally available from application of traditional statistical methods (Suhaila et al. 2011). In addition, there are also no concerns about correlation between repeated measurements since FDA treat the whole curve as a single entity. In standard statistical modelling, analysis of variance (ANOVA) is normally used to compare the population means or treatments. In the context of FDA, functional ANOVA method is appropriate when the data consist of functions that are expected to differ according to some sets of categorical factors. Since we are interested in looking on how the pattern of daily variation of rainfall can be explained by its geographical regions, functional analysis of variance is one of the ways to quantify these effects.

Our main objective in this paper is to employ functional ANOVA in analysing the rainfall curves between studied regions in Peninsular Malaysia based on the smoothing curves constructed from FDA. Additional information such as the number of basis functions required in describing the rainfall patters of the regions and the seasonal rainfall peaks could be determined. The outcome of this study is expected to be useful to policy makers, climatologist, and water resource planners dealing with climate change for the sustainable development and planning of water resources.

2. Functional data smoothing and analysis

Rainfall is one of the climate variables that often display fluctuations and changes over time. The goal of FDA is to transform discrete data, $y_i, i = 1,2,\ldots,T$ at discrete time interval into a continuous rainfall process at each rainfall station with a smooth curve, $f(t_i)$. The first step in FDA is to create a set of basis functions that best represents the functional data. A linear combination of basis function is used for representing the functions, given as

$$f(t) = \sum_{k=1}^{k} d_k \emptyset_k(t) \tag{1}$$

where d_k refers to the basis coefficient, $\emptyset_k$ is the known basis function while K is the size of the maximum basis required. The best known basis function for periodic data is given by the Fourier series. Smoothing through regression analysis will give the estimated value of the basis coefficient.

The functional ANOVA model can be written within the framework of general linear model, $y = \boldsymbol{X}\beta + \varepsilon$ with y is a vector of observation, β is parameter vector while matrix $\boldsymbol{X}$ incorporates observed covariates or independent variables. Since we are interested in looking on how the pattern of daily variation of rainfall can be explained by its geographical regions, analysis of variance is one of the ways to quantify these effects. Basically,

classical analysis of variance is used to determine whether two or more sets of data are identical, independent and coming from the same population. The same idea goes for functional data in which the verification is done by comparing their functional means. Functional analysis of variance models are appropriate when the data consist of functions that are expected to differ according to some set of categorical factors (Ramsay and Silverman 2005). In this paper we consider response models where the responses are functions indexed by groups, with the goal to learn if the functions differ across groups and, if so, how they differ. Suppose that we have a number of stations in each group g, and the model for the mth rainfall function in the gth group, indicated by $Rain_{mg}$, is given as

$$Rain_{mg}(t) = \mu(t) + \alpha_g(t) + \varepsilon_{mg}(t) \quad (2)$$

The function μ is the grand mean function which indicates the average rainfall profile across the studied stations of Peninsular Malaysia while α_g are the specific effects on rainfall of being in group g that satisfy the constraint

$$\sum_{g=1}^{4} \alpha_g(t) = 0 \; for \; all \; t \quad (3)$$

The residual function ε_{mg} is the unexplained variation specific to the mth rainfall station within climate group g. Similarly, Eq. (2) can be rewritten as

$$Rain_{mg}(t) = \sum_{j=1}^{5} x_{(mg)j} \beta_j(t) + \varepsilon_{mg}(t) \quad (4)$$

where $Rain_{mg}(t)$ t is the functional response with x_{mg} represent either 0 or 1 denoting the weather stations in a group and β_j, j=1,2,3,4,5 are the regression coefficient.

In this study, twenty rainfall stations across Peninsular Malaysia are divided into four zones; Northwest, West, East and Southwest. The functional response is the mean daily rainfall for 32 years. Suppose that we define 20 x 5 design matrix **X**, with one row for each individual station. The first column has all entries equal to 1, representing Peninsular mean rainfall, while the remaining four columns contain 1 if the weather station is in the corresponding climate region and zero otherwise. A corresponding set of five regression functions β_j, gives $\beta_1 = \mu, \beta_2 = \alpha_1, \beta_3 = \alpha_2, \beta_4 = \alpha_3, \beta_5 = \alpha_4$. Minimizing the residuals through the method of least square subject to the constraint given in Eq. (3) gives the least squares estimate $\hat{\beta}$ of the functional parameters μ and α_g.

The final step in Functional ANOVA is to determine whether there exist any significance differences between the groups based on their functional means. The permutation approach as mentioned in Ramsay et al. (2009) is employed to signify the difference. Basically, the procedure involved rearranging the vector of responses while keeping the covariates in the same order and tries to fit the model again. Functional version of the univariate statistic is given as follows: (5)

$$Ftest(t) = \frac{Var(\hat{y}(t))}{\frac{1}{n}\sum(y_i(t) - \hat{y}(t))^2}$$

where $\hat{y}$ is the vector of predicted responses. The procedure is repeated for hundreds time, for instance B=1000, and a new F-test ($F_{permute}$) is computed each time. The p-value is defined as

$$p - \text{value} = \frac{no.of\ F_{permute} \geq F_{observe}}{B + 1} \quad (6)$$

The test is found to be statistically significant if the p-value is less than or equal to the α significance level

3. Result and Discussion

Daily rainfall data for twenty rainfall stations over the duration of 32 years (1980–2011) were obtained from the Malaysian Meteorological Services. Figure 1 displays a physical map indicating the locations of the selected stations that will be used in the analysis.

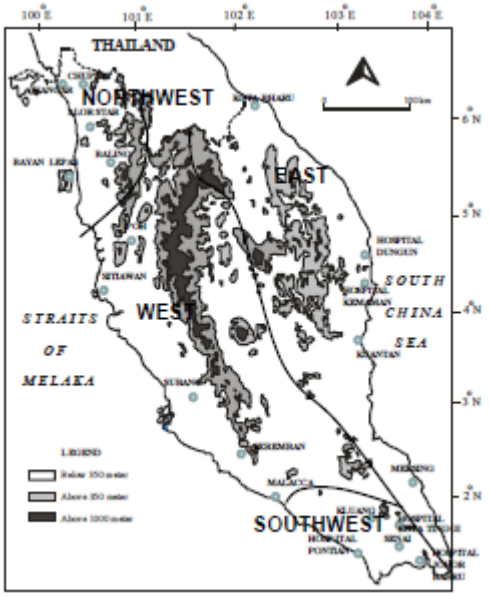


Figure 1. Map of the geographical location of the stations

To avoid having a large fraction of days with no rain, rainfall amount is pooled over 5 days periods and the mean rainfall amount per 5 days is computed over the studied period. In this study, the number of days is defined as T=73 days. Fourier basis is used to identify a suitable number of basis functions that is required to explain the behaviour and characteristics of the rainfall and pattern of each region.

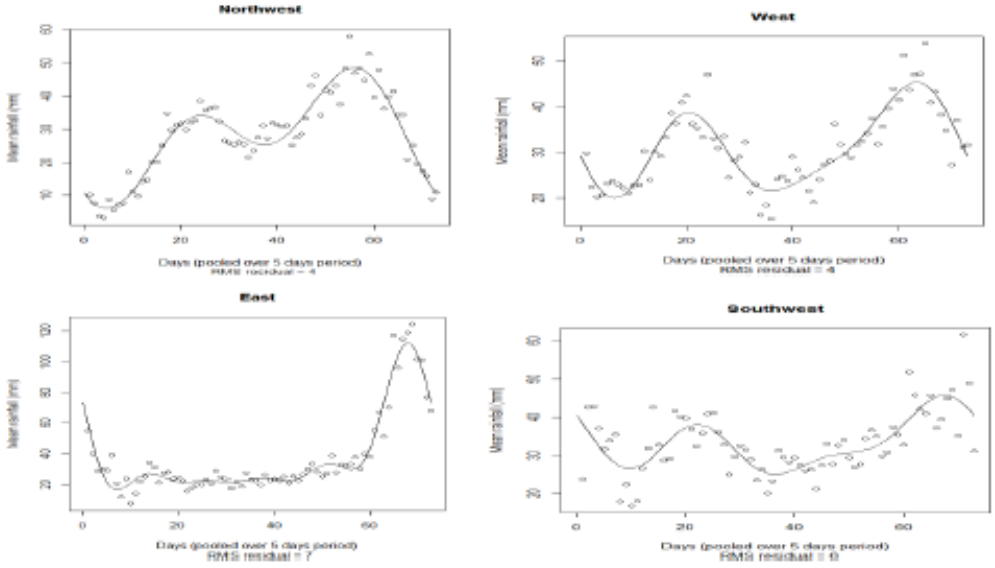


Figure 2. Observed daily mean rainfall with their corresponding smoothing curves.

Figure 2 shows the discrete observed mean rainfall over 5 days in four regions and their resulting smooth curves based on the number of basis functions that have been obtained. Bimodal rainfall patterns with two rainfall peaks are observed for the Northwest, West and Southwest regions. The rainfall peaks for the Northwest region are found during the inter-monsoons with the first peak occurs in April and May while the second peak is detected during September and October. Similar bimodal patterns are observed for West region, however the second peak is found at the early of November meanwhile the wet period for the Southwest region is found during the second rainfall peak which occurs between end of November and December. As shown in Figure 2, the second peak is larger than the first peak. A different rainfall pattern can be seen for stations at the Eastern region as shown in Figure 2. The rainfall values ranged from 20 to 120 mm and the daily mean rainfall changes rapidly throughout the year. A unimodal rainfall pattern with the highest peak is observed between November and December. Large numbers of basis functions between nine and eleven basis are required in describing the rainfall pattern for stations at the Eastern region which implies that four and five harmonics are needed to capture the rainfall variability for

these stations. In contrast to other stations located at the Northwest, West and Southwest, only five and seven basis functions are required in describing the rainfall pattern of the stations.

The concept of Functional ANOVA is then used to verify whether the differences in rainfall patterns between regions are statistically significant or not. We would like to investigate on how much of the pattern of daily variation of mean rainfall in a station is explainable by its geographical locations. Twenty rainfall stations are divided into four regions, Northwest, West, East and Southwest. The number of stations for each region is given by nNorthwest=10, nSouthwest =10, nWest =10, and nEast =10.

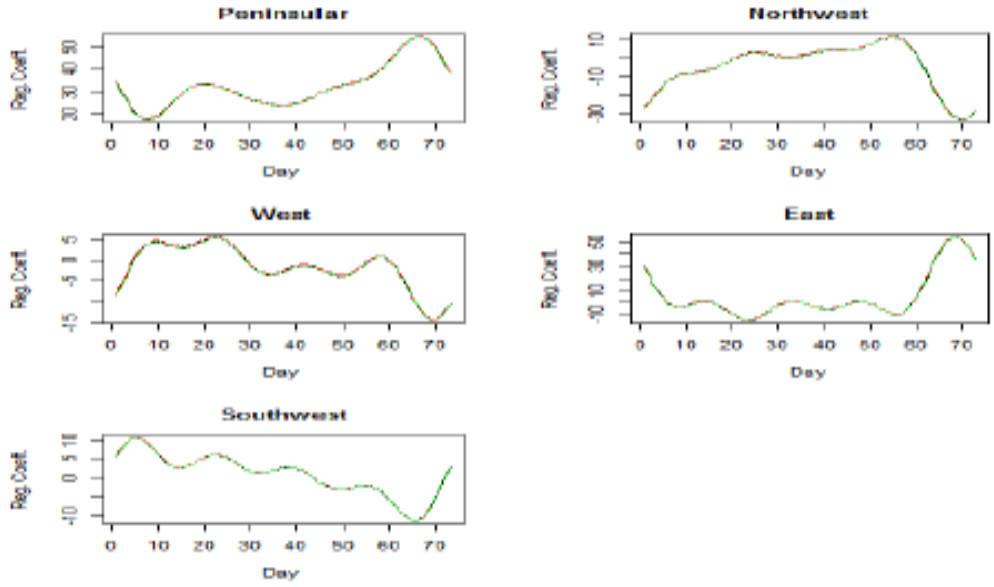


Figure 3. The solid line represents region effects α_g for the rainfall functions in the functional analysis of variance model

Figure 3 displays the resulting estimated region effects of the rainfall functions for the regions. The first panel is the intercept function, corresponding to the Peninsular mean rainfall. Peninsular Malaysia appears to have high mean rainfall at the end and early of the year. Based on the regression coefficient values, the East effect on rainfall functions are higher at the early and at the end of the year but negative effect during April to May and September to October. On the other hand, different patterns are observed for West and Northwest regions. Negative effects on rainfalls are observed at the early and at the end of the year while positive effects are seen during April to May and September to October for these two regions. In case for the Southwest region, positive effect is shown at the early and at the end of the year but turns to be negatively effect during other days of the year. In testing

the significance of the functional means between regions, the permutation tests of functional hypothesis are implemented. A test for no effect of geographical region on rainfall profile is conducted, and the results are shown in Figure 4.

The test results displayed in Figure 4(a) indicate that there exist significance differences between all regions with the probability value achieved as p-value = 0.00 < 0.05. Comparison between two regions also indicate significance difference exist in rainfall patterns particularly during the northeast monsoonal flow (Nov to Feb) as p-value achieved is less than 0.05. No indications of significant differences are observed between West and Southwest regions as shown in Figure 4(f) since the results are not statistically significant.

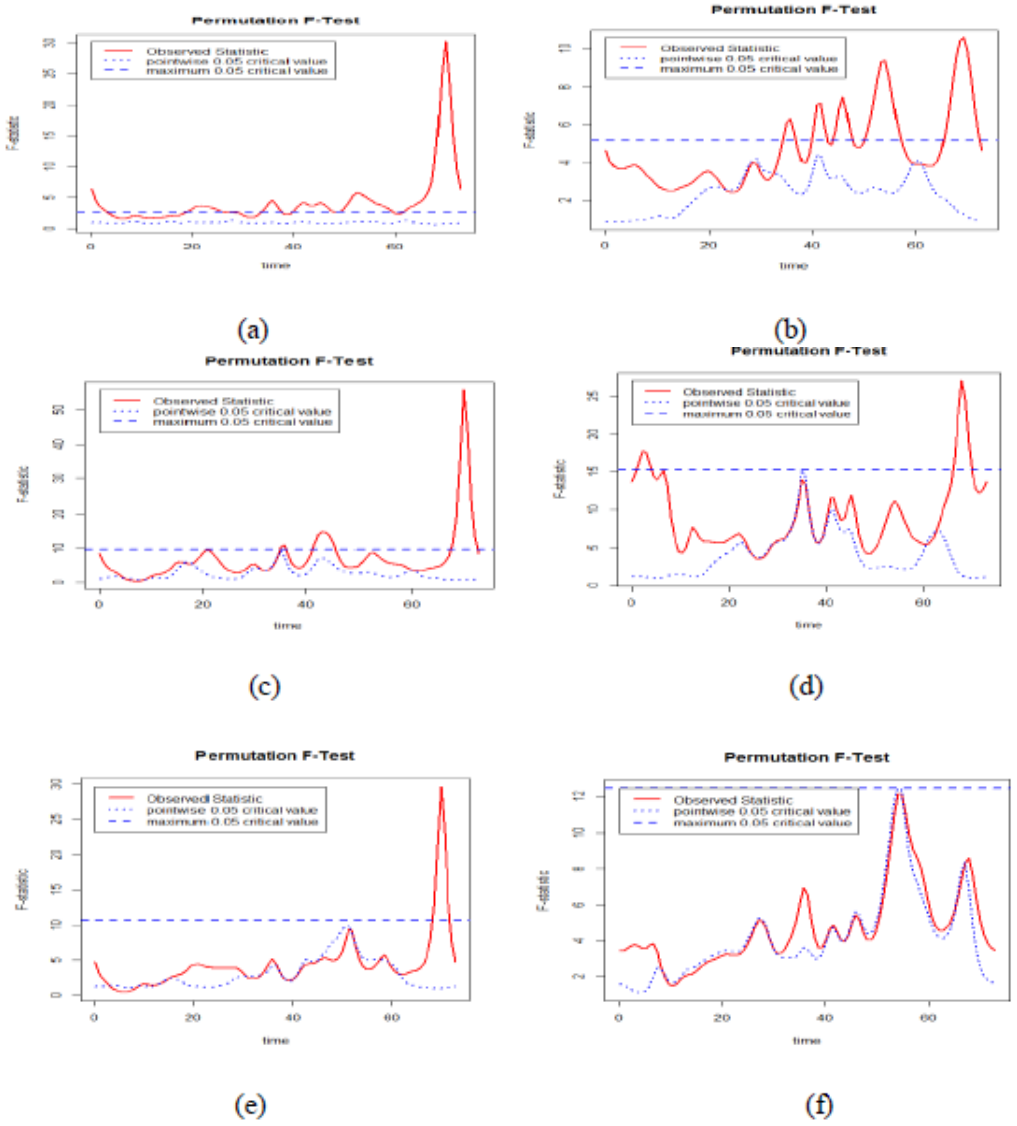


Figure 4. Permutation test for the difference in functional mean rainfall between (a) All four regions (*p*-value = **0.000**) (b) Northwest and West (*p*-value = **0.002**) (c) Northwest and East (*p*-value = **0.002**) (d) Northwest and Southwest (*p*-value = **0.002**) (e) West and East (*p*-value = **0.000**) (f) West and Southwest (*p*-value = 0.074). The dashed line gives the permutation 0.05 critical value for the maximum of *F*-statistic and the dotted the permutation critical value for the point-wise statistic.

4. Conclusion

This study has presented a technique based on functional data analysis which can represent rainfall data in the form of smooth curves that give meaningful information. Large numbers of basis functions are required in describing the rainfall patterns for rainfall stations at the Eastern regions compared to other regions. Using the mean rainfall as the functional response, a functional analysis of variance is employed to analyse the effect of geographical location on the rainfall functions. The findings can be derived such that there are significant effects of the geographical locations on rainfall profiles. The largest effect of East region on the rainfall patterns during the months of the northeast monsoonal flow while small effect was observed for the rest of the months. The main contribution of this study is focused on how the FDA technique could be used for rainfall analysis. Instead of utilizing discrete data, a smoothing curve can be created for the data which could be analysed over any time interval.

References

1. Suhaila, J., Jemain, A.A, Hamdan, M.F., and W.Z.W. Zin,. 2011. Comparing rainfall patterns between regions in Peninsular Malaysia via fumctional data analysis technique. *Journal of Hydrology*. 411, 197-206.
2. Ramsay, J.O., Silverman,B. Functional data analysis, Springer Series in Statistics, Springer, New York, 2005.
3. Ramsay, J.O., Hooker, G., Graves, S. Functional data analysis with R and MATLAB, Springer, New York, 2009.

mpcmp: Mean-Parametrized Conway-Maxwell Poisson (COM-Poisson) Regression

Thomas Fung[1], Aya Alwan[1], Justin Wishart[2], Alan Huang[2]
[1]Department of Mathematics and Statistics, Macquarie University
[2]School of Mathematics and Physics, University of Queensland

Abstract

Conway-Maxwell-Poisson (CMP) distributions are flexible generalizations of the Poisson distribution for modelling overdispersed or underdispersed counts. The main hindrance to their wider use in practice seems to be the inability to directly model the mean of counts, making them not compatible with nor comparable to competing count regression models, such as the log-linear Poisson, negativebinomial or generalized Poisson regression models. In this note, we review how CMP distributions can be parametrized via the mean, so that simpler and more easily interpretable mean-models can be used, such as a log-linear model. Moreover, this note introduces the *R* package: *mpcmp* which provides a collection of functions for estimation, testing and diagnostic checking for the proposed model. The performance of the *R* routine against the earlier proposed *MATLAB* routine will also be discussed

Keywords

count data; generalized linear model; overdispersion; R programming, underdispersion

1. Introduction

Conway–Maxwell–Poisson (CMP or COM-Poisson) distributions have seen a recent resurgence in popularity for the analysis of dispersed counts (see for instance: Shmueli et al. (2005); Lord et al. (2008); Lord et al. (2010); Sellers and Shmueli (2010)). Key features of COM-Poisson distributions include the ability to handle both overdispersion and underdispersion, containing the classical Poisson distribution as a special case, and being a continuous bridge between other classical distributions such as the geometric and Bernoulli distributions. COM-Poisson distributions are also full probability models, making them particularly useful for predictions and estimation of event probabilities. See Shmueli et al. (2005) for a recent review of the CMP distributions.

The CMP distribution was first used by Conway and Maxwell (1962) as a model for queueing system with dependent service times. A random variable is said to have a (standard) CMP distribution with rate parameter λ and dispersion parameter ν if its probability mass function (pmf) is given by

$$P(Y = y\,|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu}\frac{1}{Z(\lambda,\nu)},\ y = 0, 1, 2, \ldots,$$

Where $Z(\lambda,\nu) = \sum_{y=0}^{\infty}\frac{\lambda^y}{(y!)^\nu}$ is the normalizing constant. Obviously, one can recover the Poisson special case by setting ν=1.

One of the major limitations of CMP distributions is that it does not have closed-form expression for its moments in terms of the parameters λ and ν but satisfy recursive formulas:

$$E(Y^{r+1}) = \lambda E(Y+1)^{1-\nu}, r = 0; = \lambda\frac{d}{d\lambda}E(Y^r) + E(Y)E(Y^r), r > 0.$$

For the first two moments, approximation can be obtained as

$$E(Y) = \frac{\partial \log Z}{\partial \log \lambda} \approx \lambda^{\frac{1}{\nu}} - \frac{\nu-1}{2\nu};$$
$$Var(Y) = \frac{\partial^2 \log Z}{\partial (\log\lambda)^2} \approx \frac{1}{\nu}\lambda^{\frac{1}{\nu}} \approx \frac{1}{\nu}E(Y),$$

and they can be particularly accurate for $\nu \leq 1$ or $\lambda > 10^{\nu}$ (see Shmueli et al. (2005)). Notice that $\mu \neq \lambda$ unless $\nu = 1$ (i.e. the Poisson special case). Generally speaking, ν<1 implies overdispersion and ν>1 implies underdispersion relative to a Poisson distribution with the same mean.

As CMP is one of a few distributions that can handle both under- and over-dispersion, the aim is to extend the Generalized Linear Model (GLM) formulation to the CMP case so that one can model the relationship between Y and the predictors X. Given a set of covariates $X \in R^q$, Sellers and Shmueli (2010) proposed a GLM for count response Y that can be specified via

$$Y|X \sim CMP(\lambda,\nu) \text{ s. t. } \log\lambda = X^T\beta,$$

where $\beta \in R^q$ is a vector of regression coefficients. This structure forms the basis of the R package COMPoissonReg of Sellers, Lotze and Raim (2017) and CompGLM of Pollock (2018). However, this model does not provide a closed form relationship between $E(Y)$ and the linear predictor, making it incompatible with other commonly used log-linear model.

In this note, we review a GLM that based on the mean-parametrized version of the CMP distribution of Huang (2017) which in turn forms the basis of our R package mpcmp in Section 2. In Section 3, we illustrate the performance of our algorithm with a real dataset and a simulation study. We will discuss and conclude our findings in Section 4.

2. Methodology

As it is more convenient and interpretable to model the mean $\mu = E(Y) > 0$ of the distribution directly, Huang (2017) proposed to parametrize the CMP distribution via the mean:

$$P(Y = y|\mu, \nu) = \frac{\lambda(\mu,\nu)^y}{(y!)^\nu}\frac{1}{Z(\lambda(\mu,\nu),\nu)}, y = 0, 1, 2, \ldots,$$

where the rate $\lambda(\mu,\nu)$ is given by

$$0 = \sum_{y=0}^{\infty}(y-\mu)\frac{\lambda^y}{(y!)^\nu}.$$

We shall call this the CMPb distribution to distinguish it from the original/standard one. We can then define a GLM that based on CMPμ via

$$Y|X \sim \mathrm{CMP}_\mu\left(\mu\left(X^{\mathrm{T}}\beta\right), \nu\right),$$

where

$$E(Y|X) = \mu\left(X^{\mathrm{T}}\beta\right) = \exp\left(X^{\mathrm{T}}\beta\right).$$

There are a few advantages of using the mean parametrization version over the standard one. The biggest advantage is that the model is now comparable and compatible with other log-linear models. If ν is considered as known, then CMPb belongs to the one-parameter exponential family, which is quite similar to a model based on the negative binomial (NB) distribution. This means that one can utilize all the theoretical and practical results of the standard GLM here. It is also easier to carry out inference when it is mean-parametrized as the mean μ and dispersion ν of the CMPb are orthogonal. In contrast, the rate λ and dispersion ν in the standard CMP are not. One can also incorporate offset terms into the model for CMPb just like any other standard GLM but it is not as easy for the standard CMP model.

In Huang (2017), the estimation of the CMPb model was implemented in *MATLAB* via *fmincon*. Here we shall discuss the implementation in our *R* package: *mpcmp*. It is a package that provides parameter estimates for a **M**ean-**P**arametrized **CMP** (mpcmp) log-linear regression and associated standard errors; a LRT for testing data dispersion, particularly on whether ν = 1; and other model diagnostic tools. To the best of our knowledge, *NLOPT_LD_SLSQP* algorithm in the *nloptr* package of Ypma (2017), which is the *R* interface to *NLopt* library of Johnson (2014), is the closest routine to *fmincon* in *R*. However, we shall demonstrate in Sections 3 and 4 that the performance is lacklustre. That is the main reason we decided to implement the iterative Fisher Scoring (FS) algorithm in our package to take advantage of the fact that CMPμ belongs to the exponential family.

Here, we will briefly describe the algorithm we used. Starting with some initial guesses of $\hat{\beta}^{(0)}$ (for example those from Poisson GLM work well), we use

FS to update $\hat{v}^{(0)}$ from 1 (i.e. Poisson special case) to a more sensible initial value for v. We can then update $\hat{\beta}^{(m)}$ and $\hat{v}^{(m)}$ together with FS until (log) -likelihood converges. If the (log)-likelihood is not increasing or $\hat{v}^{(m)}$ is out-of-bound, we will carry out a half-step correction (Marschner (2011)) to the parameters i.e.

$$\hat{\beta}^{(m)} \leftarrow \frac{1}{2}\left(\hat{\beta}^{(m)} + \hat{\beta}^{(m-1)}\right) \text{ and/or } \hat{v}^{(m)} \leftarrow \frac{1}{2}\left(\hat{v}^{(m)} + \hat{v}^{(m-1)}\right).$$

Notice that the (log-)likelihood is in terms of $\lambda(\mu,v)$, so we need to solve for $\lambda^{(m)}$ from

$$\exp(X^{\mathrm{T}}\beta) = \mu = \sum_{y=0}^{\infty} y \frac{\lambda^{y}}{(y!)^{v} Z(\lambda, v)},$$

whenever we generated a new update to $\hat{v}$ or β h in order to maintain the mean constraints, by using a combination of bisection and Newton Raphson updates. In the next section, we will use a real dataset as an example to demonstrate our package. The performance of our FS routine against the earlier proposed *MATLAB* routine will also be illustrated by a simulation study.

3. Result

The attendance dataset (from https://stats.idre.ucla.edu/stat/stata/dae/nb_data.dta and is also available as part of the mpcmp package) examines the relationship between the number of days absent from school and the gender, maths score and academic programs (General, Academic & Vocational) of 314 students from two urban high schools. From the histograms in Figure 1, it is obvious that the academic programs are having an impact to the number of days absent from school and the response variable is over-dispersed. This means that it is appropriate to apply the CMP models here. For comparison purposes, we fitted the standard CMP, the CMPμ in MATLAB with fmincon as well as in R with our FS algorithm and the negative binomial GLM. The results are summarized into Table 1 below.

	CMP	CMP_{μ} (*MATLAB*)	CMP_{μ} (R-FS)	Neg Bin
Intercept	0.018	2.715	2.715	2.707
I (Male)	-0.035	-0.215	-0.215	-0.211
I (Academic)	0.050	-0.425	-0.425	-0.425
I (Vocational)	-0.232	-1.254	-1.254	-1.253
Math Score	-0.001	-0.006	-0.006	-0.006
Dispersion	0.021	0.020	0.020	1.047

Table 1: Parameter estimates of standard CMP, mean parametrized CMP and Negative Binomial models to the attendance dataset where $I(.)$ is the indicator function

We also present the results of a simulation study here to demonstrate the performance of our algorithm. We first sample n sets (n = 50, 100, 314) of covariates from the original attendance dataset, then conditional on these covariates we generate count responses via

$$Y|\,\text{gender, program, math score} \sim \text{CMP}_{\mu}(\mu, \nu = 0.02),$$

where the mean μ is given by the fitted model:

$$\log\mu = 2.707 - 0.211\, I(\text{Male}) - 0.425 I(\text{Academic}) - 1.253\, I(\text{Vocational}) - 0.006\, \textit{math score},$$

with $I(.)$ being the indicator function.

We then fit the CMPμ model to the simulated datasets with three different implementations: the *fmincon* in MATLAB, the *nloptr* routine in *R* and the FS algorithm of our package which is also in *R*.

The simulation is repeated for 1,000 times and the performance of the three models, in terms of accuracy of the parameter estimates and runtime, are summarized in Figures 2—3 respectively.

4. Discussion and Conclusion

From Table 1, we can see that FS estimates from our package are identical to those obtained from MATLAB. As both the CMPμ and the negative binomial are parameterized via the mean, the covariate estimates are quite similar and can be interpreted in the same way. For instance, students in the General program (the reference level) are expected to miss exp(1.254) = 3.5 times more days of school compared to students in the Vocational program. The parameters of the standard CMP unfortunately cannot be interpreted easily. The package also has a range of standard accessor functions as well as S3 print, summary, and plot methods. The summary output for the attendance dataset as well as a sample of standard GLM diagnostic plots that our package can provide are shown in Figure 4—5. Please refer to our package for more details.

Let's focus on the results from the simulation studies. From Figure 2, we can see that parameter estimates obtained from our FS algorithm is just as accurate as the other two routines. In terms of runtime, the FS routine is much quicker than the one from nloptr, However, our FS routine, while competitive, is less efficient than the MATLAB routine. Using Rcpp to speed up the routine is being considered at the moment.

References

1. Conway, R. W., & Maxwell, W. L. (1962). A queuing model with state dependent service rates. J.Indstrl Engng, 12, 132—136.
2. Fung, T., Alwan, A., Wishart, J. & Huang, A. (2019). mpcmp: Mean-Parametrized Conway-Maxwell Poisson (COM-Poisson) Regression. R

package. Available on CRAN and https://github.com/thomas-fung/mpcmp.
3. Huang, A. (2017). Mean-parametrized Conway-Maxwell-Poisson regression models for dispersed counts. Statistical Modelling, 17, 359—380.
4. Johnson, S.G. (2014). The NLopt non-linear-optimization package, https://nlopt.readthedocs.io/.
5. Marschner, I. C. (2011). glm2: Fitting Generalized Linear Models with Convergence Problems. R Journal, 3, 12—15.
6. Pollock, J. (2018). CompGLM: Conway-Maxwell-Poisson GLM and Distribution Functions. R package.
7. Sellers, K. F., Lotze, T. & Raim A. (2017). COMPoissonReg: Conway-Maxwell-Poisson Regression. R package.
8. Sellers, K. F. & Shmueli, G. (2010). A flexible regression model for count data. Ann. Appl. Stat., 4, 943—961.
9. Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. J. R. Statist. Soc. C. (Appl. Statist), 54, 127—142.
10. Ypma, J. with contributions by Borchers, H.W. and Eddelbuettel, D. (2017). nloptr: R interface to NLopt. R package.

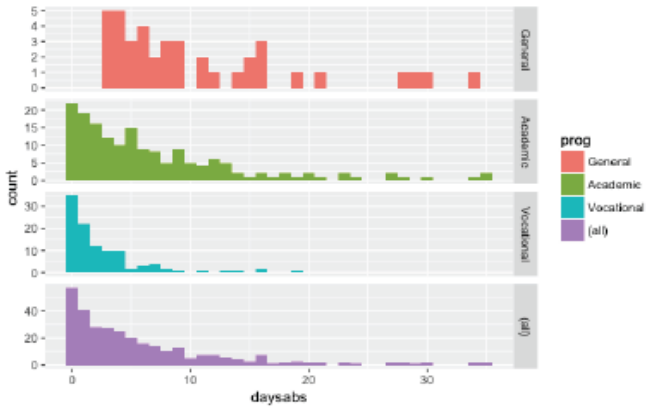


Figure 1: Histograms of the number of days absent from school for different academic programs of 314 students from two urban high schools in the attendance dataset.

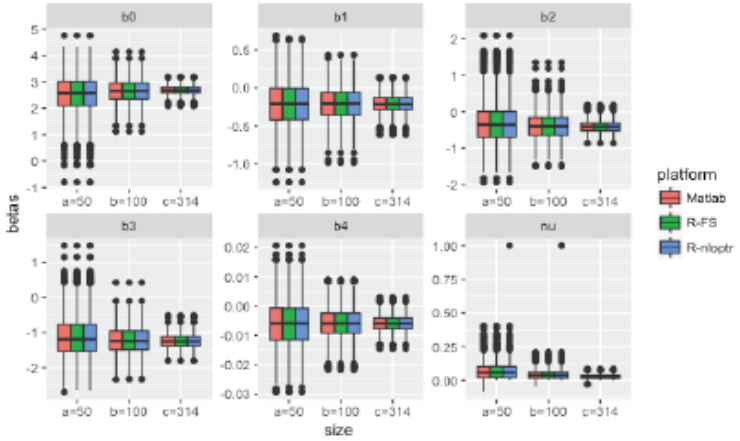


Figure 2: Boxplots of the parameter estimates of three different implementation of the mean-parametrized CMP distributions for the simulated datasets

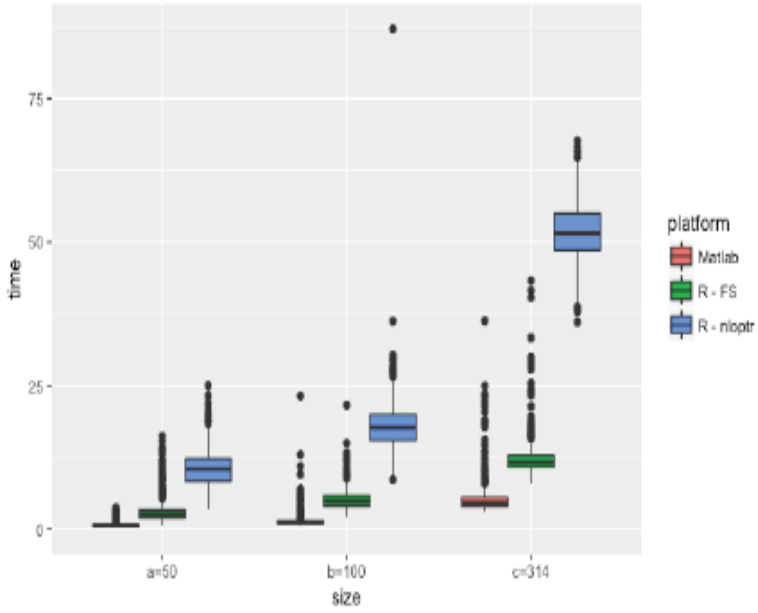


Figure 3: Runtime results of the three different implementation of the mean-parametrized CMP distribution for the simulated dataset

```
> summary(M.attendance <- glm.cmp(daysabs~ gender+math+prog, data=attendance))

Call: glm.cmp(formula = daysabs ~ gender + math + prog, data = attendance)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1923  -1.1165  -0.3974   0.2966   2.8155

Linear Model Coefficients:
              Estimate Std.Err Z value Pr(>|z|)
(Intercept)     2.7149  0.1902   14.27  < 2e-16 ***
gendermale     -0.2149  0.1172   -1.83    0.067 .
math           -0.0063  0.0024   -2.65    0.008 **
progAcademic   -0.4253  0.1694   -2.51    0.012 *
progVocational -1.2540  0.1894   -6.62  3.6e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Mean-CMP(0.02015) family taken to be 1)

    Null deviance: 456.22  on 313 degrees of freedom
Residual deviance: 377.86  on 309 degrees of freedom

AIC: 1739.023
```

Figure 4: Summary of the fitted model for the attendance dataset

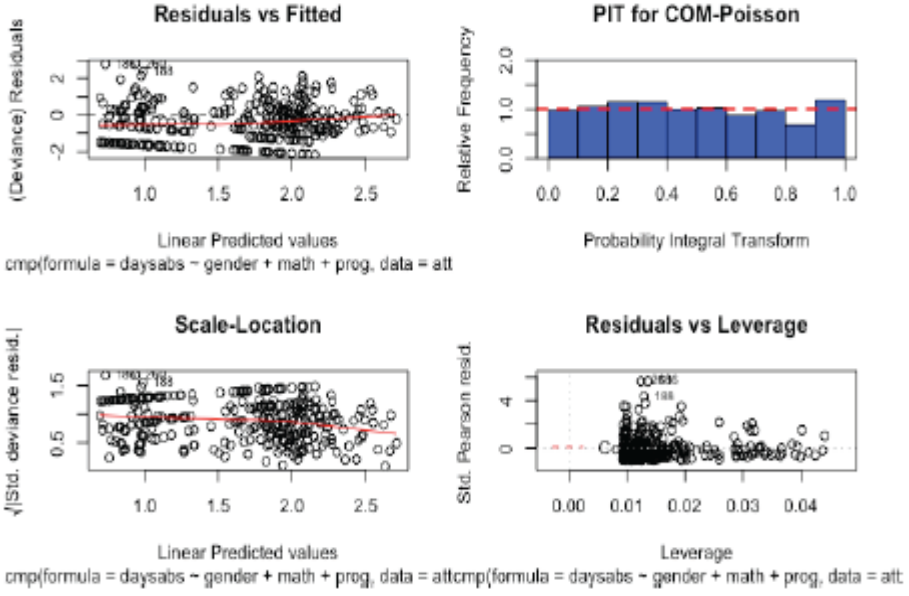


Figure 5: A sample of the diagnostic plots from the mpcmp package for the attendance dataset

Trends in the extremes of environments associated with severe US thunderstorms, and signals in their spatial dependence

Jonathan Koh; Erwan Koc; Anthony Davison
EPFL, Lausanne, Switzerland

Abstract

Concurrently high values of convective available potential energy (CAPE) and storm relative helicity (SRH) are conducive to hazardous convective weather (tornadoes, hail and damaging wind) associated with severe thunderstorms. Hence, it is highly relevant to have probabilistic models for both variables' extremes that use relevant covariate information to account appropriately for their spatial and temporal dependence structures. We consider a large area of the contiguous United States over the period 1979–2015 and use statistical extreme value models and appropriate multiple testing procedures. Various features of the data motivate a two-step model approach. In the first step, we perform pointwise modelling of each grid cell. Here, we show that there is a significant time trend in the extremes for CAPE maxima in April, May and June, for SRH maxima in May, and for the maxima of PROD = $\sqrt{CAPE}$ × SRH in April, May and August. Moreover, we show that the El Niño-Southern Oscillation explains variation in the extremes of PROD and SRH in February. Our results imply that the risk of severe thunderstorms in April and May tends to increase in parts of the US where this risk was already high and that the storm risk in February tends to be higher over the main part of the region during La Niña years. In the second stage, we perform spatial dependence modelling of each grid cell with models for spatial extremes. We incorporate covariate effects and anisotropy into the dependence modelling and show that the monthly extreme SRH events are more spatially localised in the spring and summer months than in the winter months. Furthermore, we notice that there is a signal with ENSO for the SRH maxima dependence in February.

Keywords

applied statistics; generalized extreme value distribution; extreme-value theory; max-stable processes; climate science.

1. Introduction

Annual losses from severe thunderstorms in the US have exceeded $10 billion in recent years.[14] In addition to economic losses, 2011 was marked by 552 deaths caused by tornadoes. These economic and human impacts are a

[14]http://www.willisre.com/Media_Room/Press_Releases_(Browse_All)/2017/WillisRe_Impact_of_ENSO_on_ US_Tornado_and_Hail_frequencies_Final.pdf

strong motivation for the study of how and why US thunderstorm activity varies from year to year and region to region. Two important aspects of the variability of thunderstorm activity are trends, potentially related to climate change or multi-decadal variability, and modulation by the El Niño-Southern Oscillation (ENSO). However, inadequacies in the length and quality of the thunderstorm data record present substantial challenges to addressing these questions directly (Verbout et al., 2006). Given the limitations of the historical storm record, analysis of meteorological environments that are associated with severe thunderstorms is a practical alternative approach. Environments that are favourable to severe thunderstorms, especially supercell storms, include convective available potential energy (CAPE) and vertical wind shear (see, e.g., Brooks et al., 2003; Brooks, 2013). Various measures of vertical wind shear have been used for this purpose, including storm relative helicity (SRH).

However, there are notable gaps in previous statistical studies of environments associated with severe thunderstorms. For instance, relationships with ENSO were diagnosed based on monthly averages, a quantity that is at best an indirect proxy for behaviour on the time-scale of weather. Similarly, Gensini and Brooks (2018) computed monthly accumulations of daily maxima of a significant tornado parameter. These gaps motivate the present work, which focuses on extremes of the environmental values rather than monthly means and presents results that are spatially and temporally resolved. The framework that we use is statistical extreme-value theory.

Our study covers a large part of the US for individual months from 1979 to 2015. We carefully check the relevance of the GEV and the use of explanatory variables in the location parameter of the GEV, and we account for multiple testing by implementing the false discovery rate procedure of Benjamini and Hochberg (1995). The data we investigate consist of 3-hourly time-series of 0-180 hPa convective potential energy (CAPE, Jkg^{-1}) and 0–3 km storm relative helicity (SRH, m^2s^{-2}) from 1 January 1979 at 00:00 to 31 December 2015 at 21:00. The region covered is a rectangle over the contiguous US from -110° to -80° longitude and 30° to 50° latitude and the resolution is 1° longitude and 1° latitude. These data constitute a coarse version of reanalysis data from the North American Regional Reanalysis (NARR); the original resolution is 32km longitude and 32km latitude. The region contains 651 grid points, with no data available for 32 grid points over the sea or lakes. Using these time series, we build 3-hourly times series of PROD = $\sqrt{CAPE} \times$ SRH, measured in m^3s^{-3}, which is highly representative of the risk of severe thunderstorms (see, e.g., Brooks et al., 2003, especially Figure 1 and Equation (1)). As a physical covariate we use monthly values of the NINO 3.4 index ($^{\circ}$C) from 1979 to 2015,

taken from the ERSSTv5 data set available on the NOAA Climate Prediction Center website.

In the first stage, we separately consider CAPE, SRH and the combined variable PROD = $\sqrt{CAPE}$ × SRH. We carefully check the relevance of the GEV and the use of explanatory variables in the location parameter of the GEV, and we allow for multiple testing. Finally, we consider ENSO as a covariate in the location parameter of the GEV. April and May are important months for PROD, as severe thunderstorms are frequent at this period. The corresponding time slope is positive in already risky regions of the US, which may have important impact in terms of risk assessment and management. In addition, our study reveals that ENSO can explain variation in the location parameter of the GEV for PROD and SRH maxima in February. The corresponding slope for SRH is negative over most of the region we consider, possibly suggesting an increase of storm risk in February during La Niña years.

In the second stage, we aim to adequately capture the local spatial dependence structure inherent in the observations through the theory of max-stable processes, an infinite dimensional extension of multivariate extreme value theory. The max-stable models we propose rely on a space transformation that accounts for directional effects resulting from synoptic weather patterns.

2. Methodology

2.1 First Step

Risk assessment entails the estimation of return levels associated with very high return periods and of the probabilities of observing events so extreme that they have never occurred before. Extreme-value theory provides a solid framework for the extrapolation needed to perform these tasks for the maxima of PROD, CAPE and SRH.

Let M_n denote the maximum of the independent and identically distributed random variables $X_1, \ldots, X_n$. The extremal types theorem states that if there exist sequences $\{a_n\} > 0$ and $\{b_n\} \in \mathbb{R}$ such that $(M_n - b_n)/a_n$ has a non-degenerate limiting distribution as $n \to \infty$, then this must be a generalized extreme-value (GEV) distribution,

$$\mathrm{GEV}_{\eta,\tau,\xi}(x) = \begin{cases} \exp\left[-\{1+\xi(x-\eta)/\tau\}_+^{-1/\xi}\right], & \xi \neq 0, \\ \exp\left[-\exp\{-(x-\eta)/\tau\}_+\right], & \xi = 0, \end{cases} \quad x \in \mathbb{R},$$

where ξ and η are real-valued, τ > 0 and, for any real a, $a_+ = \max\{a,0\}$. This implies that if n is large enough, we may approximate the distribution of M_n by

$$\mathbb{P}(M_n \leq x) \approx \mathrm{GEV}_{\eta,\tau,\xi}(x), \quad x \in \mathbb{R},$$

for suitably chosen η, τ and ξ. The parameters η, τ and ξ are the location, scale and shape parameters. The GEV approximation for maxima remains valid if the variables are dependent, provided that distant extremes are "nearly independent" (more formally, that Leadbetter's $D(u_n)$ condition is satisfied). Exploratory time series analyses (not shown here) show that this appears to be the case for our time series, so (2) applies. The results above provide a natural model for maxima of stationary sequences. To apply this model we split the data into blocks of equal lengths and compute the maximum of each block. Assume that we have T blocks of length n and let $M_n^{(1)},\ldots, M_n^{(T)}$ denote the corresponding maxima. If n is large enough, the distribution of the $M_n^{(t)}$ is approximately (2), upon which inference can be based; this is the so-called block maximum method. To incorporate a time trend and/or a relation with ENSO for some months, we can allow the GEV parameters to depend upon these variables. Figure 1 and results in Section 3 show that the temporal or ENSO effects only appear for certain months. For instance, time trends for PROD, CAPE and SRH are mainly present in April and May, April to June and April and May, respectively. We therefore choose our blocks to be the months and study each month separately, fitting the models

$$M_n^{(t)} \sim \mathrm{GEV}_{\eta_{\mathrm{ti}}(t),\tau_{\mathrm{ti}},\xi_{\mathrm{ti}}}, \quad \eta_{\mathrm{ti}}(t) = \eta_{0,\mathrm{ti}} + \eta_{1,\mathrm{ti}} t, \quad t = 1,\ldots,T, \tag{3}$$

and

$$M_n^{(t)} \sim \mathrm{GEV}_{\eta_{\mathrm{en}}(t),\tau_{\mathrm{en}},\xi_{\mathrm{en}}}, \quad \eta_{\mathrm{en}}(t) = \eta_{0,\mathrm{en}} + \eta_{1,\mathrm{en}}\mathrm{ENSO}_t, \quad t = 1,\ldots,T, \tag{4}$$

where $\eta_{0,ti}$, $\eta_{1,ti}$, $\eta_{0,en}$, $\eta_{1,en}$, ξ_{ti} and ξ_{en} are real-valued, τ_{ti} and τ_{en} are positive, ENSO_t is the value of ENSO in that month for year t, and n equals the number of days in the month, as we have eight observations per day.

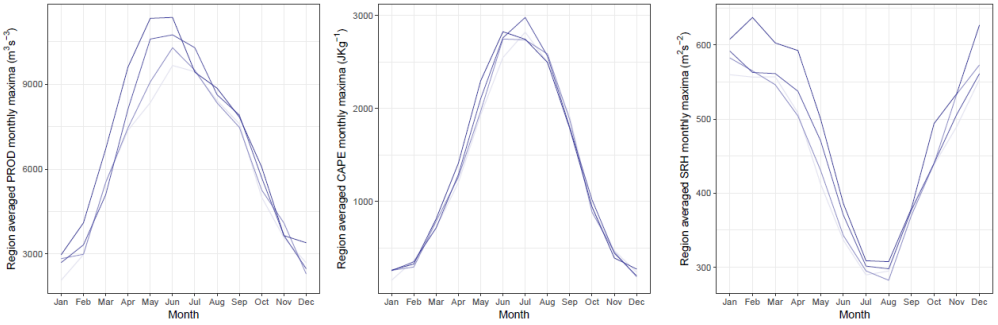


Figure 1: Whole region-averaged monthly maxima of PROD (left), CAPE (centre) and SRH (right). The four lines coloured from light blue to dark blue correspond to the time periods 1979–1987, 1988–1996, 1997–2005 and 2006–2015, respectively.

We compute the monthly maximum for each month and a given grid point and thereby obtain the maxima $M_{31}^{(1)},\ldots,M_{31}^{(37)}$ for January, say. We then fit the models (3) and (4) by numerical maximum likelihood estimation for each month and grid point.

The last step in the first stage is to assess whether time and ENSO affect the location parameter of the fitted GEV for the three variables PROD, CAPE and SRH. However, as this is assessed at 619 grid points, we must make some allowance for multiple hypothesis testing. The statistic used to test the significance of time and ENSO, respectively, in (3) and (4). In the first case, we have to test the null hypothesis

$$H_0 : \eta_{1,\mathrm{ti}} = 0 \quad \text{versus} \quad H_A : \eta_{1,\mathrm{ti}} \neq 0,$$

by comparing the fits of the models

$$\mathcal{M}_0 : \eta_{\mathrm{ti}}(t) = \eta_{0,\mathrm{ti}}, \quad \mathcal{M}_1 : \eta_{\mathrm{ti}}(t) = \eta_{0,\mathrm{ti}} + \eta_{1,\mathrm{ti}} t, \quad t = 1, \ldots, 37,$$

and similarly for ENSO. We let $\ell_0(M_0)$ and $\ell_1(M_1)$ denote the maximized log-likelihoods for the models M_0 and M_1 and compute the signed likelihood ratio statistic $\tilde{T} = \mathrm{sgn}(\hat{n}_{1,\mathrm{ti}})[2\{\ell_1(M_1) - \ell_0(M_0)\}]^{1/2}$, where $\mathrm{sgn}(\hat{n}_{1,\mathrm{ti}})$ is the sign of the estimated trend under model M_1; $\bar{T}$ has an approximate standard Gaussian distribution under H_0, and the corresponding p-value is $p = 2\Phi(-|\bar{t}|)$, where $\bar{t}$ is the observed value of $\bar{T}$ and Φ denotes the standard Gaussian distribution function. In this study, we use the Benjamini-Hochberg (BH) procedure Benjamini and Hochberg (1995) to control the false discovery rate (FDR), namely the expected proportion of false rejections of the null hypothesis H_0 out of all rejections of it.

2.2 Second Step

The procedure of fitting the GEV at each grid point, as described in the first stage, inherently assumes that the observations are independent between grid points. Thus, it does not account appropriately for the spatial dependence structure inherent in the maxima. To do so in the spatial setting, we will use max-stable processes as models for the maxima. The Poisson process has an important role in extreme value theory (EVT); its relationship with max-stable models in the spatial settings is best understood by looking at the construction of max-stable processes through the spectral representation (de Haan, 1984)

$$Z(x) = \sup_{i=1}^{\infty} R_i W_i(x), \quad x \in \mathcal{X}, \tag{5}$$

where the R_i are generated sequentially by setting $R_i = (E_1 + \cdots + E_i)^{-1}$, with $E_i \overset{iid}{\sim} \exp(1)$. The $W_i(x)$are independent replications of a non-negative stochastic process $\{W(x), x \in X\}$ taking values in a suitablespace of functions F and are independent of the R_i. In our study, we consider parametric max-stable

models with parametric forms for $W(x)$ based on the Gaussian process, such as the Smith (Smith, 1990), Schalther (Schlather, 2002) and Brown-Resnick (Kabluchko et al., 2009) processes. Furthermore, geostatistical techniques that deal with anisotropy (Blanchet and Davison, 2011) are also applied in our setting. Given D locations, likelihood inference for max-stable models is problematic as it involves D-fold differentiation of the joint cumulative distribution function of the max-stable process; this leads to a combinatorial explosion. Inference procedures to circumvent this problem have been proposed by looking at composite likelihood procedures for the maxima (Padoan et al., 2010), and this is the approach we will use. Let $z_{x,t,m}$ denote the maximum in month m for year t at grid cell $x \in X$, transformed (using the marginal pointwise parameter estimates from the first stage) so that the the time series $\{z_{x,t,m}\}_t$ is unit-Fréchet distributed. The log-composite pairwise likelihood function for month m is:

$$l_m(\psi) = \sum_{x \neq x'} \sum_t w_{ij} \log f_\psi(z_{x,t,m}, z_{x',t,m}),$$

where ψ is the vector of dependence parameters, f is the pairwise probability density function and w_{ij} are chosen weights that are tapered (Sang and Genton, 2014), on the basis of computational feasibility and a bias-variance trade-off. We will see in Section 3 that the chosen final model is the Brown-Resnick process. The law of the Brown-Resnick process only depends on the variogram 2γ of the underlying Gaussian process used in the construction of the max-stable model. To incorporate covariate effects and anisotropy into the dependence modelling, we model the variogram as

$$2\gamma(h) = (\|Ah\| / \rho)^{\alpha_0 + \alpha_1 \mathrm{ENSO}_t},$$

where h is the spatial lag, A is a transformation matrix with scale parameter r and rotation parameter δ, $\rho > 0$ and $0 < \alpha_0 + \alpha_1 \mathrm{ENSO}_t \leq 2$, for all years t.

3. Result

Table 1 shows our results from the first stage and quantifies the effects of time and ENSO in the location parameter of the GEV and study their significance, using $q = 0.05$ and $q = 0.2$, corresponding to the control of the false discovery rate at the nominal levels 5% and 20%. We only show our results from the second stage for SRH here. Comparing the Takeuchi Information Criteria (Takeuchi, 1976) for each model, the best max-stable process is the Brown-Resnick process. From Figure 2,

Variable	Covariate	q	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
ROD	Time	0.05	7	0	1	41	36	0	0	36	2	0	0	22
	Time	0.2	40	0	4	313	203	81	13	148	23	0	0	98
	ENSO	0.05	0	58	10	0	0	1	0	0	0	0	0	1
	ENSO	0.2	1	172	26	0	3	3	0	0	0	0	0	1
CAPE	Time	0.05	37	13	28	109	60	89	18	55	4	0	30	1
	Time	0.2	92	37	73	268	273	206	75	133	35	40	134	16
	ENSO	0.05	15	0	0	0	0	2	2	0	0	0	1	1
	ENSO	0.2	27	11	21	0	0	3	16	14	0	1	6	13
SRH	Time	0.05	0	1	0	7	43	2	1	7	0	0	0	0
	Time	0.2	15	44	4	138	230	14	50	45	6	0	0	27
	ENSO	0.05	0	255	0	0	1	0	0	0	0	0	0	0
	ENSO	0.2	3	384	59	18	4	0	8	7	4	1	0	82

Table 1: Number of grid points where $\hat{\eta}_{1,ti}$ and $\hat{\eta}_{1,en}$ are significant for PROD, CAPE and SRH maxima for each month (top). We have accounted for multiple testing using the BH procedure with the values of q displayed.

there is clear statistical evidence against spatial isotropy, i.e., the null hypotheses of r and δ being equal to 0 are rejected at the 5% level for all months. There is also evidence for non-stationarity in ENSO for February, i.e., $\alpha_1 > 0$. Furthermore, the estimates for the dependence parameters vary across months.

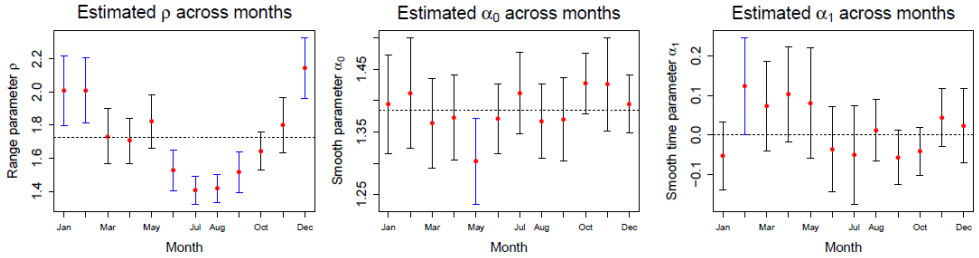


Figure 2: Estimated Brown-Resnick parameters for SRH across months. The 95% confidence bands are given by the whiskers. The dotted lines show the average estimated parameter value across all months (left and middle panel) and the value 0 (right panel).

The extremal dependence among two fixed locations can be summarised by the pairwise extremal coefficient (Schlather and Tawn, 2003). Figure 3 shows that the high SRH values are more localised in the winter than in the spring/summer months. Moreover, there is clear statistical evidence against spatial isotropy, i.e., the null hypotheses of r and δ being equal to 0 are rejected at the 5% level for all months.

4. Discussion and Conclusion

In the first stage, we quantify the effects of time and ENSO on the distribution of monthly maxima of PROD, CAPE and SRH, which are highly relevant to the risk of severe thunderstorms. Using an appropriate treatment

of multiple testing, we point out the existence of a significant time trend in the location parameter of the GEV for PROD maxima in April, May and August, CAPE maxima in April, May and June and SRH maxima in April and May. The latter months are prominent for PROD, as severe thunderstorms are recurrent at this period. The corresponding time slope is positive in parts of the US where the risk was already high, which may have important consequences. We also found ENSO to be a good covariate in the location parameter of the GEV for PROD and SRH maxima in February. The corresponding relationship for SRH is negative over most of the region we consider, perhaps implying that storm risk increases in February during La Niña years. Various in-sample and out-sample checks were performed to validify the good fit of the GEV. In the second stage, we show the existence of anisotropy and covariate effects in the dependence structure of the best fitted max-stable process. Furthermore, we see that extreme SRH events are more spatially localised in the spring/summer. Our validation procedures confirmed that the best simple

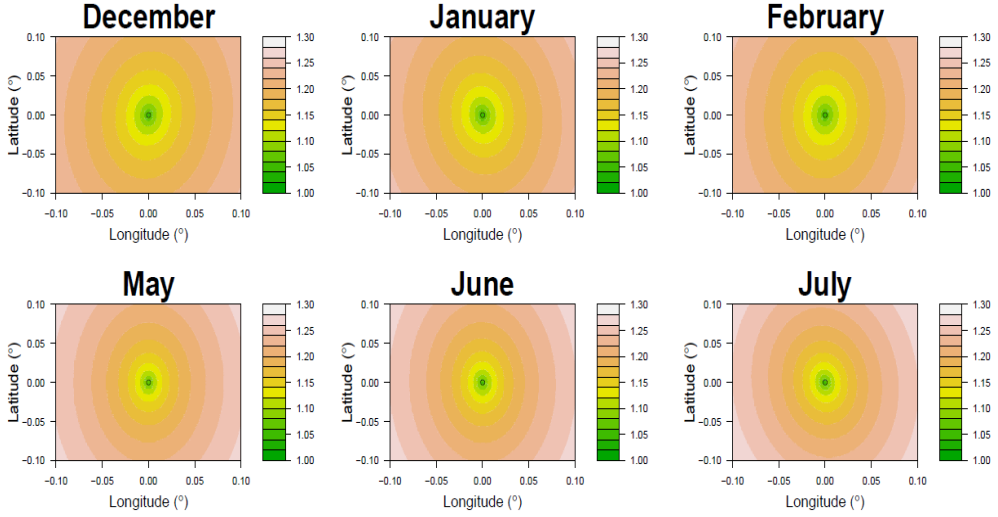


Figure 3: Estimated SRH extremal coefficient (with one point fixed at the origin) for the winter months (top panel) and for the spring/summer months (bottom panel).

max-stable model provides a better fit to the joint behaviour of the extremes than do independence or full dependence models. Future work should address another approach that also takes the dependence between the two variables into consideration.

References

1. Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 57(1):289–300.
2. Blanchet, J. and Davison, A. C. (2011). Spatial modeling of extreme snow depth. The Annals of Applied Statistics, 5(3):1699–1725.
3. Brooks, H. E. (2013). Severe thunderstorms and climate change. Atmospheric Research, 123:129–138.
4. Brooks, H. E., Lee, J. W., and Craven, J. P. (2003). The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. Atmospheric Research, 67-68:73–94. deHaan, L.(1984). A spectral representation for max-stable processes. The Annals of Probability, 12(4):1194– 1204.
5. Gensini, V. A. and Brooks, H. E. (2018). Spatial trends in United States tornado frequency. npj Climate and Atmospheric Science, 1:38.
6. Kabluchko, Z., Schlather, M., and de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. The Annals of Probability, 37(5):2042–2065.
7. Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010). Likelihood-based inference for max-stable processes. Journal of the American Statistical Association, 105(489):263–277.
8. Sang, H. and Genton, M. G. (2014). Tapered composite likelihood for spatial max-stable models. Spatial Statistics, 8:86–103.
9. Schlather, M. (2002). Models for stationary max-stable random fields. Extremes, 5(1):33–44.
10. Schlather, M. and Tawn, J. A. (2003). A dependence measure for multivariate and spatial extreme values: properties and inference. Biometrika, 90(1):139–156.
11. Smith, R. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.
12. Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. SuriKagaku (Mathematical Sciences), 153:12–18.
13. Verbout, S. M., Brooks, H. E., Leslie, L. M., and Schultz, D. M. (2006). Evolution of the U.S. tornado database: 1954-2003. Weather and Forecasting, 21:86–93.

Comparing the household final consumption expenditure in national accounts to the household budget survey – or vice versa?

Katri Soinne[1, 2]
[1]Statistics Finland
[2] University of Helsinki

Abstract
The Household Budget Survey (HBS) is the main data source for the household final consumption expenditure (HFCE) calculations in National Accounts (NA). Since the HBS is not usually conducted every year, we need other sources and methods to calculate the household consumption figures for years without the HBS. To evaluate the quality and reliability of those other methods and sources, I am comparing levels and changes in the national accounts calculations and the Household Budget Survey. On the other hand, it is known that the HBS is not reliable in all products and I am trying to recognise those products, because other sources are needed to estimate the consumption of those products even for the years with the HBS. As an example in this paper, I am looking at costs related to health.

Keywords
household consumption; average yearly change; COICOP 06 Health

1. Introduction

1.1 The System of National Accounts and the household consumption

The System of National Accounts (SNA) is an international recommendation for measuring economic activities, and it is probably the most widely used statistical system in the world. National accounts is based on a theory by Keynes, and it covers, for example, the supply and demand of goods and services. The first version of SNA was published in 1953 and already that version included the very basic idea: production, income and consumption are all equal. Revised versions of SNA have been published in 1968, 1993 and 2008, as an answer to changing economic realities. The very basic ideas are still the same, but since, for example, the production processes and international relationships are getting more complicated, the statistical system needs to be updated to be able to cover these changes.

The role of household final consumption expenditure (HFCE) has been constantly growing and, at the moment, it is the biggest single demand component almost all over the world. The challenge for household calculations in the national accounts is that most of the sources are indirect, there is very little direct information. The most important source for the

household final consumption expenditure calculations is the Household Budget Survey (HBS), which is a direct source, but which is not usually conducted every year. Since the HBS is available less frequently, but the national accounts figures are calculated yearly, other sources and methods are needed to calculate household final consumption expenditure in national accounts. At the moment, there are a lot of different sources and methods used to compile the HFCE calculations in annual national accounts in Finland. The statistics on turnover of trade is one of the main sources for products and the production calculations of the national accounts is one of the main sources for services. Furthermore, there are dozens of other data sources covering for example alcohol, tobacco, housing, cars, household appliances, travelling, communication etc.

My plan is to compare national accounts figures based on other sources and methods to the figures from the HBS to evaluate the quality and reliability of those other sources and methods. On the other hand, it is known that the HBS is not reliable in all products – for example the amount of money Page used on alcohol is both historically and internationally known to be unreliable – and I try to recognise those products, because other sources are needed for those even for the years of the HBS. The fact that there are some differences in definitions is causing difficulties in comparisons, too.

1.2 Household Budget Survey

The Household Budget Survey (HBS) is a sample survey, which produces data on the consumption expenditure of households, also according to household types and income groups. The HBS is collecting data usually with interviews (either by telephone or face-to-face), diaries and receipts. Often some data from administrative registers are used, too. The Household Budget Survey has been conducted in Finland in comparable form since 1966, in about every fifth year. The latest years, which are used in this comparison, are 1998, 2001, 2006, 2012 and 2016.

In the Finnish HBS, there is an interview and then households are asked to collect their receipts and keep a diary for two weeks. The non-response rate has been growing, and it affects the results, both at the total, but especially at the more detailed level. Although the sample for the HBS is planned very carefully, the non-response rate is causing bias to the sample. Non-response is not totally random, there are differences in relation to income levels, size of household, region and the education level, which affect the results. (Statistics Finland, 2014, p. 32-34.)

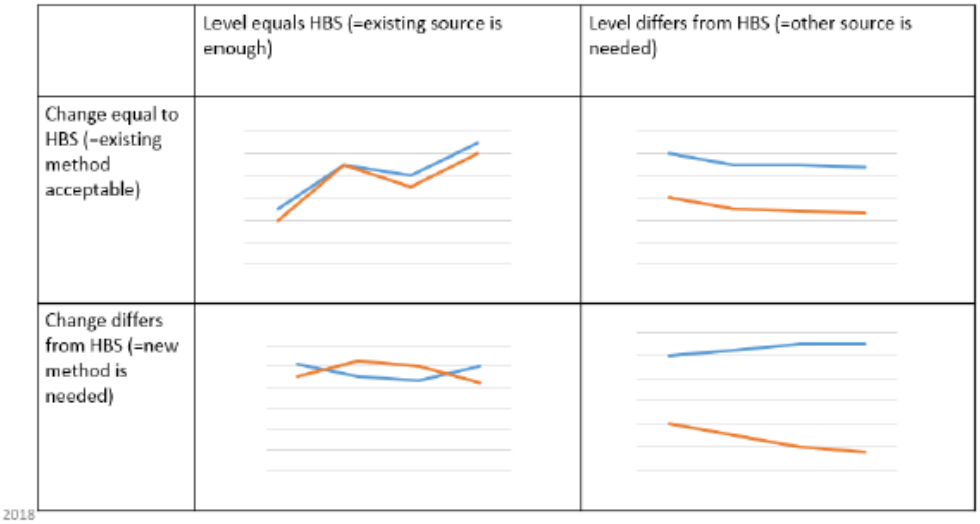
The sample of the HBS covers basically all Finnish households living in Finland, but people living in different institutions, like nursing homes or prisons, are not included.<Introduction>

2. Methodology

2.1 Comparisons to be made

Besides comparing the levels in HBS and national accounts consumption calculations, also the changes should be compared between those two. Comparing levels can give information whether the data sources for NA calculations are acceptable or not, and comparing changes can give information whether the method for NA calculations is suitable or not. If the data source or the method (or both) does not seem to produce a good result, it might be necessary to find a new data source or method. Comparing both levels and changes hopefully gives me a possibility to divide the goods and services into four groups, as in picture 1, based on which I could prioritize the search for new sources and methods.

Picture 1: Division of calculations in four groups.



2018

National accounts figures are not independent from the HBS figures, since the HBS is used as a main source, but I am comparing figures from the HBS to the figures calculated in national accounts based on other sources and methods during the years without the HBS. Behind the national accounts figures, there is always the previous HBS, but I believe that especially for defining the quality of the existing method this data is sufficient.

The figures used in my comparisons describe the actual amount of money used by households in a certain year. Since there is often inflation, the real consumption has not always grown as fast as it might seem. When comparing the consumption at aggregated level, it might make sense to calculate a relative difference instead of using absolute levels, so that the impact of inflation will be left out, but when making comparisons at the more detailed level, the absolute figures might give the clearest picture.

To compare the changes I have calculated the average yearly changes both from the HBS results and from the methodologically produced NA calculations for the same years. I have compared the differences in changes (instead of the changes themselves) to leave the effect of inflation out.

2.2 Data used in comparisons

The HBS data used in my calculations is the original HBS data, which was published as an average per household in local currency (in euros in Finland nowadays). The HBS results are calculated and published once, those are not changed afterwards (unless a significant mistake is noticed). I have calculated the totals for all households by multiplying the averages with the amount of households. (Official Statistics of Finland (OSF): Households' consumption.)

National Accounts figures are calculated yearly, but those figures are also revised backwards when more data is available or some bigger changes are made for the whole time series. From National Accounts figures I have taken the first version of yearly calculations for comparison of the level. For comparison of the yearly changes, I have used the first version of the latter year, but the figures for the starting year are from revised NA series instead of the first version. In Finnish National Accounts the HFCE is calculated only for households as total, in million euros. (Official Statistics of Finland (OSF): Annual national accounts.)

2.3 Case: COICOP 06 Health

The classification used in household consumption is called a Classification of Individual Consumption According to Purpose (COICOP). COICOP is hierarchical and coded, and it is organized in 2-digit divisions, 3-digit groups and 4- and 5-digit classes. In this paper I look at the results for COICOP-division 06 Health, which includes both goods and services. The division is divided in three 3-digit groups:

06.1 Medical products, appliances and equipment;
06.2 Non-hospital medical and paramedical services;
06.3 Hospital services.

In the Finnish NA calculations this division is calculated in eight different 5-digit level classes. Those 5-digit glasses are then summed up in seven 4-digit classes, which are summed up in three 3-digit level groups, and those are further summed up for a 2-digit division 06. Half of calculated 5-digit classes are goods and another half services.

In Finland only the pharmacies are allowed to sell medicines, and pharmacies are reporting to the social insurance institution in Finland (called KELA according to Finnish abbreviation). KELA is publishing data on those sales, divided between public buyers and households, so we have quite good data on purchases of medicines by households in Finland. The rest of the

products include other medical products as well as therapeutic appliances and equipment, and the main data source for calculations in Finland is the turnover data, mainly for NACE classes 47730 Dispensing chemist in specialised stores and 47783 Retail sale of optical goods, but also parts of some other NACE classes are included.

For the services the main data sources are NA production calculations (including both private sector and general government) combined with data from the National Institute for Health and Welfare (called THL according to Finnish abbreviation). In Finland we have

a) public healthcare system, which is mainly funded by general government, and patients are paying certain fixed fees for the services, and those fees are part of household consumption;

b) an occupational health care, organized by private companies and paid by employers, so those costs do not belong to household consumption and

c) private healthcare services, which are paid by households either directly or via health insurances: the direct costs are part of household consumption, but the costs of health insurances are not included in division 06, they are part of insurances (and insurances are calculated as part of division 12).

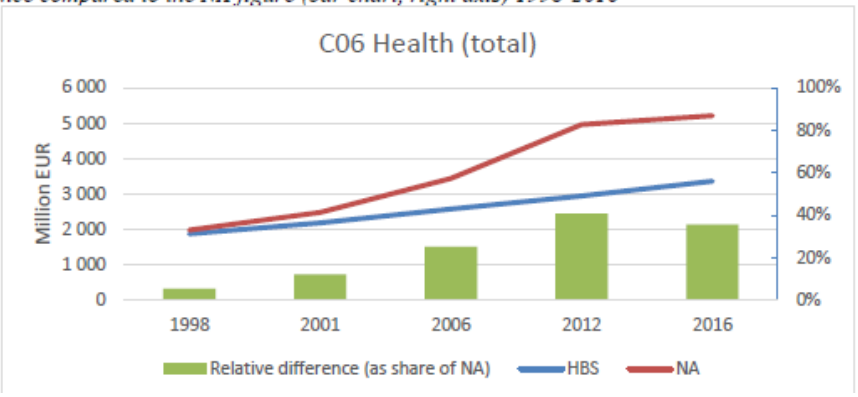
3. Result

3.1 The total level of household expenditure on health

In 1998, the expenditure on health was quite the same both in the HBS and NA, but since then the growth has been estimated to be significantly higher in NA. In 2016 the amount of money used for health was 1.5 times higher in NA than in the HBS, and the relative difference – meaning the difference in absolute figures compared to the NA figure – was about 35 percent. The results of comparison are shown in picture 2.

According to the HBS results, the total health costs are quite equally divided between goods and services in period from 1998 to 2016. In NA calculations the share was quite equal at the beginning of this period, but since then the share of services has been growing really fast, so that in 2016 goods were about 1/3 and services about 2/3 of the total costs in NA.

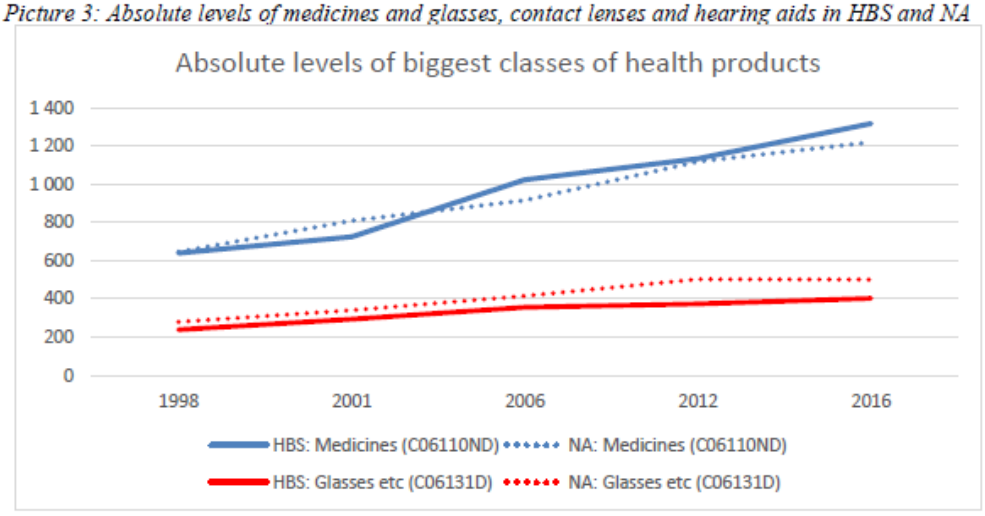
Picture 2: Total level of consumption expenditure on health (line chart, left axis) and the relative difference compared to the NA figure (bar chart, right axis) 1998-2016



3.2 Consumption expenditure at more detailed level

For products related to health, four classes are calculated in Finland. Two biggest classes are medicines (on average 71% of health products in the HBS and 66% of health products in NA) and glasses, contact lenses and hearing aids (on average 25% in the HBS and 29% in NA). The other two classes, other pharmaceutical products and other therapeutic appliances and equipment, are much smaller. The levels of medicines and glasses etc. seem reasonably similar in both the HBS and NA (picture 3). The solid lines are the HBS figures and the dotted lines are NA figures.

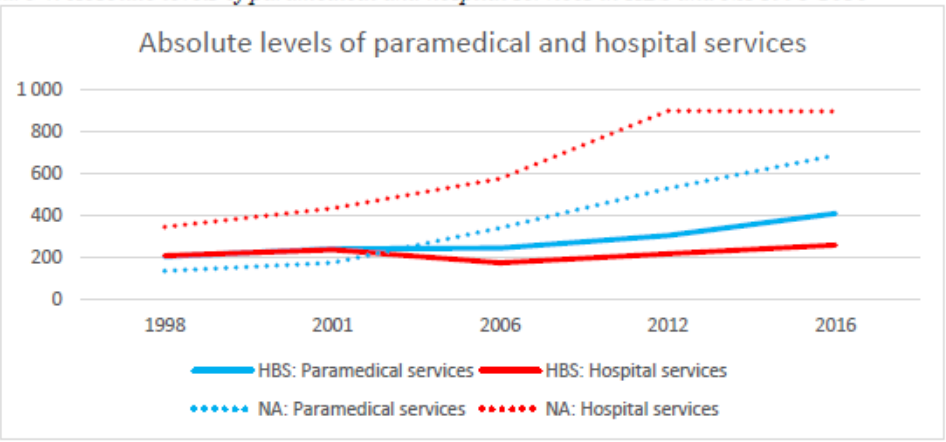
Picture 3: Absolute levels of medicines and glasses, contact lenses and hearing aids in HBS and NA



Also, for services related to health, four classes are calculated in Finland. All four classes are more or less the same size with each other on the one hand in the HBS and on the other hand in NA calculations, so I chose here the hospital services, where the relative difference is the biggest and the paramedical services, where the relative difference is the smallest. The other two groups are called medical services and dental services.

In all four groups the expenditure has been growing clearly faster in NA figures, and the level is much higher in all of them nowadays (picture 4). The solid lines are the HBS figures and the dotted lines are NA figures.

Picture 4: Absolute levels of paramedical and hospital services in HBS and NA 1998-2016

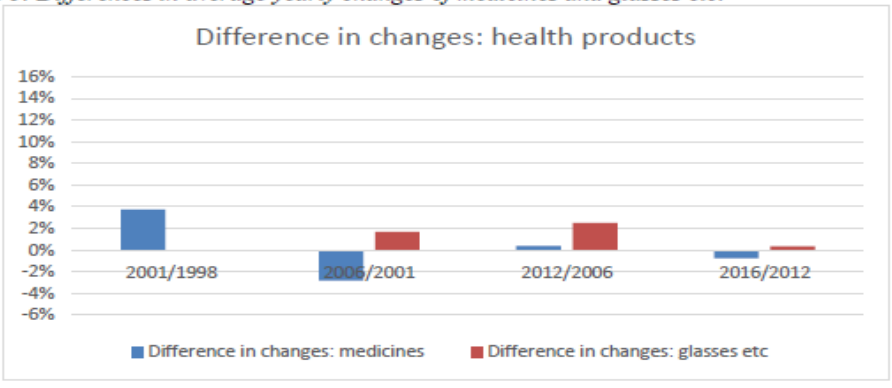


3.3 Average yearly changes

In the following pictures, the average yearly change in the HBS has been deducted from the average yearly change in NA calculations, so if the difference is positive, the average yearly change has been higher in NA (and vice versa).

Concerning the health products, the differences in average yearly changes are not very big. The change for glasses etc. has almost every time been little bigger in NA, but for medication the direction of difference has been changing. As a total, the differences in average yearly changes seem to be getting smaller.

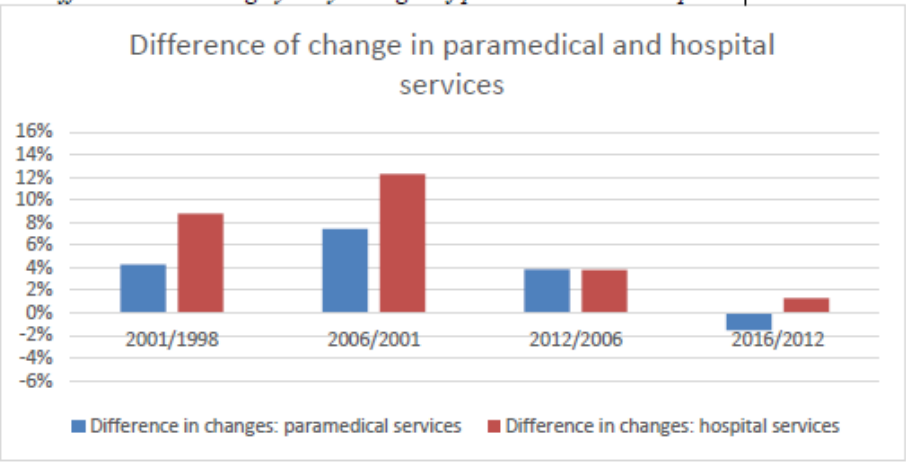
Picture 5: Differences in average yearly changes of medicines and glasses etc.



Concerning the health services, the differences in average yearly changes are significant, in both cases. The differences seem to be getting smaller, which is a good thing, and in the latest years the difference is reasonable sized, but

the earlier years have caused the difference in total level – both in absolute and relative terms – to grow very big.

Picture 6: Differences in average yearly changes of paramedical and hospital services



4. Discussion and Conclusion

According to these results, it seems that both sources and methods for calculating health products in National Accounts calculations are acceptable, but the calculations for health services should be reconsidered. Of course also the reliability of the HBS should be taken into account – is it possible that, for example, hospital services are underestimated because of non-response: households using a lot of hospital services might not have energy to participate into the survey? The difference in changes of health services seems to be getting smaller, but the absolute level is significantly higher in NA calculations – the right level should be confirmed somehow.

So far, the HBS has been the main source – as being the only real data on household consumption – for the household final consumption expenditure calculations in national accounts, but its value in the future is unclear. The descending response rate, the more divergent consumption, the new family constructions, and the new environment (like the Internet and social media) might make the results less reliable. The reliability is questioned also by the fact that consumption used to be more clearly related with purchases – you paid and you received the product, which was kind of clear and easy to write down – but nowadays, the time between purchase, payment and receiving the product might be longer, and people are shopping through "apps" without thinking of it as shopping and consumption?

The importance of consumption does not seem to diminish, rather the opposite. The yearly and quarterly calculations for household final consumption expenditure in national accounts will be done in the future, as well. At least some new data sources and methods will be needed, but I am

hoping to find out, where to start – which are the goods and services that seem to need new sources or methods or both of them. With limited resources it would be best to start with products that actually need something new, not with products where the results are acceptable already at the moment.

References

1. Household Budget Survey (figures): Official Statistics of Finland (OSF): Households' consumption [e-publication]. ISSN=2323-3028. Helsinki: Statistics Finland [referred: 20.9.2018]. Access method: http://www.stat.fi/til/ktutk/index_en.html
2. Household Final Consumption Expenditure, National Accounts (figures): Official Statistics of Finland (OSF): Annual national accounts [e-publication]. ISSN=1798-0623. Helsinki: Statistics Finland [referred: 16.8.2018]. Access method: http://www.stat.fi/til/vtp/tau_en.html
3. Statistics Finland. 2014. Kulutustutkimus 2012 – Käyttäjän käsikirja (available in Finnish, translation: Household Budget Survey 2012 – User's manual). Statistics Finland, Helsinki.
4. The System of National Accounts 1993 (SNA). 1993. https://unstats.un.org/unsd/nationalaccount/sna.asp

Contribution and growth of selected economic activities in the non-oil real GDP in the Emirate of Abu Dhabi 2007-2018

Dr. Akram Musallam Alshawawreh, Wadeema Mohamed Alkhoori
Statistics Centre-Abu Dhabi, Abu Dhabi, United Arab Emirates

Abstract

Gross Domestic Product (GDP) is the main indicator for measuring the performance of the economy in any country; due to its importance, many researchers work to determine the details, components, and changes that occur during a specific time period. Non-oil activities contribute around 50% of GDP. Due to the significant contribution of the extraction industry activity, which includes oil and natural gas, it is good to have a separate measure of the GDP. Non-oil activities in the Emirate of Abu Dhabi are the true measure of the economy; they provide a measure of the economy away from the main activity and its often price- related effects. Non-oil real GDP was also measured to neutralize price and inflation effects.

In this study, we would like to focus on the contribution and size of economic activities in the non-oil GDP and annual growth rates over a period of twelve years in the Emirate of Abu Dhabi. The paper depends on a series of statistical data published by Statistics Centre - Abu Dhabi from the year 2007 until 2018, with 2018 data being preliminary estimates. Moreover, it will include statistical tables showing the GDP, value, contribution and growth rate of Non-oil activities.

Keywords

Construction, non-oil, transportation, health, manufacturing

1. Introduction

GDP is defined as the value of all final goods and services produced in an economy over a given time period, usually one year. This definition shows that the GDP data must include both a time and place component. The time component is to determine the time represented by the data, such as during a year or part of the year (e.g. quarterly). The place component indicates the place. The production of goods and services represents the goods and services produced within the boundaries of a specific geographical area by economic units residing within the economic boundaries of the studied geographical area (Emirate of Abu Dhabi), of which individuals and enterprises are resident in the economy, it being the centre of their economic interest. This means that they will support the economy and their specific economic activities as well as remain in the Emirate for a long time.

The Emirate of Abu Dhabi is the capital of the Federal Government. It also has the largest coastline in comparison with the other emirates, which runs from the emirate of Dubai. The Emirate of Abu Dhabi has witnessed significant development over the last 40 years due to the abundance of oil and natural gas resources, which have transformed it into a global hub and a major player in the business and economic fields on the international stage.

The Emirate of Abu Dhabi is one of the seven emirates of the United Arab Emirates (UAE). Abu Dhabi is the capital of the Emirate and the capital of the UAE. The Emirate of Abu Dhabi is the largest emirate in terms of area and population. It is equivalent to 87% of the total area of the country and is projected to have a population of about 3 million by the middle of 2018. The emirate consists of three regions: the Abu Dhabi capital, Al Dhafra and Al Ain. The emirate lies on the border with Saudi Arabia, Oman and the Arabian Gulf. It has 200 islands with a coastline of 700 kilometres.

2. Methodology

This paper is based primarily on the statistical data published by the Statistics Centre - Abu Dhabi (SCAD) in the annual statistical book; the annual publication of the national accounts published by SCAD; in addition to the data published in the annual periodicals of the SCAD publications. From these publications we selected the activities in this paper. The paper is also based on data published in international organizations such as the World Bank, the International Monetary Fund (IMF) and some of the published publications on the subject. The economic activities were selected through their size in the non-oil GDP, as well as their importance to the society after the exclusion of the extraction industry, which includes crude oil and natural gas, as well as some activities that do not greatly affect the economy of the Emirate of Abu Dhabi but are important to society.

3. Abu Dhabi Emirate Economic Structure

The composition of the GDP in the Emirate of Abu Dhabi is unique in terms of economic activities and their contribution to the GDP. This combination is similar to that of some Gulf countries. This is primarily due to the predominance of the oil activity in the majority of the GDP. The other reason is the novelty of the other economic activities, which began some time after the oil industries. The oil activity accounts for about 50% of GDP in the Emirate of Abu Dhabi at constant 2007 prices.

Oil was discovered in the Emirate of Abu Dhabi in 1958; exports started in 1962. The leadership of the Emirate of Abu Dhabi realized the importance of economic diversity and having a sustainable economic development without excessive dependence on the oil activity. A number of strategically important

activities in the economy were identified as key pillars of future economic growth.

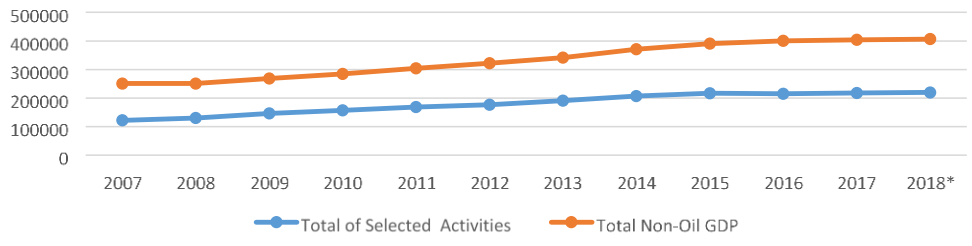
Of the economic activities that depend on the oil activity, manufacturing is the primary one. Oil is the basic material in these industries, for example the petrochemical industry, which is well-known in the Emirate of Abu Dhabi. As for the rest of the industries, they are not fully reliant on the oil industry although it remains the main source of government revenue in the Emirate.

The rest of the economic activities vary in the economy of Abu Dhabi Emirate through their size in the economy and their importance in society; this was the basis for selecting the activities in this paper.

In 2007 the GDP of the Emirate of Abu Dhabi reached AED 545,367 m at constant 2007 prices, equivalent to USD 148,500 m, reaching AED 797,278 m in 2018 or USD 217,095 m.

Figure 1 indicates that the selected activities are running almost parallel to the non-oil GDP during the study period, the latter being directly affected by the size of these activities. It shows that the total value added of the selected activities was about AED 121,997 m in 2007 and these activities reached about AED 219,682 m in 2018. The non-oil GDP was AED 239,979 m in 2007 and reached about AED 406,524 m in 2018.

Figure (1) Value added of selected activities compared to non-oil GDP



*2018 data are preliminary estimates

4. The Contribution of Selected Economic Activities

As illustrated in the attached table on the participation of the selected economic activities in the GDP of Abu Dhabi, the Construction activity is the largest activity in terms of participation, which amounted to 19.9% of the GDP in 2018. The highest percentage of construction contribution during the study period was 27.5% 2010, indicating the Emirate is overall witnessing a large increase in infrastructure, with 2010 especially highlighting the importance of this activity.

The Financial services activity, which reached 14.3% in 2018, was the second highest in the study period, with the lowest value in 2011 reaching 8.1%. The Manufacturing activity, which reached 12.1% in 2018, was highest at 14.2% in 2007 and the lowest in 2009 at 10.6%. The Transportation and storage

activity reached a 3.8% contribution in 2018, with a peak of 6.0% in both 2012 and 2013.

The remaining two activities chosen are the Health and Education activities because of their importance to society. Education reached 2.2% in 2018, while 1.9% in 2015 was the lowest percentage. Healthcare reached 1.8% in 2018, a significant increase from 0.7% in 2007.

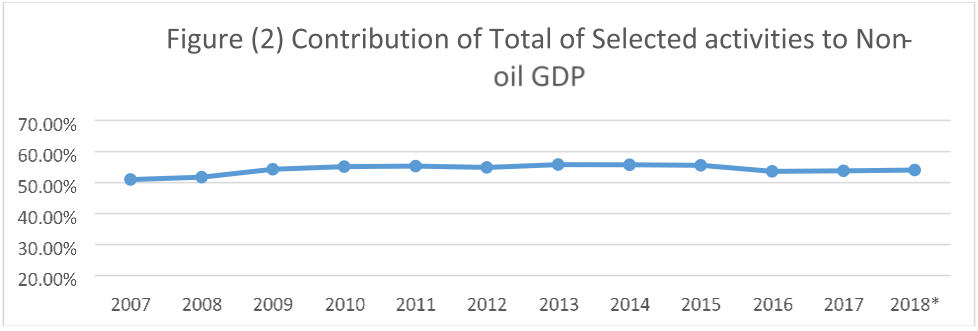
Table (1) Value Added Contribution of Selected Economic Activity in Abu Dhabi Emirate at Constant Prices 2007 - 2018*

%

ISIC4	Activities	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018*
C	Manufacturing	14.2%	12.6%	10.6%	10.8%	12.0%	11.4%	11.1%	10.9%	10.9%	11.0%	11.5%	12.1%
F	Construction	18.3%	21.5%	26.6%	27.5%	26.0%	24.5%	22.8%	21.8%	21.7%	20.6%	20.3%	19.9%
H	Transportation and storage	4.8%	5.1%	4.7%	5.6%	5.8%	6.0%	6.0%	6.0%	5.5%	4.1%	3.9%	3.8%
K	Financial and insurance activities	10.7%	9.3%	9.3%	8.3%	8.1%	9.3%	12.0%	13.1%	13.7%	14.0%	14.2%	14.3%
P	Education	2.4%	2.3%	2.3%	2.2%	2.2%	2.1%	2.1%	2.0%	1.9%	2.1%	2.2%	2.2%
Q	Human health and social work activities	0.7%	0.9%	0.8%	0.8%	1.1%	1.5%	1.9%	1.9%	1.9%	1.9%	1.8%	1.8%
	Total of Selected Activities	51.0%	51.8%	54.3%	55.2%	55.3%	54.9%	55.8%	55.7%	55.5%	53.7%	53.8%	54.0%
	Total Non-Oil	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

*2018 data are preliminary estimates

Figure (2) shows the percentage of total contribution of selected economic activities in GDP was between 54.0% and 51.0% during the study period, indicating that the contribution rate is almost constant.



*2018 are preliminary estimates

5. The Growth of the Selected Economic Activities

The data indicates that Health is the fastest growing activity, with a growth rate of 348.1% during the study period. The value added of this activity was AED 1,589 m in 2007 and reached AED 7,123 m in 2018. This indicates the importance of this activity, which is linked to the population size, society and health status of the community.

The Financial services activities grew by 127.2%; their value added was AED 25,514 m in 2007 reaching AED 57,968 m in 2018.

The Education activity grew by 55.1% during the study period. It reached AED 8,882 m in 2018 which was the highest value during the study period.

Construction is considered one of the most important non-oil activities in the economy due to its size and importance in the urban development witnessed during the study period; the value added was AED 43,768 m in 2007 and AED 81,086 m in 2018.

The Transportation and storage activity grew 35.0% over the period reaching AED 15,362 m in 2018 with a peak of AED 22,103 m in 2014.

The data indicates that the selected economic activities grew by 80.1% during the study period while nonoil GDP increased by 69.9% during the period mentioned. This indicates that the selected activities have a significant impact on non-oil GDP during the study period.

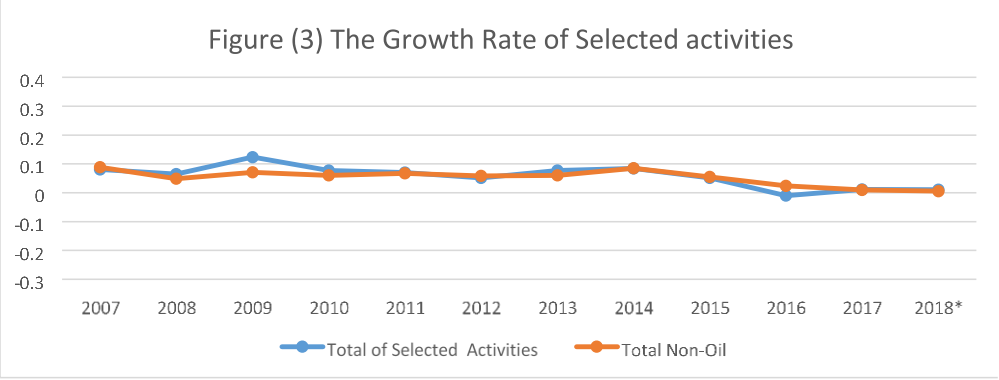
Table (2) Value Added of Selected Economic Activity in Abu Dhabi Emirate at Constant Prices 2007 - 2018*

Millions AED

ISIC4	Activities	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018*	Growth from 2007 to 2018*
C	Manufacturing	34,014	31,681	28,484	30,683	36,514	36,730	37,966	40,481	42,791	44,140	46,506	49,260	44.8%
F	Construction	43,768	53,893	71,408	78,505	79,121	78,956	77,849	80,917	84,678	82,643	81,925	81,086	85.3%
H	Transportation and storage	11,383	12,864	12,742	15,834	17,735	19,417	20,382	22,103	21,329	16,463	15,596	15,362	35.0%
K	Financial and insurance activities	25,514	23,409	24,925	23,588	24,737	30,038	40,808	48,598	53,493	55,949	57,427	57,968	127.2%
P	Education	5,728	5,731	6,139	6,388	6,740	6,836	7,051	7,346	7,560	8,284	8,805	8,882	55.1%
Q	Human health and social work activities	1,589	2,332	2,282	2,238	3,384	4,846	6,454	7,100	7,298	7,445	7,157	7,123	348.1%
	Total of Selected Activities	121,997	129,909	145,980	157,236	168,231	176,822	190,511	206,546	217,149	214,924	217,416	219,682	80.1%
	Total Non-Oil GDP	239,279	250,981	268,837	284,969	304,119	321,925	341,439	370,653	390,944	400,446	404,248	406,524	69.9%

*2018 data are preliminary estimates

Figure 3 shows that the growth rate of the selected activities is almost consistent with the non-oil GDP growth, indicating the significant impact of these activities on the non-oil GDP.



*2018 data are preliminary estimates

6. Conclusions

The selected activities are significant to the non-oil GDP of the Emirate of Abu Dhabi, which is affected by these activities and by up and downs in their growth rates. These activities are considered the main influence in the growth of the non-oil GDP and beyond to the GDP.

It is very clear that the selected activities are the main activities in the composition of GDP and that their contribution is considered essential in its composition. The non-oil GDP increases by increasing these activities and decreases when they decrease.

The data shows that there are effective economic activities such as the construction activity, real estate activities and financial services activities. These three activities show that they have a large impact on the GDP and can drive it up and down depending on their performance.

Some selected economic activities are not affected by non-oil GDP, but their importance to the economy is due to their importance to the society; this is reflected in their significant growth during the study period, for example the health activity.

Reference

1. Statistic Centre – Abu Dhabi (2018). Statistical Year Book.
2. Statistic Centre – Abu Dhabi (2013). Abu Dhabi in a half century.
3. Statistic Centre – Abu Dhabi (2018). Annual National Account publication.
4. https://www.imf.org/ar/news/searchnews#q=gdp&sort=%40imfdate%20descending
5. https://data.albankaldawli.org/indicator/NY.GDP.MKTP.KD.ZG?locations=AE
6. Gulf Cooperation Council efforts to diversify economic activity / General Secretariat of the Cooperation Council for the Arab States of the Gulf / September 2014
7. Mohamad A Alosomi, Indicators of the performance of the UAE economy in 2012 and prospects
8. Gulf Cooperation Council Falling oil prices highlight the need to diversify Gulf economies Electronic Fund Bulletin Dec 23 2014
9. Abdulla Afifi, Diplomatic Center Report: GCC countries are achieving successes in the economic diversification of their government revenues
10. Abdul Ghani, M. Efforts of the Gulf Cooperation Council (GCC) in the diversification of economic activity. Bahrain News Agency. Retrieved from http://www.bna.bh/portal/news/63294

Let the PDEs guide you to new insight into and fast inference for complex models in space and space-time

Haakon Bakka, Havard Rue
CEMSE Division, King Abdullah University of Science and Technology,
Thuwal-Jeddah, Saudi Arabia

Abstract
New non-stationary spatial models and non-separable space-time models have been developed recently in the INLA-SPDE framework (Bakka et al. (2018), Krainski et al. (2018)), and more are in development. The INLA framework uses integrated nested Laplace approximations to compute Bayesian inference quickly, while the SPDE approach uses stochastic partial differential equations (PDEs) to represent the continuous model through numerical solutions of these equations. In this presentation, we aim to explain the fundamental ideas behind this use of PDEs. PDEs have large advantages for interpretation, where we can gain insight from the physical systems these PDEs describe. PDEs also produce sparse matrices that enable efficient sampling, computation of Gaussian multivariate densities, and statistical inference.

Keywords
SPDE; INLA; filters; spatio-temporal

1. Introduction

To motivate the use of PDEs, we outline a few recent model developments to show the sucess of the framework. More details on these examples can be found in Lindgren et al. (2011) and Bakka et al. (2018).

1.1 Non-stationary models

One example we often use to illustrate the possibilities with PDEs for spatial models is

$$\alpha(s)u(s) + \nabla H(s)\nabla u(s) = \tau(s)f(s),$$

where $u(s)$ is the solution, representing the Gaussian random effect, $s \in \mathbb{R}^2$, $\alpha(s)$ is a known function, $H(s)$ is a known 2 by 2 matrix function, ∇ is the gradient/divergence operator, $\tau(s)$ is a known function, and $f(s)$ is a place-holder for the Gaussian white noise.

Different versions of this PDE has been used to solve several different problems. First, we highlight the work of Ingebrigtsen et al. (2014), where a non-stationary model is constructed to include the effect of spatial covariates. This permits, e.g., the dependency structure for rainfall data to vary according

to altitude. This model uses the differential equation above, where log(α(s)) and log(τ(s)) are linear functions of a few covariates, and $H(s)$ is the constant identity matrix. Second, Bakka et al. (2019) developed the Barrier model and an extension. The Barrier model was developed to get a robust and computationally efficient solution to the problem of internal boundaries (or holes) in the study area. This model uses the differential equation above, with α(s),τ(s) are constant, $H(s) = \phi(s)\mathrm{I}$ where I is the identity matrix. Further, this $\phi(s)$ is locally constant for the Barrier model, taking only two different values. An extension is also mentioned where $\phi(s)$ can take more than two values. Third, Fuglstad et al. (2015) model the structure of H(s) as varying slowly and continuously in space, to produce spatially varying anisotropy.

1.2 Non-separable space-time models

In space-time modeling, we study how to construct interpretable and computationally efficient space-time models with good theoretical and practical properties. One solution to this was developed by Krainski (2018), using

$$\left(\gamma_t \frac{\partial}{\partial t} - \Delta\right)^{\alpha t} (1 - \gamma\varepsilon\Delta)^{\alpha\varepsilon/2} u(s,t) = f(s,t) \qquad (1)$$

with smoothness parameters (α_t,α_ε) > 0, and scale parameters $\gamma_t, \gamma_\epsilon$. More work on this and similar models are currently underway, with very good preliminary results.

1.3 Covariance functions, filters and the "equivalent ridge regression"

The typical way to investigate space/space-time models is through the covariance function C(u(s_1,t_1),u(s_2,t_2)). Properties of the model, like stationarity and separability and "decay of dependency", are often described through the covariance function. However, interpretation of covariance or correlation is not straight forward. This is partly due to the non-uniformity of the correlation scale; e.g. the difference between correlation 0.99 and 0.97 seems larger than the difference between 0.4 and 0 when plotting samples. For space-time models, separability is equivalent to the space-time covariance being a product of spatial and of temporal covariance, but whether this is a sensible property or not in an application is difficult to ascertain.

For notational convenience, we now turn to the discrete version u of the Gaussian space/space-time model, we assume it has full rank (not intrinsic), and write it on the form $u = Fz$. This u can be a discrete representation on a grid or a mesh, the F is a matrix, and z is an iid Gaussian. This F can for example be computed as a Cholesky factorisation of the covariance matrix of u (computationally this is ill advised, but we postpone the discussion of computational efficiency to the end).

One way to view, and to construct, the model for u is by thinking of F as a matrix of covariates in a ridge regression. Assume you have data on the grid/mesh, and that $y_i = u_i + \varepsilon_i$ gives the additive model for the data y. An equivalent model is using ridge regression with

$$y_i = F_i\beta + \varepsilon_i,$$

i.e. we include a penalty term with $||\beta||_2$ in the likelihood. This can be seen easily in the Bayesian framework, where simulating a vector of iid Gaussian β's and then multiplying with F to produce $F\beta$ is equivalent to simulating from the Gaussian random effect u with covariance matrix $F^T > F$.

2. PDE based filters

A partial differential equation (PDE) is essentially an equation based on 1) a differential operator L, 2) a right hand side function f, and 3) a solution u,

$$Lu = f.$$

The operator $L(u)$ is a function of the solution u. Assuming the operator is linear (in u) we write Lu. The operator can involve non-linear functions of partial derivatives (i.e. derivatives in any spatial or temporal direction). An example of a linear differential operator is

$$L_u = a\frac{d}{dt}u + b\frac{d^2}{dx^2}u + c\frac{d^2}{dy^2}u.$$

The linear operator L can be discretised into a matrix T. Studying good discretisations is a large field of mathematics and physics, centered around analysis and numerics. Essentially, a good numerical discretisation converges to the continuous solution according to some definition of distance, when the largest square/element of the grid/mesh becomes smaller and smaller. Practically, this results in all good discretisations giving essentially the same numerical approximation as long as the approximations have a high fidelity.
After the numerical discretisation, we write

$$Tu = f,$$

where T is the matrix discretisation of the operator L, and u and f are the corresponding functions discretised (approximated) on the grid/mesh, using the same notation as for the continuous functions (giving rise to some confusion about notation).

Next, we replace f with an iid Gaussian z. This matrix T is now a filter, filtering possible u's into z's, and the inverse $F = T^{-1}$ is a filter, filtering z's into u's. Any discretised operator L gives us a materix T; which gives a Gaussian random effect $u = Fz$. The advantages of constructing the filters via a PDE is that we can interpret the filters, and equivalent ridge regressions, through the physical interpretation of the continuous PDE, and that seemingly different filters which are discretisations of the same PDE are understood to represent the same Gaussian random effect.

3. Physical interpretation of PDEs

The interpretation of PDEs is a large part of fields like physics and fluid mechanics. As an example, we interpret the equation $Lu = f$ with

$$L_u = \frac{d}{dt}u - \rho\frac{d^2}{dx^2}u - \rho\frac{d^2}{dy^2}u.$$

This is the well-known diffusion equation, explaining heat diffusion, movement of a large number of small particles, and many other phenomena. In mathematics it is closely related to Gauss's divergence theorem, and relates closely to the stochastic differential equation for Brownian motion. Without going in detail about any of these perspectives, we make the point that all of these scientific fields can be leveraged to help us understand what our model really does, and that the PDEs are the way to communicate ideas between disciplines.

We can think of our example $Lu = f$ as a surface of heat with random sources and sinks. In the case of a grid, each grid cell adds or removes a random amount of heat energy at each point in time (given by the Gaussian distribution). The strangeness of these random sources and sinks, suggest that we smooth the Gaussian noise with a spatial filter (using another operator) before we use it in the space-time model, which is what Lindgren et al. (2011) suggest. Since heat is exchanged through the plate with a speed depending on ρ in the PDE, the strength of the space-time dependence in the Gaussian random effect will depend on ρ. A larger ρ represents a faster diffusion, and the dependency (space-time correlation) will therefore be stronger.

4. Fast computations - sparsity in T

To achieve fast computational time, sparse matrices are important. In sparse matrices, most elements are zeroes that are not stored by the computer, and not accessed when computing products or solving linear systems. To simulate quickly from the random effect Tu = z we use a sparse matrix T, simulate from an iid Gaussian z and solve the linear system with a sparse matrix solver. To simulate more than once, to compute Gaussian multivariate densities, and to perform inference, we compute the Cholesky factor C of TT^{T}. The Gaussian random effect u in

$$\mathrm{C}u = z$$

is equivalent in distribution to the u in $Tu = z$. The benefit is that C is lower (or upper) triangular and allows for extremely fast computations. Additionally, with a well chosen re-ordering of the rows, the C is also very sparse. How to perform inference with Gaussian and non-Gaussian likelihoods based on sparse T are described in Rue et al. (2009) and Rue et al. (2017), where the notation is focused on the precision matrix Q = TT^{T}. The last connection we

need is how to make sparse T-matrices from PDEs. Fortunately, this has been very widely studied, and we can re-use results, algorithms and software from numerical mathematics and physics. Unfortunately, statisticians are typically not trained in numerical solutions of PDEs, but we hope this improves in the future.

References

1. Bakka, H., Rue, H., Fuglstad, G. A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., and Lindgren, F. (2018). Spatial modelling with R-INLA: A review. WIREs Computational Statistics, 10:e1443(6).
2. Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D., and Rue, H. (2019). Non-stationary gaussian models with physical barriers. Spatial Statistics.
3. Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015). Exploring a new class of non-stationary spatial Gaussian random _elds with varying local anisotropy. Statistica Sinica, pages 115{133.
4. Ingebrigtsen, R., Lindgren, F., and Steinsland, I. (2014). Spatial models with explanatory variables in the dependence structure. Spatial Statistics, 8:20{38.
5. Krainski, E. T. (2018). Statistical Analysis of Space-time Data: New Models and Applications. PhD thesis, Norwegian University of Science and Technology.
6. Krainski, E. T., G_omez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilio, D., Simpson, D., Lindgren, F., and Rue, H. (2018). Advanced Spatial Modeling with Stochastic Partial Di_erential Equations using R and INLA. CRC press. Github version www.r-inla.org/spde-book.
7. Lindgren, F., Rue, H., and Lindstrom, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4):423{498.
8. Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(2):319{392.
9. Rue, H., Riebler, A., Sorbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. Annual Review of Statistics and Its Application, 4:395{421.

An analysis of the contribution of women in Abu Dhabi

Khaddouj Abu Baker Abdulla, Thumna Al Rashedi, Saleh Almansouri
Statistics Centre Abu Dhabi, UAE

Abstract

This paper describes how Statistics Centre – Abu Dhabi (SCAD) is paying special attention to the Statistics related to women in the Emirate of Abu Dhabi. This attention is in line with the attention the government of Abu Dhabi and key decision makers pay to the role Emirati women play in all aspects of society. Women's empowerment is increasingly becoming a major objective of the government. In this regard, 28th August has been announced as "Women's Day". In response to the government's particular interest in women's empowerment, and as a part of its commitment to meet stakeholders' and customer needs, SCAD has produced an encompassing publication presenting official statistics on women and their contribution to the society, economy and workforce, using administrative data obtained through collaboration with other Abu Dhabi Government Entities (ADGEs) and data from SCAD field surveys. The publication is entitled "Al Mar'a Al Emaratya Bain Al Ams Wa Al Yaum" (Arabic for UAE Women: Past and Present). SCAD is also playing an important role providing consultancy services to the committee responsible for the design and launch of the "Women Strategy in the UAE". The Emirati woman has had many achievements at the Abu Dhabi Emirate, United Arab Emirates (UAE), regional and international levels. She has reached the highest decision-making positions, entered the university as a student and a professor, graduated from postgraduate studies. She has become minister, judge, doctor, engineer and ambassador. The high importance of human development in Abu Dhabi Emirate has had a positive influence on women by improving the quality of education, health, housing, and social welfare. Efforts made to enhance Abu Dhabi Emirate development show results in key indicators for women's progress. The literacy rate in Abu Dhabi among Emirati women has increased more than five times in less than fifty years. The number of Emirati female students enrolled in higher education has doubled in the recent decade. The gender ratio of enrolments shows that more than twice as many Emirati women as Emirati men are enrolled in higher education. The participation of Abu Dhabi women in the labour force increased over four decades from just 2% to more than 30%, and the majority of employed Emirati women are working in highly skilled and professional occupations. In Abu Dhabi Emirate official statistics tables by both gender and

citizenship show clear evidence of the progress and rapid improvement in the status of Emirati women in Abu Dhabi.

Keywords
SCAD; Females; progress; UAE; Society; Administrative Data.

1. Introduction

Demographic statistics analyses of women often focus on the strife, challenges and difficulties of women (BIIP 2012, Islah 2018 and Marinari 2019). Reports of the success and progress of women are less common. In many countries international migrants account for more than half of the population change (Lee 2011). In these countries with a large dynamic and fluctuating non-citizen population, the female citizen population may be better measure of their characteristics and trends. Statistics on the long-term more permanent resident population of female citizens is often difficult to find.

In this paper we use comprehensive statistics from SCAD to analyse how the official statistics show evidence of Emirati women's progress. The progress of Emirati women in Families, Health, Education, Work and decision making is shown. The results show a rapid change and improvement in the status of Emirati women in Abu Dhabi. We conclude that in Abu Dhabi statistics showing change over time for the Emirati women can be easily found because SCAD routinely collect and disseminate statistics by both gender and citizenship.

The SCAD is responsible for the collection, classification, storage, analysis and dissemination of official statistics covering social, demographic, economic, environmental and cultural indicators. Our vision is leadership and innovation in statistics.

2. Methodology

SCAD provides important statistical documentation of the remarkable development achievements by Emirati women in the Emirate of Abu Dhabi in various areas of society. This is accomplished by the Abu Dhabi governments support for the collection, processing, cleaning, analysis and dissemination of statistics.

This paper shows examples of how Emirati women's progress has been measured with official statistics. Statistics on Emirati women has been compiled using multiple sources of data:

- Administrative data obtained through collaboration with other ADGEs;
- Censuses of Population and Housing; and
- SCAD field surveys, such as the annual Labour Force Survey.

The UAE constitution guarantees its full rights to social freedom, equality between men and women, and the same legal right to education and

employment. Based on the firm belief in the vital role played by Emirati women in promoting comprehensive development, our leadership has paid great attention through long-term plans and programs. Women's issues have long been at the forefront of the Abu Dhabi government's agendas, which are empowering women to occupy their rightful place in society.

In this paper we use comprehensive statistics from SCAD to analyse how the official statistics show evidence of Emirati women progress.

3. Result

Emirati Women in the Population and Families

Traditionally, women have always been the backbone of family life and the social structure of the Abu Dhabi Emirate maintaining the Islamic heritage and national culture.

The study and analysis of women in the population, and in the families, is an essential step for the planning, policy-making and development programs to provide job opportunities and to provide services, such as education, health, housing and other daily necessities. In order to achieve this, it is necessary to identify the demographic facts.

The Emirati women's in Abu Dhabi Emirate has had a remarkable development. The Emirati women population has grown 55 times since 1960 until mid-year 2016, reaching 268,903 from 4,853 in 1960. The age distribution of Emirati women is characterised by a large proportion in younger ages, with 38.1% children (under the age of 15 years), while 59.6% are of working age (15-64 years). As for geographical distribution, Abu Dhabi Region attracts more than half of Emirati women (54.3%), while the Dhafra Region is only home to 4.5% of the Emirati women, and with a significant proportion living in the Al Ain Region (42.2%).

As for marriage rates, the average rate of marriage for female citizens at the emirate level was 11.6 marriages per 1,000 females in 2017. The general marriage rate was 18.7 marriages per 1000 females 15 years and over. The refined marriage rate was 42.6 marriages per 100 unmarried females aged 15 and over. The median age at the first marriage of Emirati women has gradually increased and reached 24.2 years in 2017.

Fertility among Emirati women remained high in an international comparison. The general fertility rate was 113.7 births per 1,000 women aged 15 to 49 years in 2017. The total fertility rate was 3.7 children per women aged 15 to 49 years in 2017.

The Family Development Foundation was established in 2006, and acts in cooperation with local and federal entities and specialised NGOs for the promotion of the holistic development of families, women and children, and to develop means and mechanisms to more effectively integrate public work and social welfare, and to coordinate with relevant domestic and international

organisations and experts to exchange information and expertise. The Family Development Foundation is a rather unique institution in UAE only, and it operates through an extensive network of branches to achieve its goals.

Emirati Women and their Health

As a result of the health progress witnessed by the emirate and the significant decrease in mortality, the median age of Emiratis increased from 15.5 years in 1975 to 20.3 years in 2017. This was reflected in the increase in life expectancy at birth for female citizens, which reached 80.2 years in 2017.

A vison of a healthier Abu Dhabi has been reflected in significant investment to significantly develop the healthcare capacity to meet its current and future healthcare demands. Since the end of 2010 there has been a 17% average annual growth in the number of licensed clinicians and 12% growth in the number of licensed facilities. Although the current healthcare capacity can meet the demand from Emirati women, there are currently 10 hospitals under construction, which are more than 50% complete, providing more than 1,000 additional hospital beds, ensuring that healthcare can meet future demands.

Emirati Women and Education

Efforts made to enhance Abu Dhabi Emirate development show results in key indicators for women's progress. The literacy rate in Abu Dhabi among Emirati women (10 years and above) increased from 10.7% in 1970 to 94.9% in 2016. The number of Emirati female students enrolled in higher education increased from 16,619 in 2008 to 28,821 in 2017. The gender ratio of female students in higher education reached 206.6 females per 100 male students in 2017.

Education is a cornerstone of the development process, which in essence is based on an integrated set of benefits that positively impact both individuals and society. It contributes to the creation of the creative potential that leads to the desired progress and prosperity. It is the catalyst for the creation of solutions to emerging problems, and the optimal way to define the future path and to absorb the existing elements to shape a bright future.

The Emirati education desired goal is reflected in the upgrading of educational outputs to keep pace with the highest international standards adopted. Abu Dhabi Emirate has achieved qualitative leapfrogging in the eradication of illiteracy, especially among females. It is the leading country in the proportion of educated female citizens compared to the total population. It has a pioneering educational experience that best reflects the requirements of the labour market. This contributes to the advancement of economic development.

The number of national students by gender in public and private schools shows evidence of gender equality in Abu Dhabi Emirate schools. The number of national students for the academic year 2016/17 reached 165,631 students, 49.8% of whom are females. Women attend government schools at a higher rate than in private schools. In public schools, female students accounted for 67.2% of the total female students in the Emirate, 53.0% of them in public schools.

The total number of teachers is 4,379, of whom 3,941 (90.0%) are female, with the absolute majority (99.1%) working in public schools.

As a result of intensive efforts and interest in education, literacy rates among Emirati women (10 years and over) increased steadily to reach 94.9% in 2016. In contrast, illiteracy among Emirati women continued to decline over the years to 5.1% in 2016 from 89.8% in 1970.

The advancement of women in the field of higher education has not been immune to their advancement in all areas of life. The UAE has been keen to expand higher education in all subject fields in order to provide research and employment. The number of female students enrolled in higher education increased from 16,619 in 2008 to 28,821 in 2017, Average annual growth during that period reached 8.2%. It should be noted here that Emirati females enrolled in higher education reached 206.6 females per 100 Emirati students in 2017.

Emirati women have achieved outstanding progress in education. Emirati women account for over 70 percent of university graduates.

Emirati Women Working and Participating in the Labour Force

The government's vision for Emirati women, understanding the importance for women to contribute to the development process of the Emirate, is to provide them with the necessary tools to achieve professional excellence. The priorities set by the government enables Emirati women better education, and to work in the public and private sectors. All careers should be open to Emirati women.

Efforts made to enhance Abu Dhabi Emirate development show results in key indicators for women's progress. The participation of Abu Dhabi women in the labour force increased between 1975 and 2017 from 2.2% to 30.2%. The majority of employed Emirati women are in professional occupations (52.8%).

Hard work is the cornerstone in the existence of diverse human civilisations. Considering that women constitute half of the society, it is important to expand the prospects of Emirati women's participation in the work force and in multiple areas and disciplines.

"Nothing pleased me more than seeing Emirati women take their role in society and achieve their rightful place. Nothing should stand in the way of its

progress. Women, like men, have the right to choose the highest positions, commensurate with their abilities and qualifications. "
The late Sheikh Zayed bin Sultan Al Nahyan, may Allah have mercy on him.

To have a meaningful and productive work is a human purpose and a social duty in life. Therefore, all countries seek to secure employment opportunities for all members of society, according to their qualifications and potential. Work guarantees a decent life for the inhabitants and for the society. Since women are half of society, they are men's partners in work and construction. In this sense, our wise leadership pays great attention to securing all means of women's work in order to preserve their dignity and religion without affecting the welfare and progress of their homes and children.

We note that the proportion of female UAE citizens out of the total UAE citizens' labour force in Abu Dhabi Emirate rose significantly between 1975 and 2017 from 2.2% to 35.3%. The proportion of female citizens employed among the total number of employed citizens rose from 2.2% to 30.2% between 1975 and 2017, respectively.

The majority of female citizens work as managers and professionals (52.8%), and with the largest proportion in the activities of public administration, defence and social security (42.9%). This is also reflected in that the public sector accounted for 83.6% of female employees in 2017.

Emirati Women and Power and decision-making

The traditional role of Emirati women has changed over the last two generations. The Abu Dhabi society has changed in many ways allowing Emirati women to create a bridge between the traditional and the modern, without sacrificing the heritage and culture that defines this society's identity.

Emirati women participate actively in the political sphere through representation in the Abu Dhabi Emirate government bodies and UAE Federal National Council (FNC).

Emirati women proved their worth and demonstrated their right to be present in various aspects of practical life. Creativity was a constant source of their passion for excellence. Emirati women have penetrated areas and sectors that have been limited to men for a long period of time, and have achieved tremendous achievements in this regard during a record period. The UAE is currently full of women leaders with deep knowledge. This stems from the fact that the UAE society is more open minded and more understanding of the principle of women's work. Statistics show that Emirati women and their active involvement in the labour market is a fundamental element of economic development in the UAE.

The government views the empowerment of women as crucial and wish to see women to more visibly contribute in the political arena. This is shown by the role played by women in the FNC elections of December 2006. Through

the elections, Emirati women demonstrated their ability to move into the national political arena and compete as equals with men. Female candidates ran effective campaigns with issues ranging from health and social welfare to education and jobs.

"While women made up 17.7 percent of the Electoral College, 63 of the 452 candidates who contested the polls were women. The voter turnout among women was also extremely high across the UAE. Interaction with women ahead of the elections revealed that many of them were keen to participate because they were setting precedents for women's political participation in the future.

Dr Amal Al Qubaisi was the first woman in the UAE's history to win a seat on the FNC, elected by the Abu Dhabi Electoral College. In order to ensure fair representation, the government nominated eight other women across the remaining six emirates to the 40-member FNC, which translates into a 22.5 percent share of the seats – way above the Arab world average of 9.3 percent and the world average of 17 percent. This affirmative action from government agreements specifically relating to women and children, including the Convention on the Elimination of all Forms of Discrimination Against Women (CEDAW), an international benchmark for high standards of non-discrimination. The UAE is also a signatory to the United Nations Convention on the Rights of the Child and the International Convention on the Elimination of All Forms of Racial Discrimination." (MoFNCA, 2008)

4. Discussion and Conclusion

The statistics show clear evidence of the progress of Emirati women in Abu Dhabi. The women also have more success than men in some walks of life in the Emirati society.

One of the most amazing findings is the very rapid change and improvement in the status of Emirati women in Abu Dhabi. Like the structural development of Abu Dhabi has been faster than many other parts of the world, so has the progress of Emirati women in Abu Dhabi.

In Abu Dhabi it is easy to find statistics showing change over time for the Emirati women. In Abu Dhabi statistics is routinely collected and disseminated by both gender and citizenship. In Abu Dhabi Emirate official statistics tables, data items are generally identifiable as for women and citizens of UAE. This practise of presenting tables showing if the statistics refers to citizens or non-citizens allows for easy identification of Emirati women as opposed to all women living in Abu Dhabi Emirate.

As government policy seeks further improvements in the contribution of women in Abu Dhabi, so are the ambitions of the statistical system of Abu Dhabi also seeking further improvements for collection and dissemination of statistics on Emirati women.

References

1. Bureau of International Information Programs (BIIP), United States Department of State (2012). Global Women's Issues: Women in the World Today, extended version. https://opentextbc.ca/womenintheworld/
2. Islah, J (2018). Palestinian Women's Activism: Nationalism, Secularism, Islamism
3. Lee, R.D. (2011). The outlook for population growth. In Science (pp. 569–573).
4. Mairinari, M. et. al. (2019). A Nation of Immigrants Reconsidered: US Society in an Age of Restriction, 1924-1965
5. Statistics Centre - Abu Dhabi. (2018). Emirati Women: Past and Present.
6. United Arab Emirates Ministry of State for Federal National Council Affairs (MoFNCA). (2008). Women in the United Arab Emirates: A Portrait of Progress.
7. United Nations Statistics Division. (2015). The World's Women 2015 Trends and Statistics. https://unstats.un.org/unsd/gender/worldswomen.html.

Epidemiology of acute kidney injury in critically ill patients in a South African intensive care unit

Sisa Pazi[1], Gary Sharp[1], Elizabeth van der Merwe[2,3]

[1]Department of Statistics, Nelson Mandela University, Port Elizabeth, South Africa
[2]Adult Critical Care Unit, Livingstone Hospital, Port Elizabeth, South Africa
[3]Department of Medicine, Walter Sisulu University, Umtata, South Africa

Abstract

Acute Kidney Injury is associated with substantial morbidity and mortality in Intensive Care Units. There are wide variations in the reported incidence of Acute Kidney Injury in high income country Intensive Care Units. In Sub-Saharan Africa, where there is a substantial burden of HIV, there is a paucity of data concerning Acute Kidney Injury and its incidence, aetiology and effect on mortality and functional renal recovery. The aim of this study was to prospectively investigate the occurrence and outcomes of Acute Kidney Injury in a closed multidisciplinary Intensive Care Unit in South Africa. This was the largest prospective study of Acute Kidney Injury in critically ill patients from sub-Saharan Africa, with very few patients lost to follow up.

Keywords

Logistic regression; Survival analysis; Cox proportional hazards regression

1. Introduction

Acute Kidney Injury (AKI) is commonly encountered in the Intensive Care Unit (ICU) but with a widely variable reported incidence due to non-standardisation of its definition (Koeze, Keus, Dieperink, van der Horst ICC, Zijlstra and van Meurs, 2017). Regardless of the definition used, AKI is a well-recognised independent risk factor for mortality, is associated with substantial morbidity and is a current major cause for global concern (Zeng, McMahon, Brunelli, Bates and Waikar, 2014). Furthermore, AKI requiring dialysis is now recognised as a risk factor for end stage kidney disease in the long term and is associated with poor long-term quality of life after ICU discharge (Villeneuve, Clark, Sikora, Sood, Bagshaw, 2016). The main objective of this research study was to describe the epidemiology of AKI in a closed multidisciplinary critical care unit.

2. Methodology

Study setting

The Livingstone Tertiary Hospital adult ICU and high care is a closed, multi-disciplinary 16-bed unit. It serves a region of approximately 1.6 million people

and is located in the Nelson Mandela Bay Municipality in the Eastern Cape Province of South Africa (Census, 2011). Of the 1.6 million, 1.15 million people live in an urban setting, with 12.3% living in informal dwellings and with an unemployment rate of about 36.6%. The remaining 450 000 people live in surrounding rural areas within a radius of 150-250km.

Study design

An observational prospective design was used. Cohorts were divided into those patients who developed AKI (prior to and/or after admission to the ICU) and those who did not. All patients older than 12 years admitted to the Livingstone Hospital ICU between 3 January 2017 and 3 January 2018 were included. Patients who died within 6 hours of being admitted to ICU, those who were brain-dead awaiting organ harvesting, and patients with known or presumed end stage kidney disease were excluded from the study.

Definition of acute kidney injury

Acute Kidney Injury was diagnosed when a patient's serum creatinine increased to more than 1.5 times the baseline within 7 days or more than 26.5 micromol/litre in 48 hours and was staged according to the Kidney Diseases Improving Global Outcomes (KDIGO) definition. A normal serum creatinine not older than 90 days was assumed to be the baseline where available (Siew and Matheny, 2015). The urine output criterion of less than 0.5millilitres/kilogram/hour for 6 hours was also used where possible although patients were not always weighed on admission or catheterised. The cause of AKI was determined by the treating intensivist/nephrologist and more than one cause could be assigned.

Definition of outcomes

AKI was recorded as resolved once the creatinine improved to the known or presumed baseline. If renal function had not recovered by hospital discharge, patients were followed up in the renal unit for at least 90-days following ICU discharge or until renal recovery. Patients who had not recovered their renal function by this time were deemed to have chronic kidney disease.

Ethical approval

Approval for the study (protocol number: 067/2016) was granted by the Walter Sisulu University's Postgraduate Education, Training, Research and Ethics Unit. Since we were conducting a non-experimental study that would not influence clinical decision-making or patient management, the need for study participant consent was waived.

Data collection and management

Demographic data including age, sex, race and details related to co-morbidities were collected on a data collection form. For patients known with HIV, a premorbid CD4 count was recorded where available. The Simplified Acute Physiology Score 3 (SAPS 3) was calculated using clinical and laboratory information collected within the first hour of ICU admission. The Sequential Organ Failure Assessment (SOFA) was calculated 24 hours after admission and every third day thereafter, or sooner if the patient's condition deteriorated. Vasopressor and mechanical ventilation requirements were also recorded. Cause of AKI, renal replacement modality and admission, peak and discharge creatinine were recorded for patients who were admitted to the ICU with AKI or who developed AKI whilst in the ICU. Sepsis was defined using Sepsis-3 criteria (Shankar-Hari, Phillips, Levy, 2016).

Statistical analysis

Study data were captured using the Research Electronic Data Capture (REDCap) electronic data capture tools hosted at the University of Cape Town in South Africa. REDCap is a secure, web-based application designed to support data capture for research studies, providing an interface for validated data entry, audit trails and the export/importing of data. Then, Data were exported from REDCap and analysed with RStudio, 2017. Rstudio: Integrated development environment for R (Version 3.5.1).

Continuous data were tested for normality using the Kolmogorov-Smirnov, Shapiro-Wilk, Anderson-Darling and Pearson's chi-squared tests. Normally distributed data are reported as means (standard deviation) and skewed data as medians (inter quartile range). Discrete data are presented as numbers (percentages). The student's t-test and the Mann-Whiteney U test were used to compare continuous data and the Chi-square and Fischer's exact test were used for discrete data, as appropriate. Missing outcome data (n = 12) were analysed using multiple imputation. Hazard ratios for mortality by AKI and HIV status, were calculated using the Cox proportional hazards regression model. Multivariate logistic-regression models were used to determine associations of developing AKI and dying. Purposeful variable selection method was used to select variable to be included in the models. All the variables with 0.20 $<p-value$ <0.25 (Baendel and Afifi, 1977; Constanza and Afifi, 1979; Mickey and Greenland, 1989) in the bivarite analysis, as well as those with clinical importance were included.

3. Result

Figure 1 details the number of patients admitted to the Livingstone ICU during the study period.

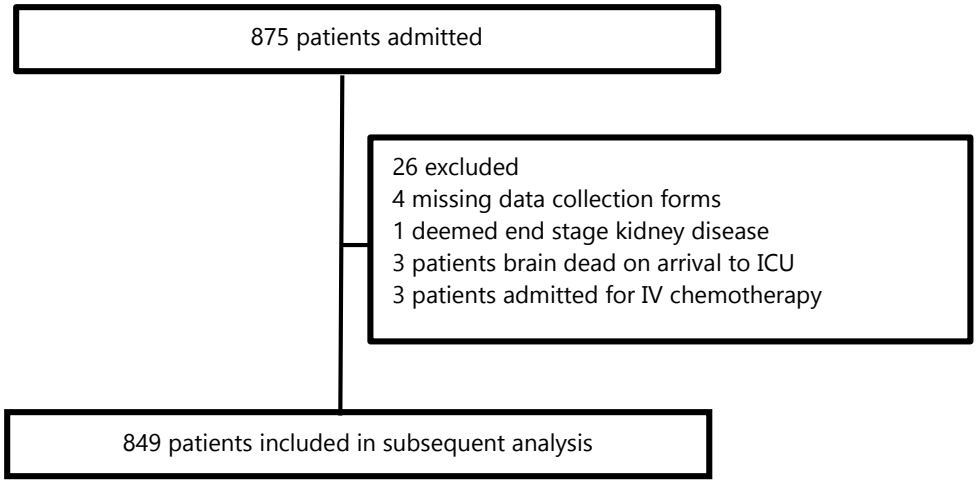


Figure 1. diagram of patients included in the analysis

Overall, 181 181 of 837 (21.6%) patients died in hospital, 12 were lost to follow up. Table 1 shows the baseline characteristics of all patients admitted to the ICU and by AKI status. Risk factors associated with AKI are presented in Table 2.

Table 1. Baseline characteristics of all patients admitted to the ICU.

Characteristic	**All patients n=849**	**AKI n=497 (58.5%)**	**No AKI n=352 (41.5%)**	**p-value**
Demographics				
Mean age, years (SD)	42.5 (16.8)	43.7 (16.8)	40.9 (16.6)	0.017
Male gender %	58.9	61.2	55.7	0.110
Race %				
Black African	58.0	60.65	54.3	0.067
Mixed Ancestry	27.4	26.0	29.6	0.248
Caucasian	13.8	12.7	15.3	0.267
Other	0.8	0.8	0.8	0.827
Co-morbidities				
Diabetes %	13.2	16.1	9.1	0.003
Mean HbA1c % (SD)	9.9 (3.2)	10.3 (3.1)	9.1 (3.3)	0.142
Hypertension %	31.6	32.2	30.7	0.641
Ischaemic heart disease %	4.2	4.6	3.7	0.506
Active Tuberculosis %	6.1	6.6	5.4	0.457
Chronic kidney disease %	7.7	6.4	9.4	0.113
Mean CKD eGFR (SD)	32 (19)	40 (16)	21 (1)	<0.001
Epilepsy %	4.8	5.0	4.6	0.746
Malignancy*%	3.3	3.6	2.8	0.53
HIV				

Known status, n = 472 (56.1%)				
Positive, n (%)	155 (32.6)	105 (35.1)	50 (28.3)	0.01
Negative, n (%)	321 (67.4)	194 (64.9)	127 (71.8)	0.382
In HIV positive cohort:				
Median CD4 count, cells/μL	318(156-493)	205 (119-391)	404 (308-513)	0.025
(IQR)	2.0 (1.98-4.04)	2.0 (1.88-3.62)	2.5 (2.0-4.64)	0.234
Median viral load, log	11.9	14.5	8.2	0.006
(IQR) Receiving HAART %				
Severity of illness				
Emergency admissions %	87.0	92.8	79.0	<0.001
Mean SAPS 3 score (SD)	48.1	54.0	39.7	<0.001
Median highest SOFA score (IQR)	4 (1 - 6)	6 (3 – 9)	2 (1 – 4)	<0.001
Sepsis and septic shock %	30.2	43.1	11.9	<0.001
Ventilated %	53.7	64.8	38.2	<0.001
median ventilator days (IQR)	3 (1 – 7)	3 (1 – 8)	2 (1 – 4)	<0.001
Required vasopressors %	25.0	38.8	5.4	<0.001
ARDS %	4.4	6.8	0.9	<0.001
Median ICU days (IQR)	3 (1 – 6)	4 (2 – 8)	2 (1 – 4)	<0.001
ICU length of stay > 7 days %	24.7	33.2	12.8	<0.001
Deceased in ICU %	13.0	20.5	2.3	<0.001

Abbreviations: AKI, Acute Kidney Injury; CI, confidence interval; SD, standard deviation; IQR, interquartile range; CKD, Chronic Kidney Disease; eGFR, Estimate Glomerular Filtration Rate; HAART, Highly Active Antiretroviral Therapy; SAPS 3, Simplified Acute Physiology Score 3; SOFA, Sequential Organ Failure Assessment; ARDS, Acute Respiratory Distress Syndrome.

Table 2: Crude and adjusted odds ratios determined by multivariate analysis for risk factors associated with AKI

Covariate	**Crude odds ratio (95% CI)**	**Adjusted odds ratio (95% CI)**	**p-value**
Age*	1.01 (1.002; 1.018)	1.01 (0.999; 1.023)	0.079
Male gender	1.22 (0.92; 1.608)	1.44 (1.008; 2.058)	0.045
Diabetes	1.88 (1.228; 2.945)	1.69 (0.982; 2.947)	0.059
Baseline chronic kidney disease	0.64 (0.382; 1.066)	0.32 (0.163; 0.634)	0.001
Receiving HAART	1.95 (1.245; 3.137)	1.64 (0.940; 2.920)	0.086
Admission SOFA score*	1.48 (1.391; 1.585)	1.33 (1.215; 1.456)	<0.001
SAPS 3 score*	1.07 (1.062; 1.088)	1.03 (1.013; 1.047)	<0.001
ICU length of stay in days*	1.12 (1.081; 1.155)	1.04 (1.007; 1.072)	0.026
Required mechanical ventilation	3.05 (2.299; 4.072)	0.56 (0.359; 0.860)	0.003
Required vasopressors	11.15 (6.954; 18.878)	2.52 (1.392; 4.700)	0.003
Developed ARDS	12.39 (3.734; 76.752)	3.2 (0.840; 21.071)	0.137
Developed sepsis or septic shock	5.5 (3.839; 8.038)	1.81 (1.119; 2.929)	0.016
HIV Infection**	1.42 (0.952; 2.15	0.6 (0.272; 1.305)	0.198

The development of AKI was associated with a higher in-hospital mortality rate of 31.8% compared to 7.23% in those without AKI (Hazards Ratio 4.07, 95% CI 2.66 to 6.21; Logrank p < 0.001 [Figure 2]). The Hazards Ratio of 4.07 means that patients with AKI were approximately 4 times more likely to die compared to those without AKI.

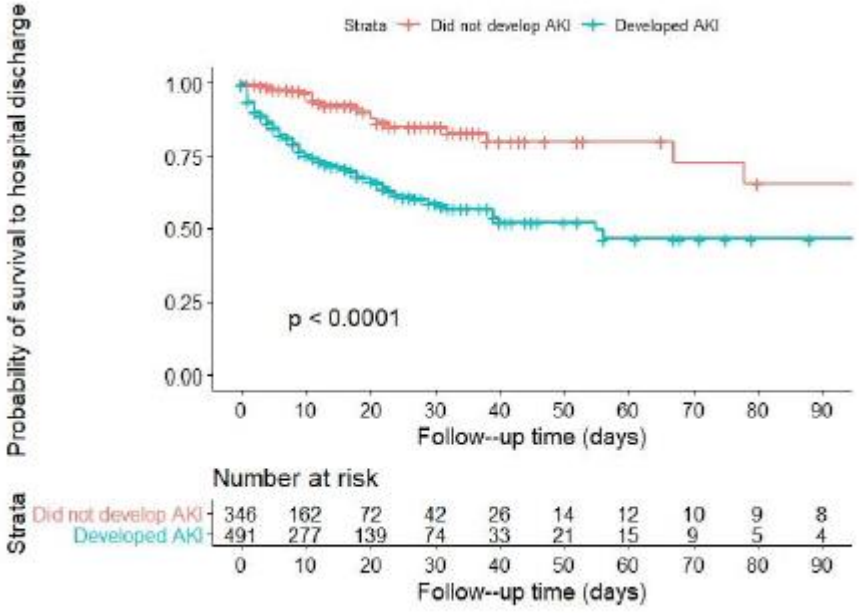


Figure 2. Kaplan Meier plot showing survival probability

4. Discussion and Conclusion

This is the largest prospective study of acute kidney injury in critically ill patients from sub-Saharan Africa, with very few patients lost to follow up. Consistent with other low to middle income countries, patients were younger than in high income country cohorts (mean age 40.6 years) with lower comorbidity rates (Hoste et al.,2015), mostly of black race, and were representative of the local community that we serve.

AKI was common in our cohort and affected nearly two thirds (58.5%) of all patients admitted to the ICU and was associated with a higher severity of illness, more sepsis, longer ICU stay, the need for vasopressors as well as male gender. While pre-existing CKD was negatively associated with AKI [OR (95%CI) = 0.32 (0.163; 0.634), p = 0.001], this reflects an admission bias against very ill patients with CKD to the unit due to a lack of resources to continue with chronic renal replacement therapy in most. This also explains the higher baseline eGFR in those patients with CKD that developed AKI.

This study had few limitations. Firstly, the incidence AKI may have been underestimated based on the urine output KDIGO criterion. While urine output has been shown to significantly improve the sensitivity for the diagnosis of AKI, we were unable to reliably utilise this criterion for diagnosis

as patients admitted to the ICU were not always weighed or catheterised. Only one patient was documented to have developed AKI based on urine output criteria alone. Secondly, 43.9% of all patients admitted to the ICU were not tested for HIV. All patients who were able to consent were encouraged to have an HIV test, however, due to the high turnover of rotating ICU doctors, testing was not consistently offered. Patients that were moribund or confused were not tested HIV without indication or consent. This was seen as a major limitation of this study in a high HIV burden setting especially since there is paucity of such data in the literature. There are several strengths to the study as well. To the best knowledge of the authors, it is the largest prospective study of AKI in critically ill patients in Sub-Saharan Africa and as such, contributes important information regarding the epidemiology of AKI in Africa. The provision of AKI outcome data at 90 days is also very valuable. The study was inclusive of every admission to the unit, the loss to follow up was low at 1.4% and the study included all stages of AKI.

References

1. Delannoy B, Floccard B, Thiolliere F, et al (2009). Six-month outcome in acute kidney injury requiring renal replacement therapy in the ICU: a multicentre prospective study. Intensive Care Med, 35(11), 1907-1915.
2. Hoste EA, Bagshaw SM, Bellomo R, Cely CM, Cruz DN, et al. (2015). Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. Intensive Care Med, 41(8), 1411-1423.
3. Koeze J, Keus F, Dieperink W, van der Horst ICC, Zijlstra JG, van Meurs M. Incidence, timing and outcome of AKI in critically ill patients varies with the definition used and the addition of urine output criteria (2017). BMC Nephrology,18(1), 70.
4. 4.Lameire NH, Bagga A, Cruz D, et al. Acute kidney injury: an increasing global concern (2013). The Lancet, 382(9887),170-179.
5. 5.Lewington AJ, Cerdá J, Mehta RL (2013). Raising awareness of acute kidney injury: a global perspective of a silent killer. Kidney international, 84(3), 457-467.
6. 6.Linder A, Fjell C, Levin A, Walley KR, Russell JA, Boyd JH (2014). Small acute increases in serum creatinine are associated with decreased long-term survival in the critically ill. Am J Respir Crit Care Med, 189(9), 1075-1081.
7. Shankar-Hari M, Phillips GS, Levy ML, et al (2016). Developing a New Definition and Assessing New Clinical Criteria for Septic Shock: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 315(8), 775-787.
8. Siew ED, Matheny ME (2015). Choice of Reference Serum Creatinine in Defining Acute Kidney Injury. Nephron, 131(2), 107-112.
9. 9.Villeneuve PM, Clark EG, Sikora L, Sood MM, Bagshaw SM (2016). Health-related quality-of-life among survivors of acute kidney injury in the intensive care unit: a systematic review. Intensive Care Med, 42 (2), 137-146.
10. Zeng X, McMahon GM, Brunelli SM, Bates DW, Waikar SS (2014). Incidence, outcomes, and comparisons across definitions of AKI in hospitalized individuals. Clin J Am Soc Nephrol,9(1), 12-20.

Monitoring unit root in sequentially observed autoregressive processes against local- to-unity hypotheses

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

Abstract

We consider unit root test under monitoring sequential sample of autoregressive (AR) processes. We examine the property of the sequential autoregressive coefficient estimator evaluated at the stopping time based on the observed Fisher information, which was firstly introduced by Lai and Siegmund (1983) for fixed accuracy estimation. We propose three kinds of test: a t-type test (T test), a test using the distribution of the stopping time (ST test), and a Bonferroni test (BON test). We consider a local-to-unity hypothesis and adapt an approximation by Orstein-Uhlenbeck process. The sequential autoregressive coefficient estimator is found to be approximating to a Dambis, Dubins, and Schwarz (DDS) Brownian motion under unit root hypothesis and to a DDS Brownian motion with drift under a local-to-unity hypothesis. The asymptotic distribution of the stopping time is characterized by a Bessel process of dimension 3/2 under unit root hypothesis and also by a Bessel process of dimension 3/2 with drift under a local-to-unity hypothesis. Henceforth, T test turns out to be possessing local asymptotic normality (LAN). We implement Monte Carlo simulations and numerical computations to examine their small sample properties.

Keywords

Local-to-unity hypothesis; Stopping time; Ornstein-Uhlenbeck process; Bessel process; Local asymptotic normality

1. Introduction

Suppose a discrete time series $\{x_n, n = 1,2,\cdots\}$ is generated from a first-order autoregressive process (AR(1)) with initial value x_0;

$$(1-\beta L)x_n = \epsilon_n$$

where $\boldsymbol{\epsilon}_n$ are independent and identically distributed with expectation 0 and finite variance $\sigma^2 > 0$ ($\epsilon_n \sim$i.i.d.$(0,\sigma^2)$). We would like to test the null hypothesis of $H_0 : \beta = 1$. As the alternative hypothesis, we can local alternatives,

$$H_0 : \beta = 1,\ H_1 : \beta = 1-\delta/\sqrt{c}$$

for $\delta > 0$ or $\delta \neq 0$ and some positive $c \to \infty$. In standard sampling theory, c is typically the sample size, then the H_1 is called Pitman local alternatives. In the present sequential framework, the sample size turns out to be random so that it is inappropriate. In our case, c is the level of the observed Fisher information

predetermined in advance. This local alternative setting is also useful to examine the statistical properties of the test; T test using the sequential autoregressive coe-cient estimator turns out to be possessing local asymptotic normality (LAN).

As in Dickey=Fuller unit root test in Dickey and Fuller (1979), we estimate the following transformed model, letting $\phi \equiv \beta - 1$, $\Delta x_n = \phi x_{n-1} + \epsilon$.

We now explain how we stop sampling and test the hypothesis using the observations. Suppose σ^2 is known for now. We propose to stop sampling at time

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

for some predetermined c > 0. This is the same stopping time considered in Lai and Siegmund (1983). If n are normally distributed, the left side of the inequality in the braces coincides with the observed Fisher information for ϕ. Therefore, we can interpret that this stopping time guarantees the estimation accuracy c. Given a sample $x_0, x_1, x_2, \cdots . x_{\tau c}$ in hand, we obtain its ordinary least squares (OLS) estimator $\widehat{\emptyset}_{\tau_c}$ of ∅, where $\widehat{\emptyset}_N = \sum_{n=1}^{N} x_{n-1}\Delta x_n / \sum_{n=1}^{N} x_{n-1}^2$ for $N \geq 1$. Lai and Siegmund (1983) show that this estimator is asymptotically normally distributed as c $\rightarrow\infty$ uniformly on [0,2]. It can be used to test if ϕ = 0. When σ^2 is unknown, a feasible stopping time can be dened as follows. Let $s_N^2 = \sum_{n=1}^{N}(\Delta x_n - \widehat{\emptyset}_N x_{n-1})^2/N$, and

$$\tau_{2c} = \inf\left\{N > 1 : \sum_{n=1}^{N} x_{n-1}^2/s_N^2 \geq c\right\}. \tag{3}$$

Then we obtain a feasible sequential OLS estimator $\widehat{\emptyset}_{\tau 2c} = \sum_{n=1}^{\tau 2c} x_{n-1}\Delta x_n / \sum_{n=1}^{\tau 2c} x_{n-1}^2$.

2. Methodology

We show the asymptotic properties of the estimator $\hat{\phi}_{\tau 2c}$ and stopping time τ_{2c} under the null of β = 1, namely

$$(1-L)x_n = \epsilon_n \; n = 1,2,..., \tag{4}$$

where $\epsilon_n \sim$ i.i.d.$(0,\sigma^2)$ with $\sigma^2 \in (0,\infty)$. Suppose $x_1, x_2, ...$ are generated by the model (4) with an initial value x_0 written as $x_0 = X(0)c^{1/4}\sigma$ for some $X(0)$ independent of ϵ_n, $n \geq 1$. By the functional central limit theorem for martingale dierences (Theorem 18.2 in Billingsley (1999)), as $c \uparrow \infty$,

$$\frac{\sum_{n=1}^{\lfloor \sqrt{c}t \rfloor} \epsilon_n}{\sigma c^{1/4}} \Rightarrow W_t \tag{5}$$

in the sense of $D[0,\infty)$ where W is a standard Brownian motion and $\Rightarrow$ indicates weak convergence in the sense of J_1 topology in D[0,∞) the space of right continuous functions with left limits. Let

$$X_c(t) = \frac{x_{\lfloor \sqrt{c}t \rfloor}}{c^{1/4}\sigma} \tag{6}$$

and

$$X(t) = X(0) + W(t). \tag{7}$$

Then, under the null,

$$X_c \Rightarrow X, \tag{8}$$

as c ↑∞ in the sense of $D[0,\infty)$.

Next, we examine the properties of the test statistics under the local alternatives

$$\left(1 - (1 - \frac{\delta}{\sqrt{c}})L\right) x_n = \epsilon_n \tag{9}$$

$n = 1,2,...$, where $c > 0$, $\epsilon_n \sim i.i.d.\ (0, \sigma^2)$ and are independent of the initial value x_0. We obtain parallel results to the null case shown in the previous section. We use a functional central limit theorem on $D[0,\infty)$.

In this setup, we test $H_0 : \delta = 0$ vs $H_1 : \delta > 0$ or $H_0 : \delta = 0$ vs $H_1 : \delta < 0$. The alternative hypothesis we consider here is stationarity in the former case, while explosive in the latter. Let $\phi^c = -\delta/\sqrt{c}$, then (9) can be rewritten as

$$\Delta x_n = \phi^c x_{n-1} + \epsilon_n. \tag{10}$$

Now we provide a Theorem with respect to AR(1) process approximating to a OU process under the alternative.

Theorem 1. *Suppose $x_1, x_2, \ldots$ are generated by the model (9);*

$$\left(1 - (1 - \frac{\delta}{\sqrt{c}})L\right) x_n = \epsilon_n \quad n = 1, 2, \ldots \tag{11}$$

where $\epsilon_n \sim$ i.i.d. $(0, \sigma^2)$ with $\sigma 2 \in (0, \infty)$. We assume that the initial value x_0 can be written as $x_0 = X(0)c^{1/4}\sigma$ for some $X(0)$ independent of ϵ_n. Let

$$X_c(t) = \frac{x_{\lfloor \sqrt{c}t \rfloor}}{c^{1/4}\sigma} \tag{12}$$

and X be the OU process;

$$X(t) = X(0) - \delta \int_0^t X(s)ds + W(t), \tag{13}$$

where $W(t)$ is the standard Brownian motion in (5). Then,

$$X_c \Rightarrow X,$$

as $c \uparrow \infty$ in the sense of $D[0,\infty)$.

Remark 2. X of (8) coincides with X in (13) with $\delta = 0$.

Let W be the Brownian motion in (5) and X be in (7) or (13). Define the martingale $M_t = \int_0^t X(u)dW_u$ and its quadratic variation $\langle M \rangle_t = \int_0^t X^2(u)du$. Put

$$U_s = \langle M \rangle_s^{-1} = \inf \left\{ t \geq 0 : \int_0^t X^2(u)du = s \right\}. \quad (14)$$

Then, by DDS Theorem (Theorem 1.6 in Revuz and Yor (1999) pp181),

$$B_s = M_{U_s} \quad (15)$$

is a Brownian motion with respect to the filtration $\mathcal{G}_s = \mathcal{F}_{U_s}$ and this is called a DDS Brownian motion.

3. Result

The following theorem is obtained in Hitomi, Nagai, Nishiyama, and Tao (2018), but also could be proved by setting $\delta = 0$ in Theorem 4.

Theorem 3. Suppose x_n is generated by the model (4) with an initial value x_0 independent of ϵ_n, $n \geq 1$. Then, if we put $\rho_s = X^2_{U_s}/2$, ρ_t is a 3/2 dimensional Bessel process;

$$\rho_t = \rho_0 + B_t + \int_0^t \frac{1}{4\rho_s} ds. \quad (16)$$

where $\rho_0^\delta = X^2(0)/2$. The asymptotic behavior of the stopping times τ_{ic} (*(i* = 1,2) in *(2)* and in *(3)*) and the sequential estimators $\hat{\phi}\tau_{ic}$ is given as follows: as $c \uparrow \infty$,

$$\left(\sqrt{c}\hat{\phi}_{\tau_{ic}}, \frac{\tau_{ic}}{\sqrt{c}} \right) \Rightarrow \left(\int_0^{U_1} X_u dW_u, U_1 \right) = \left(B_1, \int_0^1 \frac{1}{2\rho_s} ds \right) \quad (i = 1, 2). \quad (17)$$

Here is the main theorem of this section.

Theorem 4. Suppose x_n is generated by the model (9) with an initial value x_0 independent of ϵ_n, $n \geq 1$. Then letting $\rho_s^\delta = X^2(U_s)/2$, ρ_t^δ is a $3/2$ dimensional Bessel process with drift $-\delta$;

$$\rho_t^\delta = \rho_0^\delta + B_t + \int_0^t \left(\frac{1}{4\rho_s^\delta} - \delta \right) ds \quad (18)$$

where $\rho_0^\delta = X^2(0)/2$. The asymptotic behavior of the stopping time τ_{ic} (i = 1,2) in (2) and in (3)) and the sequential estimators $\hat{\phi}\tau_{ic}$ is given as follows: $\tau_{ic} \to_p \infty$, $\hat{\phi}\, \tau_{ic} - \phi \to_p 0, s^2_{\tau 2c} \to_p \sigma^2$,

$$\left(\sqrt{c}\hat{\phi}_{\tau_{ic}}, \frac{\tau_{ic}}{\sqrt{c}} \right) \Rightarrow \left(-\delta + \int_0^{U_1} X(u)dW(u), U_1 \right) = \left(-\delta + B_1, \int_0^1 \frac{1}{2\rho_s^\delta} ds \right) \quad (19)$$

where W(t) is the standard Brownian motion in (5).

The following theorem gives the asymptotic distribution of $s^2_{\tau 2c}$.

Theorem 5. Under the same assumption of Theorem 3 or Theorem 4, we have $s^2_{\tau 2c} \to_p \sigma^2$. Suppose $\omega^2 = E[(\epsilon_1^2 - \sigma^2)^2] < \infty$. Then as $c \uparrow \infty$,

$$\frac{c^{1/4}}{\omega}\left(s^2_{\tau_{2c}}-\sigma^2\right)\Rightarrow\frac{1}{U_1}W'(U_1) \tag{20}$$

where $W\prime$ is a Brownian motion satisfying $E(W\prime W_1) = E(\in_1^3)/\omega\sigma$ with W being the Brownian motion in (13). Furthermore, if $E(\in_1^3) = 0$, then W and $W\prime$ are independent and

$$\frac{\tau_{2c}^{1/2}}{\omega}\left(s^2_{\tau_{2c}}-\sigma^2\right)\Rightarrow N(0,1). \tag{21}$$

Corollary 6. Under the same assumptions as in Theorem 4 with $\delta = 0$ and $X(0)= 0$, we get the asymptotic expectation

$$E(\tau_{ic}/\sqrt{c})\to 2E(\rho_1)=2\frac{\sqrt{2}\Gamma(5/4)}{\Gamma(3/4)}=2.0921.$$

Table 1: Size of tests when alternative hypothesis is stationary

$\beta = 1$

c = 2500	rejection rate	mean of $\tau_{2c}/\sqrt{c}$	c = 10000	rejection rate	mean of $\tau_{2c}/\sqrt{c}$
T	0.0465	2.0693040	T	0.0488	2.0932680
ST	0.0438	2.0399180	ST	0.0463	2.0628650
BON	0.0296	2.0571760	BON	0.0300	2.0801700

Table 2: Power of tests when alternative hypothesis is stationary

$\beta = 0.99$

c = 2500	rejection rate	mean of $\tau_{2c}/\sqrt{c}$	c = 10000	rejection rate	mean of $\tau_{2c}/\sqrt{c}$
T	0.1288	2.5729220	T	0.2630	3.1839720
ST	0.1129	2.4831220	ST	0.2357	2.9486160
BON	0.0829	2.5293460	BON	0.1909	3.0518460

$\beta = 0.95$

c = 2500	rejection rate	mean of $\tau_{2c}/\sqrt{c}$	c = 10000	rejection rate	mean of $\tau_{2c}/\sqrt{c}$
T	0.8019	5.4098460	T	0.9996	10.026283
ST	0.7457	3.9848560	ST	0.9989	4.239397
BON	0.7213	4.3702700	BON	0.9989	4.788537

4. Discussion and Conclusion

Now we propose three testing procedures of $\beta = 1$ or $\phi = 0$, namely, t test (T test), stopping time test (ST test), and Bonferroni test (BON test). We explain the procedures only for one sided test against the alternatives of $\beta < 1(\phi < 0)$ or $\beta > 1(\phi > 0)$.

First, T test uses the asymptotic normality of $\hat{\phi}_{\tau 2c}$. Since the null hypothesis is $\phi = 0$ and its asymptotic variance equals to unity, the T simply looks at $\sqrt{c}\hat{\phi}_{\tau 2c}$. Let z_α be the α quantile of the standard normal distribution. We reject the null when $\sqrt{c}\hat{\phi}_{\tau 2c} < z_\alpha$ for left sided test with the alternative of $\beta < 1$, and when $\sqrt{c}\hat{\phi}_{\tau 2c} > z_{1-\alpha}$ for the right sided test.

Second, ST test is constructed by using the distribution of U_1 (14) which is the limit of $\tau_{2c}/\sqrt{c}$ under the null. Its distribution is not standard, but we can easily obtain its quantiles by numerical computation. Under the stationary alternatives, the stopping time tends to wait longer, while if the process is explosive or $\beta > 1$, the sequential procedure stops earlier than unit root case. Therefore, letting $u\alpha$ be the α quantile of U_1, we reject the null if $\tau_{2c}/\sqrt{c} > u_{1-\alpha}$ against the alternative of $\beta < 1$. If $\tau_{2c}/\sqrt{c} < u_\alpha$ we reject the null against $\beta > 1$.

Third, we can combine both statistics by BON Bonferroni test as follows. When we want to test the existence of unit root against the alternative of stationarity, we reject the null when $\sqrt{c}\hat{\phi}_{\tau 2c} < z_{\alpha/2}$ or $\tau_{2c}/\sqrt{c} > u_{1-\alpha/2}$. Obviously, it will be a conservative test as is always the case with Bonferroni tests.

Considering $\mathrm{AR}(1)$ process, we obtain the asymptotic distribution of the OLS estimator of the $\mathrm{AR}(1)$ parameter and the Fisher information based stopping time under a sequential sampling both under the unit root process and near unit root process. The t statistic is asymptotically normally-distributed and the stopping time is characterized by Bessel processes. We

Table 3: Power of tests when alternative hypothesis is explosive

$\beta = 1.01$

c = 2500	rejection rate	mean of $\tau_{2c}/\sqrt{c}$	c = 10000	rejection rate	mean of $\tau_{2c}/\sqrt{c}$
T	0.1196	1.7052720	T	0.2624	1.4189670
ST	0.0823	1.7052720	ST	0.1550	1.4189670
Bon	0.0978	1.7052720	Bon	0.2060	1.4189670

$\beta = 1.05$

c = 2500	rejection rate	mean of $\tau_{2c}/\sqrt{c}$	c = 10000	rejection rate	mean of $\tau_{2c}/\sqrt{c}$
T	0.8010	0.94602	T	0.9996	0.610923
ST	0.4100	0.94602	ST	0.8233	0.610923
Bon	0.7300	0.94602	Bon	0.9990	0.610923

employ functional central limit theorem to prove the results which enables us to analyze the asymptotic properties in the case of local alternatives of near unit root processes. Based on the results, we propose three kinds of unit root tests using the t statistic, the stopping time, and the both (Bonferroni). When the alternative is a stationary process, we show that the stopping time can be a useful test statistic especially when the sampling cost is large. If the alternative is an explosive process, t test is shown to perform the best. The Bonferrroni approach is one way of using both of t value and stopping time, but it is likely to be able to construct a better test exploiting both information.

References

1. Billingsley, P. (1999) *Convergence of Probability Measures*, 2nd Ed., John-Wiley and Sons.
2. Dickey, D. A. and W. A. Fuller (1979). _Distribution of the estimates for autoregressive time series with a unit root," *J. Amer. Statist. Assoc*. 74, no. 366, part 1, 427-431.
3. Nagai, K., Nishiyama, Y. and Hitomi, K. (2018). Sequential test for unit root in AR(1) model, *Kyoto Institute of Economic Research Discussion Paper*, No. 1003, Kyoto University
4. Lai, T.L. and D. Siegmund (1983). Fixed accuracy estimation of an autoregressive parameter, *Annals of Statistics* 11, 478-485.
5. Revuz, D. and M. Yor (1999). *Continuous Martingale and Brownian Motion*, 3rd ed. Springer-Verlag: New York.

An investigation into parametric and non-parametric modelling of LGD to estimate extreme percentiles of the loss distribution with respect to defaulted loans

Janette Larney[1]; Gerrit Grobler[2]
[1] Centre for BMI. North-West University, South Africa
[2] School for Mathematical and Statistical Sciences, North-West University, South Africa

Abstract
Very little research has been devoted to the variability of Loss Given Default (LGD), or in other words, the "unexpected" part of recovery risk. Our aim is to investigate models for the estimation of economic capital for the risk of recoveries in respect of non-performing loans being different from expected. As we are specifically interested in the extreme percentiles of the loss distribution, we first set out to model the unique distribution of LGD, which includes its bimodal nature, and having most of its density near zero and one. We model LGD both parametrically, using a mixture of beta distributions as well as non-parametrically, where we use the reflection kernel estimator to effectively reduce boundary bias and improve the accuracy of the estimator. Both methods show promising results and the mixture of beta distributions offer improvements on other widely used competitor distributions. We also attempt to model specifically the tail dependence of individual exposures by using the Gumbel copula and t-copula. Our results indicate, non-surprisingly, that both the Gumbel and t-copula improve on the Gaussian copula in capturing tail dependence and offer a way to accurately reflect the dependence between individual obligors.

Keywords
Boundary Bias; Copula; Kernel Density; Recovery Risk

1. Introduction

One of the key parameters used in determining a bank's regulatory capital under the Basel accord is Loss Given Default, or LGD, which is defined as the economic loss incurred in the event of default, including workout costs and material discount effects. Research on LGD is relatively scant when compared to the much studied Probability of Default (PD), and the predominant bulk of publications on LGD is aimed at modelling its expected value. This makes sense, as not only is the expected value needed to determine regulatory capital amounts (the LGD component enters the risk weight formula of Basel II only by its expected value), but the LGD parameter is an important component in determining expected credit losses for impairment, or reserving calculations.

In the Basel II framework, the expected loss of a portfolio of loans is assumed to comprise three components. They are the PD, which is the proportion of borrowers that is expected to default within a given time frame, the exposure at default (EAD), and the LGD. The expected loss (EL) of a portfolio of loans can then be expressed as:

$$EL = PD \times LGD \times EAD.$$

The Basel II regulatory capital formula is based on the amount by which the unexpected loss of a portfolio exceeds its EL. Under the Asymptotic Single Risk Factor (ASRF), a structural model adapted from Merton's (1974) single asset model, the capital requirement (K) in respect of a portfolio of loan exposures can be expressed as a function of the three parameters on the right hand side of (1) as well as an asset correlation parameter, ρ, and a smoothed maturity adjustment,($\mathcal{PD}$), which a is a function of maturity (M) and PD. The maturity adjustment maps a specific rating grade and maturity to the Value-at-Risk measures for a range of maturities and rating grades, where the rating grades represent the probabilities of default (Basel Committee on Banking Supervision, 2005). The model sets the standard maturity to 2.5 years:

$$K = \left(LGD \times \Phi\left[\sqrt{1/(1-\rho)} \times \Phi^{-1}(PD) + \sqrt{1/(1-\rho)} \times \Phi^{-1}(0.999) \right] - PD \times LGD \right) \times \frac{1 + (M - 2.5) \times b(PD)}{1 - 1.5 \times b(PD)}.$$

When the ASRF model is applied to defaulted loans (i.e. where $PD=1$), a zero capital charge is obtained for the case where downturn LGD is used.

Very little research has been devoted to the variability of LGD's, or in other words, the unexpected part of recovery risk. This despite a statement by the Basel Committee on Banking Supervision (2005) that a capital charge for defaulted assets is desirable "in order to cover systematic uncertainty in realised recovery rates for these exposures". Also, empirical evidence from the banking sector shows that recovery losses vary materially from what is expected due to the long-term exposure to recovery risk (Gupton et al., 2000). Our aim is therefore to investigate models for the estimation of capital against "unexpected recovery risk", i.e. the risk of recoveries in respect of non-performing loans being different from that expected. In this paper we specifically focus on modelling the unique distribution of LGD and also the dependence between individual exposures within a portfolio of defaulted loans.

In his paper Tasche (2004) proposes a single risk factor model, based on a parametric beta distribution to determine a capital charge in respect of a portfolio of defaulted loans. The model incorporates the Loss Given Default (LGD) and Loss Given Default volatility (LGDV) as parameters. Tasche's model remains a single risk factor model similar to the ASRF model in that

idiosyncratic risks related to individual exposures are assumed to diversify away and a single "risk" factor (representing the state of the economy) is used to capture the systematic risk of the portfolio.

Being bounded between 0 and 1, the beta distribution is a popular choice for modelling LGD and is also used in commercial models for the recovery rate, e.g. in CreditMetrics (Gupton et al., 1997). However, empirical evidence suggests that LGD values have an unusual distribution in that the distribution tends to be bimodal with modes close to the boundary values (Qi and Zhao, 2011). The standard beta distribution has a single mode and is unable to cope with the probability masses at 0 and 1 of the empirical distribution. As capital in respect of unexpected recovery risk is our ultimate goal, we are specifically interested in the extreme percentiles of the loss distribution and the drawbacks of the standard beta distribution are therefore considerable.

Likewise, although hailed for its simplicity, the use of the Gaussian copula to represent the dependence between individual obligors for the determination of 1-in-1,000 year losses[15] in the ASRF model is worth our consideration. Especially so because the Gaussian copula has a coefficient of tail dependence of zero.

2. Methodology

As mentioned above, the standard beta distribution presents several limitations in capturing the unique distribution of LGD. Several authors have proposed alternative distributions and modelling approaches, including fractional response regression (Papke and Wooldrige, 1996), inflated beta distributions (Ospina and Ferrari, 2012) and a parametric multi-step approach (Tanoue et al., 2017) to name but a few.

Parametric models have the significant advantage of their interpretability, but assuming a specific distribution for LGD does compromise on prediction power when compared to non-parametric methods, as highlighted by Hurlin et al. (2018). In this paper we not only investigate an alternative parametric distribution for modelling the distribution of LGD, but also a fully nonparametric method.

Parametric modelling

As mentioned before, the beta distribution is commonly used to model LGD and this is also the choice of Tasche (2004) in determining a capital charge in the case of correlated LGD's. The beta distribution is related to the gamma distribution. It is a continuous probability distribution with the probability density function (pdf) defined on the interval (0, 1). The pdf of the beta distribution is given by:

[15] A 1-in-1,000 year events is equivalent to the 99.9% percentile of the loss distribution (reduced by the amount reserved for, i.e. the expected loss)

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1},$$

where $a, b > 0$ and $0 < x < 1$.

The expected value and variance of a beta distributed variable X, is given by

$$E(X) = \frac{a}{a+b}$$

and

$$var(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

By matching the first two central moments of the beta distribution with the expected value and variance of the loss distribution a parametrisation in terms of LGD and LGDV is then given by

$$a = \frac{LGD}{LGDV}(LGD(1-LGD) - LGDV)$$

and

$$b = \frac{1-LGD}{LGDV}(LGD(1-LGD) - LGDV).$$

Therefore, the shape parameters aa and bb can be estimated by estimating the expected value and variance of the loss distribution. Tasche (2004) also proposes, for the purpose of prudence, using a fixed proportion vv of the maximally possible variance of the beta distribution, adopting a value of 0.25 for vv in his numerical examples.

Typically, LGD has most of its density close to zero and one. This makes the beta distribution only suitable for a restricted choice of parameter values. Furthermore, the shape of the beta distribution is very sensitive to the choice of parameter values. In Figure 1 we illustrate this by comparing two beta distributions with different values of LGDV (10% and 6%) at a fixed LGD value of 60%. Although LGDV does not differ greatly, the shape of the two beta distributions does.

Several authors, including Calabrese and Zenga (2010) and Chen (1999), have shown that the beta density function is unable to accurately represent the distribution of LGD, even when the boundary observations are removed from the data. To replicate the typical bimodality of loss data, we decided to rather use a mixture of two beta distributions. The density function for a mixture of two beta distributions is simply the weighted sum of two beta density functions,

$$f(x) = \pi \frac{\Gamma(a_1+b_1)}{\Gamma(a_1)\Gamma(b_1)} x^{a_1-1}(1-x)^{b_1-1} + (1-\pi)\frac{\Gamma(a_2+b_2)}{\Gamma(a_2)\Gamma(b_2)} x^{a_2-1}(1-x)^{b_2-1},$$

where $0 < \pi < 1$, a_1, a_2, b_1, $b_2 > 0$ and $0 < x < 1$. An example of a mixed beta density function is shown in Figure 1 to illustrate the possibility to obtain a bimodal density.

This mixture distribution has five unknown parameters which we estimated using a method proposed by Schröder and Rahmann (2017). The implemented algorithm combines latent variables with the method of moments. This method was used due to the popular maximum likelihood method being inappropriate for a mixture of beta distributions when observations take on values of zero or one, which is invariably the case with recovery or LGD data.

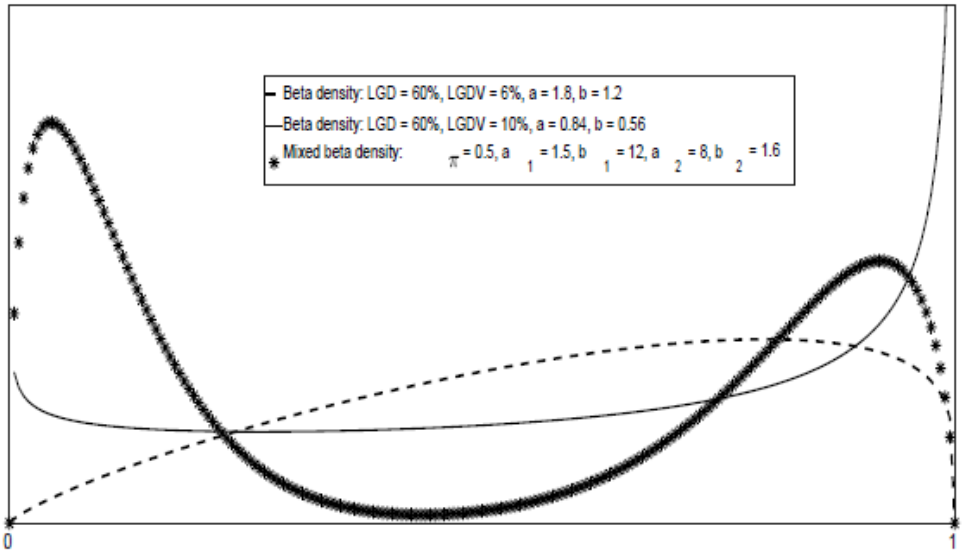


Figure 1: The density functions of two beta distributions with an LGD value of 60% and LGDV values of 10% and 6% respectively as well as a bimodal density function from a mixture of two beta distributions.

Non-parametric modelling

A natural and well-studied estimator for the density function f, based on data $X_1, X_2,\ldots, X_n$, is the kernel-density estimator given by:

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right), \qquad (2)$$

where h is the bandwith and kk is the kernel function, satisfying some regularity conditions. For an in-depth discussion on Kernel density estimators the interested reader is referred to Wand and Jones (1994).

In our set-up, because the LGD is restricted to the interval (0,1) the estimator in (2) will suffer from boundary bias, which will have a negative effect on the accuracy of the estimator. We are specifically interested in reducing boundary bias around the upper boundary as it is well reported that LGD observations often exceed one due to workout costs, but this is rarely included in datasets, as reported by Miller and Töws (2017). Calabrese and Zenga (2010)

consider a mixture of beta kernel estimators to model the LGD density of a large dataset of defaulted Italian loans. This beta kernel estimator is given by:

$$\hat{f}_B(x,h) = \frac{\sum_{i=1}^{n} K_{(\frac{x}{h}+1,\frac{1-x}{h}+1)}(X_i)}{n}$$

where $K_{(\frac{x}{h}+1,\frac{1-x}{h}+1)}$ is the beta probability density function

$$K_{(\frac{x}{h}+1,\frac{1-x}{h}+1)}(X_i) = \frac{1}{B(\frac{x}{h}+1,\frac{1-x}{h}+1)} X_i^{\frac{x}{h}}(1-X_i)^{\frac{1-x}{h}} \qquad i = 1,2,\ldots,n$$

with K denoting the kernel function and h denoting the smoothing parameter, such that $h \to 0$ as $n \to \infty$.

Now, rather than using the beta kernel estimator to alleviate the problem of boundary bias, we investigated a modification of the traditional kernel density estimator, namely the reflection kernel estimator. In its simplest form this estimator is given by:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left\{\left(\frac{x-X_i}{h}\right) + k\left(\frac{x+X_i}{h}\right)\right\},$$

and accounts for boundary bias by putting some data outside the interval (0,1). In our Monte Carlo simulation we compare the performance of the beta kernel estimator and the reflection kernel estimator under various scenarios, including the scenario of LGD exceeding one due to workout costs.

Modelling obligor dependence

Arguably the most crucial component of estimating the extreme percentiles of the loss distribution of a portfolio of defaulted loans is embedded in the modelling of the dependence between individual exposures. We have commented on the shortcomings of the Gaussian copula earlier, and therefore identified the need to investigate alternative, less familiar, dependence structures.

Here we considered the t-copula as well as the Gumbel copula, both of which are able to model tail dependence. We also considered different ways to estimate the copula parameters, including maximum likelihood estimation and the methods of moments (based on Spearman's rho).

3. Result

Based on an empirical study we concluded the following:

- The mixture of beta distributions is a viable alternative to use to accurately model the LGD. It accurately captures the "bumps" in the density of the LGD at its boundary values. The parameter estimates for the mixture of beta distributions are also more stable than that of the general beta distribution.

- The non-parametric boundary bias reduction kernel estimators showed great promise as a way to model the density of the LGD without making any distributional assumptions. We are specifically interested in reducing boundary bias around the upper boundary, and in this regard the reflection kernel estimator is effective in correcting asymmetrical boundary bias.
- In preliminary results both the Gumbel as well as the t-copula outperformed the Gaussian copula. This is not surprising, since it is well known that there exist tail dependence between individual obligors.

4. Discussion and Conclusion

In this paper we highlighted the shortcomings of the standard beta distribution in modelling the probability masses at 0 and 1 of the empirical distribution of the LGD. As capital in respect of unexpected recovery risk is our ultimate goal, we are specifically interested in the extreme percentiles of the loss distribution and these shortcomings therefore deserve specific consideration. Both the parametric and non-parametric methods applied in this paper showed great promise to improve on the conventional modelling of LGD, specifically in improving modelling the tails of the recovery loss distribution, which is crucial for determining capital in respect of unexpected recovery risk. The Gumbel and t-copulas also show potential to more accurately capture tail dependence than the popular Gaussian copula.

This study is still in its preliminary phase and future work will focus on using these alternatives in estimating economic capital in respect of non-performing loans.

References

1. Basel Committee on Banking Supervision (2005). An Explanatory Note on the Basel II IRB Risk Weight Functions. Consultation Paper, July.
2. Calabrese R. & Zenga M. (2010). Bank Loan Recovery Rates: Measuring and Nonparametric Density Estimation. Journal of Banking and Finance, 34(5), 903-911.
3. Chen S. X. (1999). Beta kernel estimators for density functions. Computational Statistics & Data Analysis 31, 131-145.
4. Gupton, G., Finger, C. & Bhatia, M. (1997). CreditMetrics. Technical report, J.P. Morgan.
5. Gupton, G., Gates, D. & Carty, L. (2000). Bank loan loss given default. Special comment, Moody's Investor Service.
6. Hurlin, C., Leymarie, J., & Patin, A. (2018). Loss functions for loss given default model comparison. European Journal of Operational Research, 268 (1), 348–360.

7. Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. Journal of Finance 29, 449–470.
8. Miller, P. & Töws, E. (2018). Loss given default adjusted workout processes for leases. Journal of Banking & Finance, Elsevier, 91(C), 189-201.
9. Ospina, R. and Ferrari, S. (2012). A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis, 56(6), 1609–1623.
10. Papke L.E. and Wooldridge J.M. (1996). Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. Journal of Applied Econometrics, 11(6), 619-632.
11. Qi, M. and X. Zhao (2011). A comparison of methods to model loss given default, Journal of Banking and Finance, 35, 2842-2855.
12. Schröder, C. and Rahmann, S. (2017). A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification, Algorithms for Molecular Biology, 12(1), 1-12.
13. Tanoue, Y., Kawada, A. and Yamashita, S. (2017). Forecasting Loss Given Default of Bank Loans with Multi-Stage Model. International Journal of Forecasting, 33(2), 513-522.
14. Tasche, D. (2004). The single risk factor approach to capital charges in case of correlated loss given default rates. Technical report, Deutsche Bundesbank.
15. Wand, M.P. and Jones, M.C. (1994). Kernel Smoothing. New York. Chapman & Hall.

Experimental statistics: A hub for data innovation in Official Statistics?

Daniel Kilchmann
Swiss Federal Staistical Office (SFSO), Statistical Methods Unit

Abstract
Innovation has always been of key interest in Official Statistics and became more and more important during the last decade due to increased political and financial pressure. However, the implementation of innovation often takes a long time in Official Statistics due to the need of a thorough validation process guaranteeing the publication of reliable and valuable results based on sound methodology. Similar to some other National Statistical Institutes and Eurostat, the SFSO decided to promote data innovation through a website dedicated to experimental statistics. The microsite www.experimental.bfs.admin.ch has been launched during the Swiss Statistics Meeting in August 2018 and presents small area estimation for the activity rate as well as SFSO's five data innovation projects in the fields of small area estimation, machine learning and deep learning. This contribution gives a short overview of the content of SFSO's microsite with its opportunities and provides an insight to the challenges of experimental statistics with respect to data innovation.

Keywords
Experimental statistics; data innovation; small area estimation; machine learning; deep learning.

1. Introduction

Innovation has always been of key interest in Official Statistics leading to moving from census' based statistics to survey sampling and including administrative data and other data sources. These changes of the data sources went hand in hand with the adaptation of the statistical production processes, innovative methodology for the treatment of missing units and values and inconsistencies and with efficiency gains due e.g. to calibration. The political and financial pressure to produce more with less during the last decade lead to re-think the implementation procedure of innovation activities. On the one hand, there are activities which prove to be more effective compared to the procedures they should replace like favouring respondents to fill in web questionnaires instead of paper questionnaires (with paper questionnaires as fall back mode). On the other hand, innovation activities in the domain of statistical methodology usually take more time for validation and therefore for

implementation. One way to shorten this validation process is to publish the corresponding results and their underlying methodology on dedicated websites, e.g. experimental statistics website. This contribution gives a short overview of the characteristics of publishing experimental statistics in paragraph 2, followed by a short description of SFSO's experimental statistics in paragraphs 3 and 4 before concluding with a general discussion in paragraph 5.

2. Experimental statistics

The SFSO characterises experimental statistics as "… produced using new methods and/or new data sources and are therefore in line with the SFSO's data innovation strategy and the Confederation's multi-annual programme for federal statistics" and are subject to "… better meet users' needs in terms of efficiency, quality and speed.". Furthermore, "…these statistics still have the potential to evolve, especially regarding their methodology, which is still being assessed". The following examples show that experimental statistics from other institutions have in common that they have not the maturity of official statistics, like mentioned above for the SFSO, and therefore they should be used with great care:

- Eurostat [3] states that "… these statistics have not reached full maturity in terms of harmonisation, coverage or methodology …".
- ISTAT [4] describes them as "… they do not respect all the steps necessary to test new methodologies, to transform them into technological and organisational solutions, to verify if quality requirements and harmonisation rules are fulfilled.".
- ONS [9] defines experimental statistics with "These are series of statistics that are in the testing phase and not yet fully developed.".

At the time of writing this contribution, Eurostat published 11 experimental statistics on its website and ISTAT 5. As ONS has not a dedicated website for the results of experimental statistics the author renounces of counting them, however, there were a lot of publication branded 'experimental statistics' available on the ONS website. The SFSO did not only choose to publish finalised experimental statistics but also projects leading potentially to experimental statistics or activities being part of the production process which is redesigned by the use of alternative methods. There is no guarantee that these projects will ever result in publishing experimental statistics for the use of the clients or even official statistics.

Hence, all these institutions invite users to give their feedback on the published products. With respect to this, the SFSO site gives descriptions of (pilot) projects currently being developed and gives the users and partners the opportunity to get involved "… at an early stage for both the development and consolidation of projects". Therefore, one of the main goals of publishing

experimental statistics is to speed up the validation procedure of these products with the aim to assess the potential of moving them to the official statistics as quick as possible. ONS published its procedure with respect to removing the experimental status [9] underpinning the fact that experimental statistics do not have the same status as official statistics.

The content of SFSO's experimental statistics website is discussed in the following to provide a better insight of SFSO's understanding of these products.

3. Small area estimation (communes) of economic activity rate in the structural survey [20]

The structural population survey provides important information on the population, including information about work. By means of a sample survey of at least 200 000 people each year in Switzerland, it is possible to make reliable estimates of the economic activity rate of groups of 15'000 inhabitants.

By combining the surveys of several years (pooling over three or five years), it is possible to reduce this limit to groups of 5000 respectively 3000 people. This can be achieved using standard methods.

It should be noted that in 2014, Switzerland had almost 2300 communes. Of these, 80% have fewer than 5000 inhabitants and 70% fewer than 3000.

The question addressed here is whether it is possible to reduce the limits mentioned above even further to produce reliable estimates by means of new methods that are based on statistical models, which use information that is available for the whole population.

3.1. Objectives

The whole purpose of Small Area Estimation (SAE) is to challenge the limits imposed by standard methods by using techniques based on modelling which make use of complementary information on the whole population.

Professor I. Molina from the University Carlos III of Madrid, co-author with J.N.K. Rao of the book "Small Area Estimation" [10], was given the madate to examine the possibility of obtaining reliable estimates of the economic activity rate at commune level. This study, based on the 2012 structural survey and on OASI (Old-age and survivors' insurance) data from 2011, showed that it is possible to obtain reliable estimates for both annual economic activity rates, as well as their precision, design mean squared error (MSE), for communes that had a sample of at least 100 people, i.e. considerably smaller than when using standard methods. The four reports of this study, which lasted for about two years, are available on SFSO's experimental statistics website, [5], [6], [7] and [8].

3.2. Procedure

Following the encouraging results obtained with Linear Mixed Models and an innovative method to estimate the design MSE using the 2012 structural survey and OASI data from 2011 [8], the SFSO then continued the study by investigating the possibility of combining this information with data pooled from the structural survey over several years, see [1]. Specifically, these estimates refer to an average target population from the structural survey, in this case over the three years 2012-2014. The main aim of the pooling is to obtain reliable estimates for a larger number of communes than can be obtained from annual surveys. The lower limit of 100 persons in the sample remains a priori the same for the pooled data and therefore the number of communes satisfying this condition can be increased by pooling the data.

3.3. Results

Based on the conclusions of the study executed by Professor I. Molina, for the structural survey the SFSO decided to publish the estimates for communes with a sample of at least 100 people. To have a sufficiently large sample and to be able to publish results for as many communes as possible, results are based on structural survey data pooled over three years, i.e. from 2012 to 2014. In contrast to standard methods, which enable results to be published for only 20% of communes, it is now possible, thanks to these new methods, to triple this figure, covering 60% of communes. At the end of 2014, this represented 80% of the permanent resident population in Switzerland aged 15 and over.

4. SFSO's data innovation projects

Five projects have been chosen to implement SFSO's data innovation strategy 1.0 [13] which focuses on the application of complementary analysis methods (e.g. predictive analysis using advanced statistical techniques, data science and machine learning) that enable the current production of official statistics to be increased or completed. These data innovation projects were launched in 2018 also with the aim of skill building inside the production units. At the time of writing this contribution the data innovation projects were still ongoing and subject to modifications with regard to the methods applied and the data used.

4.1 Area Statistics Deep Learning [16]

The SFSO's land use statistics are an invaluable tool for long-term land observation. This pilot project involves learning and mastering the use of artificial intelligence (AI) technologies to eventually automate (even partially) the visual interpretation of aerial images, which represent the core information for SFSO's land use statistics, in order to detect changes in the classification and to classify the Swiss area.

Among the learning methods that have emerged from AI for image recognition, Deep Learning is particularly adapted to statistics on land use and cover for which learning data are available in very large quantities. This pilot project is linked to an externally commissioned IT prototype aiming at implementing adequate learning methods in view of their use in a production environment.

The focus for the project team was on gathering experience of learning methods and their application in the domain of land use and cover change detection and classification because of the lack of thorough knowledge of these methods.

The external commissioned experts suggested to use a combination of a Deep Learning algorithm with a Random Forest. The first one, using a Convolutional Neural Network (CNN), is applied to RGB-data. Afterwards, the Random Forest takes the output of the CNN and auxiliary data, which were under investigation at the time of writing this contribution, as inputs.

The need to reproduce the results can be derived from principle 14.2 of Eurostat's Code of Practice [2]. Therefore, it was suggested to freeze the outcome of the deep learning algorithm and execute the final learning of the Random Forest with fixed seed. Knowing that this procedure is not standard among experts in machine learning, the need for the reliability attached to official statistics dominated even if the outcome belongs to experimental statistics at a first time.

4.2 Machine Learning SoSi [19]

This project aims at predicting trajectories inside the social security system ("Soziale Sicherheit" in german) and employment at the entrance into this system. The project was split up in two phases where the first one should provide a reduction of the number of trajectories to a few summarising meaningful trajectories by using clustering algorithms. The second phase should then allow assessing the expected trajectory and the trajectories with the highest probability by using machine learning algorithms based on demographic variables as well as on information on employment and social benefit receipt prior to entering the system. The data consisted in a cohort of about 140'000 people entering the social security system in January 2012 and their trajectories were observed monthly for four years. During these 48 months, people might keep their status, change it inside the system or become economically independent again.

At the time of writing this contribution, the first phase was still ongoing and the second had not yet started.

Concerning the clustering phase, IT performance problems had to be faced resulting in splitting the clustering in two parts. The first part consists in the application of k-means to build 3'000 clusters. For each of these clusters a characteristic trajectory had been defined as being the one with the smallest

distance to all the units in the same cluster based on the optimal matching distance. The final clustering with optimal matching was then performed on these 3'000 characteristic trajectories resulting in seven final clusters (chosen by SFSO deliberately). The impact of several parameters for the first clustering phase showed no significant impact on the final clusters of trajectories. The finetuning of the number of the final clusters is still ongoing also with the aid of trajectory indicators like the number of months of employment among others.

These trajectory indicators should help to characterise and describe the final clusters in order to give them a meaning. It is too early to say whether this approach will be successful in giving an answer to the challenge of interpretability of the clusters.

4.3 Evaluate the potential of small area estimation methods for the Job Statistics [18]

The aim of this project is to produce reliable estimates of the total number of jobs and full-time equivalents for the 26 cantons, which is not possible for all cantons with direct estimation.

The sampling design of the Job Statistics is established at the economic division level (NOGA2) and the major regions (NUTS 1). There is a considerable demand for results at detailed levels like the cantons. SFSO usually suggests to the cantons to increase the sample size through co-financing. This practice is costly for clients but also goes against the aims to reduce the burden on enterprises. Therefore, the potential of small area estimation methods to respond to these needs without increasing the sample size is evaluated in this project.

At the time of writing this contribution, the applied Linear Mixed Model showed promising results but the conceptualisation of the validation was still ongoing and the results had not yet been validated. The validation and the estimation of the design MSE are still challenges to be faced.

4.4. Data validation with machine learning [17]

The aim of this project is to extend and speed up data validation in the SFSO by means of machine learning algorithms and at the same time to improve data quality.

Statistical offices carry out data validation to check the quality and reliability of administrative data and survey data. Data that are either clearly incorrect or seem at least questionable are sent back to data suppliers with a correction request or comment. Until now, such data validation have mainly been carried out at two different levels: either through manual checks or automated processes using threshold values and logical tests. This process of two-way plausibility checks involves a great deal of work. In some cases, staff are required to manually check the data again, in other cases rules are applied that often require additional checks. This rule-based approach based on

previous experience is not necessarily exhaustive and not always precise. Machine learning could help to ensure faster and more accurate checks. This approach would rely on an algorithm using historical data at first. Based on previous data analysis, a target variable can be defined that should be predicted by the algorithm. As the final stage, the predicted and actual values of the target variables are compared and the predictive accuracy can be evaluated. Finally, a feedback mechanism has to be developed including potential error localisation to send an automatic explanation to data suppliers.

This pilot project has been presented at the UNECE Workshop on Statistical Data Editing [11] and at Eurostat's NTTS 2019 Conference [12].

4.5 Automation of NOGA coding [15]

This pilot project aims at an automation of the coding of the economic activity of enterprises using Machine Learning methods applied to data already available within the SFSO (data from surveys, descriptions in the commercial register, keywords, explanatory notes for classifications etc.) to support coding.

Because another SFSO contribution at the ISI World Statistics Congress 2019 discusses this pilot project extensively, we renounce on describing more details in this contribution.

5. Discussion and Conclusion

The presented SFSO experimental statistics showed on the example of SAE in the structural survey, paragraph (3), that research can not be speeded up if fundamental new findings are needed to produce the aimed result, even for this type of statistical product. All SFSO's experimental statistics need monitoring in the light of repeated application in order to assess their reliability and stability over time. And a crucial issue of official statistics and also for experimental statistics is its reproducibility which might be an unusual condition for the application of some algorithms like deep learning procedures.

A key condition to move experimental statistics to official statistics is their validation and if needed their improvement. All clients and everyone who shows interest in experimental statistics is therefore kindly invited to give their feedback on these published products.

References

1. Dupraz, M. and Massiani, A. and Kilchmann, D. (2018). Estimation du taux d'activité au niveau communal dans le cadre du relevé structurel: Etude de la combinaison des méthodes Small Area Estimation et du pooling (only in French). Experimental statistics report, Swiss Federal Statistical Office, Neuchâtel. https://www.bfs.admin.ch/bfs/en/home.assetdetail.6028620.html.
2. Eurostat (2018). Code of Practice. https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-0218-142.
3. Eurostat website on experimental statistics. https://ec.europa.eu/eurostat/web/experimentalstatistics.
4. Italian Statistical Institute (ISTAT) website on experimental statistics. https://www.istat.it/en/experimental-statistics.
5. Molina, I. and Strzalkowska, E. (2015). Small Area Estimation in the Structural Survey, Report part 1.1. Experimental statistics report, Swiss Federal Statistical Office, Neuchâtel. https://www.bfs.admin.ch/bfs/en/home.assetdetail.6028684.html.
6. Molina, I. and Strzalkowska, E. (2015). Small Area Estimation in the Structural Survey, Report part 1.2. Experimental statistics report, Swiss Federal Statistical Office, Neuchâtel. https://www.bfs.admin.ch/bfs/en/home.assetdetail.6028686.html.
7. Molina, I. and Strzalkowska, E. (2015). Small Area Estimation in the Structural Survey, Report part 2.1. Experimental statistics report, Swiss Federal Statistical Office, Neuchâtel. https://www.bfs.admin.ch/bfs/en/home.assetdetail.6028683.html
8. Molina, I. and Strzalkowska, E. (2015). Small Area Estimation in the Structural Survey, Report part 2.2. Experimental statistics report, Swiss Federal Statistical Office, Neuchâtel. https://www.bfs.admin.ch/bfs/en/home.assetdetail.6028685.html.
9. Office for National Statistics (ONS) website "Guide to experimental statistics". https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/guidetoexperi mentalstatistics.
10. Rao, J.N.K. and Molina, I. (2015). Small Area Estimation, Second Edition. John Wiley and Sons, Hoboken, New Jersey.
11. Ruiz, C. (2018). Improving Data Validation using Machine Learning. UNECE Workshop on Statistical Data Editing 2018, Neuchâtel. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Switzerland_RUI Z_Paper.pdf.
12. Ruiz, C. et al. (2019). Improving Data Validation using Machine Learning. Conference on New Techniques and Technologies for official Statistics

(NTTS 2019), Brussels.
https://coms.events/ntts2019/data/x_abstracts/x_abstract_170.pdf.
13. Swiss Federal Statistical Office Data Innovation Strategy (2017). https://www.bfs.admin.ch/bfs/en/home/news/whats-new.assetdetail.3862240.html.
14. Swiss Federal Statistical Office microsite on experimental statistics. https://www.experimental.bfs.admin.ch.
15. Swiss Federal Statistical Office microsite on experimental statistics – Pilot project Automation of NOGA coding (NOGAuto). https://www.experimental.bfs.admin.ch/en/nogauto.html.
16. Swiss Federal Statistical Office microsite on experimental statistics – Pilot project Area Statistics Deep Learning (ADELE). https://www.experimental.bfs.admin.ch/en/adele.html.
17. Swiss Federal Statistical Office microsite on experimental statistics – Pilot project Data Validation. https://www.experimental.bfs.admin.ch/en/data-validation.html.
18. Swiss Federal Statistical Office microsite on experimental statistics – Pilot project Evaluate the potential of small area estimation methods for the Job Statistics (JOBSTAT). https://www.experimental.bfs.admin.ch/en/jobstat.html.
19. Swiss Federal Statistical Office microsite on experimental statistics – Pilot project Machine Learning SoSi (ML_SoSi). https://www.experimental.bfs.admin.ch/en/ml_sosi.html.
20. Swiss Federal Statistical Office microsite on experimental statistics – Small area estimation (communes) of economic activity rate in the structural survey. https://www.experimental.bfs.admin.ch/en/sae.html.

Robust estimation of treatment effects in a latent-variable framework

Mikhail Zhelonkin
Erasmus University Rotterdam

Abstract

The policy evaluation is one of the central problems in modern economics. Unfortunately, it is usually impossible to perform a randomized experiments in order to evaluate the treatment effects. Hence, the data from observational studies has to be used. In this case the sample is typically non-random and one has either to correct for selectivity or to impose (conditional) independence assumtion. Since this assumption is often irrealistic, the structural latent variable model is used. The parametric estimators (although, they are straightforward to compute and to interprete) have been criticized for sensitivity to the departures from the distributional assumptions. The alternative semi- and nonparametric estimators have complex identification and are limited to estimation of certain parameter(s) of interest but do not allow for the general evaluation and interpretation of the model. In this work we employ the latent-variable framework. We study the robustness properties of the estimators of three principal parameters: average treatment effects, average treatment effects on the treated and local average treatment effects and propose the robust alternatives.

Keywords

Inuence function; Robustness; Treatment effect.

1. Introduction

The topic of causality and policy evaluation is one of the central topics in Statistics and Econometrics. In an ideal world, in order to draw inferences about certain policy we need to perform the randomized controlled experiments. Unfortunately, this is hard or sometimes impossible to do in practice. For instance, it would be unethical to force random people to smoke or to forbid potential students to attend universities. Thus, a lot of empirical research (see the review by Imbens and Wooldridge 2009 for a list of examples) in Economics, Social Sciences, and sometimes in Biology and Medicine has to deal with the observational data. In this paper we focus on the situation, when there is self-selection into treatment and unconfoundedness assumption does not hold. Assume a parametric model F. The selection equation is given by

$$D = I((w^T\gamma + e^s \geq 0), \quad (1)$$

where I(A) is the indicator function, which is equal to 1 if A is true, w is a vector of covariates, γ is a vector of parameters. The potential outcomes are defined by

$$y^1 = x^T\beta^1 + e^1, \text{ if } D = 1, \quad (2)$$

$$y^0 = x^T\beta^0 + 0, \text{ if } D = 0, \quad (3)$$

where the superscript 1 denotes the treated state and the superscript 0 denotes the untreated state, x is a vector of explanatory variables, β^1 and β^0 are vectors of parameters. Notice that x and w can overlap and can be even the same. The latter is a somewhat pathological case (although not uncommon in practice), since the model identification relies completely on parametric assumptions. It is in general desirable to have at least one significant predictor in w that is not in x. Such a variable is known as exclusion restriction variable in literature on sample selection models, or instrumental variable in policy evaluation literature. Unfortunately, it is often difficult to find such a variable in practice. In the classical textbook formulation the error terms e^s, e^1 and e^0 follow a multivariate normal distribution

$$\begin{pmatrix} e^s \\ e^1 \\ e^0 \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & \sigma_{1s} & \sigma_{0s} \\ \sigma_{1s} & \sigma_1^2 & \sigma_{10} \\ \sigma_{0s} & \sigma_{10} & \sigma_0^2 \end{pmatrix} \right\}, \quad (4)$$

where $\mathrm{Var}(e^s)$ is set to 1 to ensure identifiability. Note that, the covariance between e^1 and e^0 cannot be estimated, since the potential outcomes cannot be observed simultaneously.

The model (1)-(4) exploits very strong distributional assumptions and had been criticized for its sensitivity even to minor contamination. From the statistician's point of view, estimation of sample selection correction model requires addition of supplementary, often too restrictive and not viable assumptions, while an economist might argue, that the structural model has connection to economic theory and is therefore interesting and important. It allows to see the big picture, i.e. not only the treatment effect, of course, conditional on correct specification of the parametric model. The requirement of correct specification is perhaps the weakest point of parametric methods in this case.

The framework, when the assumed model F is contaminated by an unknown model G with some small probability $\boldsymbol{\epsilon}$, is reasonable to represent a true data-generating process. In this paper we study the robustness problem of evaluation of treatment effects using infinitesimal approach (Hampel et al. 1986). It allows to formalize the problem of sensitivity to (small) departures from assumed model and to construct the estimators and tests which are reliable in presence of contamination.

The article is structured as follows. In Section 2 we introduce the parameters of interest. In Section 3 the robustness problem is formalized. Section 4 outlines the construction of the robust estimators. The numerical

examples are given in Section 5 and Sectio 6 concludes and provides some practical recommendations.

2. Treatment Parameters of Interest

We concentrate on three treatment parameters, namely the average treatment effect (ATE), the average treatment effect on the treated (ATT), and the local average treatment effect (LATE) (Imbens and Angrist 1994). The expressions of these parameters in latent variable framework using normality assumption were provided by Heckman et al. (2001).

The ATE conditional on x is

$$\mathrm{ATE}(x) = x^{\top}(\beta^{1} - \beta^{0}). \tag{5}$$

The unconditional version can be obtained as follows

$$\mathrm{ATE} = \int ATE(x)dF(x) = E(x^{\top})(\beta^{1} - \beta^{0}), \tag{6}$$

where E denotes expectation.

The conditional ATT is given by

$$ATT\{X, W, d(w) = 1\} = x^{T}(\beta^{1} - \beta^{0}) + \left(\beta_{\lambda}^{1} + \beta_{\lambda}^{0}\right)\frac{\phi(w^{T}\gamma)}{\Phi(w^{T}\Upsilon)} \tag{7}$$

Its unconditional version is

$$\mathrm{ATT} = \int \mathrm{ATT}\{x, w, D(w) = 1\}\, dF\{x, w \,|D(w) = 1\}. \tag{8}$$

The LATE is given by

$$\mathrm{LATE}\{x, D(w) = 0, D(\breve{w}) = 1\} = x^{\top}(\beta^{1} - \beta^{0}) + (\beta_{\lambda}^{1} + \beta_{\lambda}^{0})\lambda_{l}, \tag{9}$$

Where $\lambda_{l} = \frac{\phi(\breve{w}^{T}\lambda) - \phi(w^{T}\gamma)}{\Phi(\breve{w}^{T}\lambda) - \Phi(w^{T}\Upsilon)}$. The unconditional LATE is

$$\mathrm{LATE} = \int \mathrm{LATE}\{x, D(w) = 0, D(\breve{w}) = 1\} dF(x). \tag{10}$$

The parameters of model (1)-(4) can be estimated by Heckman (1979) two-stage procedure. The first step is the probit maximum likelihood estimator with the following population version of estimating equations

$$\int \{D - \Phi(w^{T}\Upsilon)\}\frac{\phi(w^{T}\gamma)}{\Phi(w^{T}\Upsilon)\{1 - \Phi(w^{T}\Upsilon)\}} w dF = 0$$

The second step is ordinary least squares estimator with control functions correcting for selectivity. The estimating equations are

$$\int (y_{j} - x^{T}\beta^{j} - \lambda^{j}\beta_{\lambda}^{j})\begin{pmatrix} x \\ \lambda^{j} \end{pmatrix} D dF,$$

where $j = 0.1$. Then, as suggested by Heckman et al. (2003) the conditional treatment effect parameters can be evaluated by plugging in estimated parameters from the two-stage estimator into (5), (7) and (9). The unconditional versions are obtained additionally by averaging over $X's$.

3. Robustness Properties of Treatment Effect Estimators

For a statistical functional T(F) the Influence Function (IF) was defined by Hampel (1974) as

$$\mathrm{IF}(z; T, F) = \lim_{\epsilon \to 0}[T\{(1 - \epsilon)F + \epsilon\Delta_{z}\} - T(F)]/\epsilon,$$

where Δ_z is the probability measure which puts mass 1 at the point z. The IF describes the asymptotic bias on the estimator due to a small amount of contamination $\boldsymbol{\epsilon}$ at the point z. The IF is a convenient and useful heuristic tool. It allows not only to formalize the robustness problem, but also provides a way to construct a robust statistic.

Denote three statistical functionals S, T^1 and T^0, corresponding to the estimators of selection (1), treated (2) and untreated (3) equations, respectively. The IF of the estimator of probit is

$$\text{IF}(z;S,F)=\left(\int[\frac{\phi(w^T\gamma)^2ww^T}{\Phi(w^T\Upsilon)\{1-\Phi(w^T\Upsilon)\}}]dF\right)^{-1}\{D-\Phi(w^T\Upsilon)\}\frac{\phi(w^T\gamma)w}{\Phi(w^T\Upsilon)\{1-\Phi(w^T\Upsilon)\}}. \quad (11)$$

It is clear that the IF in (11) is unbounded with respect to w, hence the MLE can be arbitrarily biased by contamination.

The estimator of the parameters of treated state is the Heckman's (1979) two-step estimator. It's IF was derived by Zhelonkin et al. (2016) and is given by

$$IF(z;T_1,F)=\left\{\int\begin{pmatrix}xx^T & \lambda^1x\\ \lambda^1x^T & (\lambda^1)^2\end{pmatrix}DdF\right\}^{-1}\left\{\left(y^1-x^T\beta^1-\lambda^1\beta_\lambda^1\right)\begin{pmatrix}x\\ \lambda^1\end{pmatrix}D+\int D\begin{pmatrix}x\beta_\lambda^1\\ \lambda^1\beta_\lambda^1\end{pmatrix}\frac{\partial\lambda^1(w^T\gamma)}{\partial\gamma}dF\cdot IF(z;S,F)\right\}, \quad (12)$$

where $\lambda^1(w^T\gamma)$ denotes dependence of λ^1 on the linear predictor $w^T\gamma$. With minor modifications the result in (12) is adapted for the untreated state:

$$IF(z;T_0,F)=\left\{\int\begin{pmatrix}xx^T & \lambda^0x\\ \lambda^0x^T & (\lambda^0)^2\end{pmatrix}(1-D)dF\right\}^{-1}\left\{\left(y^0-x^T\beta^0-\lambda^0\beta_\lambda^0\right)\begin{pmatrix}x\\ \lambda^0\end{pmatrix}(1-D)+\int(1-D)\begin{pmatrix}x\beta_\lambda^0\\ \lambda^1\beta_\lambda^0\end{pmatrix}\frac{\partial\lambda^0(w^T\gamma)}{\partial\gamma}dF\cdot IF(z;S,F)\right\}, \quad (13)$$

where $\lambda^0(w^T\gamma)$ also depends on the linear predictor. The IF's for both states are unbounded with respect to x, y^0, y^1, λ^0 and λ^1. Moreover they contain the IF of probit MLE, which provides the second source of unboundedness. The control functions λ^0 and λ^1 are unbounded from the right and from the left, respectively.

a. Influence Functions of Treatment Parameters Given x and w

Denote the statistical functionals τ_c, τ_c^t and τ_c^l corresponding to the conditional on x and w estimators of the ATE, ATT and LATE, respectively. Then their IF's are given by the three following propositions. Proofs can be requested from the author.

Proposition 1. The IF of the ATE estimator conditional on x and w is

$$IF(z;\tau_c,F)=\begin{pmatrix}x^T & 0 & -x^T & 0\end{pmatrix}\begin{Bmatrix}IF(z;T_1,F)\\ IF(z;T_0,F)\end{Bmatrix}. \quad (14)$$

Proposition 2. The IF of the ATT estimator conditional on x and w is

$$IF\left(z;\tau_c^l,F\right)=\begin{Bmatrix}x^T & \lambda^1 & -x^T & \lambda^1 & (\beta_\lambda^1+\beta_\lambda^0)\frac{\partial\lambda^1}{\partial\gamma}\end{Bmatrix}\begin{Bmatrix}IF(z;T_1,F)\\ IF(z;T_0,F)\\ IF(z;S,F)\end{Bmatrix}, \quad (15)$$

where

$$\frac{\partial \lambda^1}{\partial \gamma} = \frac{-\Phi(w^T\gamma)\phi(w^T\gamma)w^T\gamma - \phi(w^T\gamma)^2}{\Phi(w^T\gamma)^2} w^T.$$

Proposition 3. The IF of the LATE estimator conditional on x and w is

$$IF\left(z;\tau_c^l,F\right) = \left\{x^T \quad \lambda^1 \quad -x^T \quad \lambda_l \quad (\beta_\lambda^1 + \beta_\lambda^0)\frac{\partial \lambda_l}{\partial \gamma}\right\}\begin{Bmatrix} IF(z;T_1,F) \\ IF(z;T_0,F) \\ IF(z;S,F) \end{Bmatrix}, \quad (16)$$

where

$$\frac{\partial \lambda_l}{\partial \gamma} = \frac{\breve{w}^T\phi(\breve{w}^T\gamma)\breve{w}^T\gamma - w^T\phi(w^T\gamma)w^T\gamma}{\Phi(\breve{w}^T\gamma) - \Phi(w^T\gamma)} - \frac{\phi(\breve{w}^T\gamma)\breve{w}^T - \phi(w^T\gamma)w^T}{\{\Phi(\breve{w}^T\gamma) - \Phi(w^T\gamma)\}^2}\{\phi(\breve{w}^T\gamma) - \phi(w^T\gamma)\}.$$

It is clear that all three IF's are unbounded and can be arbitrarily biased.

b. IF's of unconditional ATE, ATT and LATE

Denote the statistical functionals τ, τ^t and τ^l corresponding to the estimators of the ATE, TT and LATE, respectively.

Proposition 4. The *IF* of the *ATE* estimator is

$$IF(z;\tau,F) = x^T(\beta^1 - \beta^0) - \tau + E\{x^T \quad 0 \quad -x^T \quad 0\}\begin{Bmatrix} IF(z;T^1,F) \\ IF(z;T^0,F) \end{Bmatrix}.$$

Proposition 5. The *IF* of the *ATT* estimator is

$$IF(z;\tau^t,F) = (\int DdF)^{-1}\Big[D\{x^T(\beta^1 - \beta^0) + (\beta_\lambda^1 + \beta_\lambda^0)\lambda^1 - \alpha^t\} + \int D(x^T \quad \lambda^1 \quad -x^T \quad \lambda^1\}dF\begin{Bmatrix} IF(z;T^1,F) \\ IF(z;T^0,F) \end{Bmatrix} + \int D(\beta_\lambda^1 + \beta_\lambda^0)\frac{\partial \lambda^1}{\partial \gamma}dFIF(z;S,F)\Big] \quad (17)$$

Proposition 6. The *IF* of the *LATE* estimator is

$$IF\left(z;\tau^l,F\right) = (\int DdF)^{-1}\Big[D\{x^T(\beta^1 - \beta^0) + (\beta_\lambda^1 + \beta_\lambda^0)\lambda^1 - \alpha^l\} + \int D(x^T \quad \lambda_l \quad -x^T \quad \lambda_l\}dF\begin{Bmatrix} IF(z;T^1,F) \\ IF(z;T^0,F) \end{Bmatrix} + \int D(\beta_\lambda^1 + \beta_\lambda^0)\frac{\partial \lambda^1}{\partial \gamma}dFIF(z;S,F)\Big] \quad (18)$$

An obvious but important remark is that if there are outliers in the instrumental variable, then the entire estimator of the model and of the treatment effect can become arbitrarily biased.

4. Robust Estimation

The expressions of the IF's in the previous section pave the way to construct the robust estimators. We need to bound the sources of unboundedness of the IF's. The first step is to construct the bounded-influence estimator for the selection equation such that the IF in (11) is bounded. Then the second step is to bound the IF(z ; T_0,F) and IF(z ; T_1,F). Following the approach proposed by Zhelonkin et al. (2016) we have the following score functions

$$\Psi_s^R\{z_1;S(F)\} = \nu(z_1;\mu)\omega_1(w)\mu' - \alpha(\gamma), \quad (19)$$

$$\Psi_0^R\ (z_2;\lambda^0,T) = \psi_{c1}(y^0 - x^\top > \beta^0 - \lambda^0\beta^0\ \lambda)\omega(x,\lambda^0)y^0, \quad (20)$$

$$\Psi_1^R\ (z_2;\lambda^1,T) = \psi_{c1}(y^1 - x^\top > \beta^1 - \lambda^1\beta^1\ \lambda)\omega(x,\lambda^1)y^1, \quad (21)$$

where $\alpha(\beta_1) = \frac{1}{n}\sum_{i=1}^{n} E\{\nu(z_{1i};u_i)\}\omega_1(w_1)\mu_i'$ ensures Fisher-consistency, $\mu_i = \mu_i(z_{1i},\gamma) = \Phi(w_{1i}^\top\gamma)$, $\mu_i' = \frac{\partial}{\partial\gamma}\mu_i$, $\nu(z_{1i};\mu_i)$ is the weight function defined as follows:

$$\nu(z_{1i};u_i) = \psi_{c_s}\left\{\frac{y_{1i}-u_i}{V^{1/2}(u_i)}\right\}\frac{1}{V^{1/2}(u_i)},$$

where ψ_{cs} is the Huber function defined by

$$\psi_{c_s}(r) = \begin{cases} r, & |r| \le c_s, \\ c_s sign(r), & |r| > c_s. \end{cases} \quad (22)$$

Functions ψ_{c1} are the Huber functions defined above, but can have different tuning constant. The weight functions $\omega_1(w_i)$ and $\omega(x,\lambda^j)$, $j = 0,1$ can be based on the hat matrix, for instance $\omega 1(\cdot)$ is $\omega_{1i} = \sqrt{1 - H_{ii}}$, where H_{ii} is the ith diagonal element of the hat matrix $H = W(W^\top > W)^{-1}\ W^\top$, or they can be based on robust Mahalanobis distance $d(x,\lambda^j)$ such as

$$\omega(x,\lambda^j) = \begin{cases} x, & if\ d(x,\lambda^j) < c_{m,} \\ \frac{xc_{m,}}{d(x,\lambda^j)}, & if\ d(x,\lambda^j) \ge c_{m,} \end{cases} \quad (23)$$

where c_m is chosen according to the level of tolerance (say 95%), given that the squared Mahalanobis distance follows a χ^2-distribution. The estimator defined by (19)-(21) has bounded influence function and is robust for estimation of conditional treatment effects in Section 3.1.

In order to obtain robust estimator of unconditional treatment effects we need to make a further step. We propose to use a classical Huber estimator of location.

5. Numerical Examples

5.1 Simulation Study

To illustrate the robustness issue and performance of robust estimator I carry out a Monte Carlo simulation study. We generate $D_i = I(w_{1i} + 0.5w_{2i} + 0.75w_{3i} + e_i^s)$, where $w_{ji} \sim N(0,1)$, for $j = 1,2,3$. For the equations of interest we set $x = w$, and if there is an exclusion restriction we generate x_3 independently from w_3 from the same distribution. Then the treated and untreated equations are $y_i^1 = 1 + x_{1i} + 0.5x_{2i} + 0.5x_{3i} + e_i^1$ and $y_i^0 = x_{1i} + 0.5x_{2i} + 0.5x_{3i} + e_i^0$, respectively. The error terms e_i^s, e_i^1 and e_i^0 follow a multivariate normal with expectation 0, variances 1 and $\mathrm{Cov}(e^0,e^s) = 0.25$ and $\mathrm{Cov}(e^1,e^s) = -0.75$. The distribution of observations into the states can be controlled by the intercept in selection equation, for simplicity it is set to 0, which gives 50% chance to be in each state. The true values of the treatment parameters are 1 for ATE, 1.8 for ATT and -0.26 for $LATE$. The simulation setup is an adaptation of the setup in Zhelonkin et al. (2016). The setup in Heckman et al. (2003) used only one explanatory variable and the effects of contamination were small.

When the dimension increases the effects of contamination become more pronounced. Moreover the identification of outliers becomes more difficult.

We study several types of contamination. In the first scenario we generate observations from the model described above and with probability 0.01 replace the original observations by outliers, i.e. by a point mass at w = x = $(-2,-2,-2)$ and $(\mathrm{D},y^1,y^0) = (1,0,1)$. In this case we have leverage outliers in the treated state, i.e. the observation is unlikely to be in treated state but emerges there. In the second scenario we study teh effect of outliers in the untreated state, when the observation should be in treated state but appears in untreated. The point mass is at $w = x = (2,2,2)$ and $(\mathrm{D},\mathrm{y}^1,\mathrm{y}^0) = (0,1,0)$. The sample size is $N = 1000$ and we repeat the experiment 500 times. For other sample sizes ($N = 500$ or $N = 5000$) the results are similar.

We compare the classical parametric estimator with the robust alternative. When the model is correct both estimators perform well and the classical estimator is more efficient. The efficiency loss is approximately 25-27%, what corresponds to the theory. When the contamination is added, the classical estimator becomes biased, while the robust estimator remains stable. Of course, it allows some bias (it is noticeable in the case when exclusion restriction is not available), but the bias is bounded and small. Another important aspect is the variability of the classical estimator. Under contamination it is much less efficient than the robust estimator. In the context of sample selection models, Zhelonkin et al. (2016) derived the change-of-variance function for the Heckman's 2-stage estimator and showed that it was unbounded. The change-of-variance function is an analog of the IF but for the variance functional (see Hampel et al. 1981) and has similar interpretation. This can explain the effect of inflated variability of the classical estimator of treatment effects.

6. Discussion and Conclusion

From the practical point of view, the proposed robust estimators can be used in two ways. First, they should be used for, what Athey and Imbens (2017) called supplementary analysis, which is meant to convince the reader that the results of the modeling are credible. If the distributional assumptions hold, then the robust and classical estimators should be close, since both are consistent at the model. If they diverge, then it is a strong indication that the assumptions are violated. Second, the robust estimators can be used as stand-alone estimators in the cases, when analyst believes that the model assumptions hold (at least for the majority of observations), but there might be a few atypical but valid observations that can distort estimation and inference. The robust estimator should not be considered as a replacement for semi- and nonparametric methods. It is designed to deal with small departures from the model and is fully parametric by nature. If the analyst sees that the

parametric assumptions do not hold even approximately, then the semi- and nonparametric methods should be preferred.

References

1. Athey, S. and Imbens, G. W. (2017), \The State of Applied Econometrics: Causality and Policy Evaluation," Journal of Economic Perspectives, 31, 3{32.
2. Hampel, F. (1974), \The Inuence Curve and Its Role in Robust Estimation," Journal of the American Statistical Association, 69, 383{393.
3. Hampel, F., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), Robust Statistics: The Approach Based on Inuence Functions, New York: John Wiley and Sons.
4. Hampel, F., Rousseeuw, P. J., and Ronchetti, E. (1981), \The Change-of-Variance Curve and Optimal Re-descending M-Estimators," Journal of the American Statistical Association, 76, 643{648.
5. Heckman, J. (1979), \Sample Selection Bias as a Speci_cation Error," Econometrica, 47, 153{161.
6. Heckman, J. J., Tobias, J. L., and Vytlacil, E. (2001), \Four Parameters of Interest in the Evaluation of Social Programs," Southern Economic Journal, 68, 210{223.
7. |(2003),\Simple Estimators for Treatment Parameters in a Latent Variable Framework,"Review of Economics and Statistics, 85, 748{755.
8. Imbens, G. W. and Angrist, J. D. (1994), \Identi_cation and Estimation of Local Average Treatment Effects," Econometrica, 62, 467{475.
9. Imbens, G. W. and Wooldridge, J. M. (2009), \Recent Developments in the Econometrics of Program Evaluation," Journal of Economic Literature, 47, 5{86.
10. Zhelonkin, M., Genton, M. G., and Ronchetti, E. (2016), \Robust Inference in Sample Selection Models,"Journal of the Royal Statistical Society Series B, 78, 805{827.

A new additive index number system with maximum characteristicity for international price comparisons

Chang Xie[1]; Zhongshan Yang[1]; Prasada Rao[2]
[1] Dongbei University of Finance and Economics,China
[2] The University of Queensland,Australia

Abstract

Additivity and characteristicity are important properties that should be possessed by aggregation methodsused in international price comparisons and in the compilation of purchasing power parities (PPPs) of currencies. Currently, aggregation methods used for computing PPPs in the International Comparison Program (ICP) at the World Bank ICP do not satisfy both of these properties. The Gini-Elteto-Koves-Szulc (GEKS) method used in the ICP maintains characteristicity with respect to the Fisher binary index numbers but fails to satisfy additivity. In this paper we propose the Maximal Bilateral Characteristic (MBC) Method which satisfies additivity and at the same time deviates the least from the matrix of binary Fisher index numbers. The simple MBC method is then generalized leading to a Weighted Maximal Bilateral Characteristic (WMBC) method where each bilateral comparison is given a weight that reflects its reliability. Reliability is measured using three difference indicators: (i) the Laspeyres-Paasche distance; (ii) price structure (non) similarity index, and (iii) quantity structure(non) similarity index. This approach is operationalizedusing seven different weight functions. Finally, an empirical illustration using the ICP2011 data shows that the MBC results are very close to the Fisher binary indexesand the officialresultsbased on the GEKS method. Therefore, the MBC method combines the advantages of characteristicity implicit in the GEKS method and that of additivity property of the Geary-Khamis (GK) method. In practical applicationsof this new method, it is necessaryto select an appropriate weight functionreflecting the objectives of the empirical analysis. Future studies should focus on further improvement on themeasuresof reliabilityof bilateral comparison.

Keywords

Additivity; Characteristiciy; GK, GEKS; MBC

1. Introduction

The calculation of purchasing power parities (PPP) and the real income comparison based on PPP are the core concerns of the International Comparison Program (ICP). An ideal PPP aggregation method should satisfy some important index properties such as additivity, transitivity, base country invariance and characteristicity, and these properties are also used to evaluate

whether an aggregation method is good or not. Although all the aggregation methods used by ICP satisfy transitivity and base country invariance, currently there is no aggregation method that combines the two important properties of additivity and characteristicity. This means that in practice a choice has to be made as to which aggregation method will be used. Looking back at the history of several phases of the ICP, the Geary-Khamis method which satisfies additivity was selected by Kravis, Heston and Summers (1982). However, the Eurostat opted for the use of theGini-Elteto-Koves-Szulc (GEKS) method which satisfies characteristicitiy but fails additivity. Since the 2005 round of the ICP, the Geary-Khamis method was replaced by GEKS which maintains a degree of characteristicity but fails additivity.

To make the PPP aggregation method satisfy more good properties, this paper builds a new additive aggregation method. The method on the one hand will achieves additivity within the framework of GK's additivity, on the otherhand references the GEKS method, and from two different angles to approximate the binary optimal index——the Fisher index, so this method satisfies additivity and maximizes characteristicity. It will put forward both of the basic form of this new method and its generalized form.

2. Methodology

2.1 Construction of the Maximal Bilateral Characteristic Method

2.1.1 The MBC Method Based on the Extremum Optimization Angle

Here, the following variables should be defined. P_{ij} is the price of commodity i ($i = 1,2,...,I$) in country j ($j = 1,2,...,J$), and q_{ij} is the corresponding consumption quantity .

The key of the MBC method is to identify an international average price vector (π) which can realize the characteristicity to the greatest extent. It is to solve the following optimization problem:

$$\min_{\pi}(\max \mu_j) = \min\left\{ \max\left(\left| \frac{\sum_{i=1}^{I} \pi_i q_{ij}}{\sum_{i=1}^{I} \pi_i q_{ik}} / Q_{kj}^{F} - 1 \right| \right) \right\} \quad (1)$$

$$\pi_i > 0 (i = 1,2,...,I); \forall k, j = 1,2,...,J$$

Q_{kj}^{F} is the Fisher volume index.

When getting the optimal international average price vector of π, we can use the following equation to calculate each country's PPP which can maintain additivity:

$$ppp_j = \frac{\sum_{i=1}^{I} p_{ij} q_{ij}}{\sum_{i=1}^{I} \pi_i q_{ij}} \tag{2}$$

2.1.2 The MBC Method Based on the Least-Squares Angle

The MBC method based on the least-squares angle is equal to solve the following objective function:

$$\min \sum_{j=1}^{J} \sum_{k=1}^{J} \left(\frac{\sum_{i=1}^{I} \pi_i q_{ij}}{\sum_{i=1}^{I} \pi_i q_{ik}} - Q_{kj}^{F} \right)^2, with, \pi_i > 0, i = 1, 2..., I \tag{3}$$

Similar to the least-squares issue of the GEKS method, the solutions of $\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_J$ in the MBC method based on least-squares angle is equal to the OLS estimators of the following non-linear regression model:

$$Q_{kj}^{F} = \frac{\sum_{i=1}^{I} \pi_i q_{ij}}{\sum_{i=1}^{I} \pi_i q_{ik}} + u_{kj} \tag{4}$$

$$\text{with}, \pi_i > 0, i = 1, 2..., I; E(u_{kj}) = 0; v(u_{kj}) = \sigma^2; \forall k, j = 1, 2, ..., J, j \neq k$$

In addition, if it assumes that the Fisher volume index is decomposed into the product of the volume index under multilateral comparison and a random error term. At this point, the solutions of the MBC method based on the least-squares angle are equivalent to the OLS estimators of the following logarithmic nonlinear regression model:

$$\log Q_{kj}^{F} = \log \sum_{i=1}^{I} \pi_i q_{ij} - \log \sum_{i=1}^{I} \pi_i q_{ik} + u_{kj} \tag{5}$$

$$with, \pi_i > 0, i = 1, 2..., I; E(u_{kj}) = 0; \; v(u_{kj}) = \sigma^2; \forall k, j = 1, 2, ..., J, j \neq k$$

Here, the above mentioned MBC method can be called the MBC method based on the logarithmic least-squares angle.

2.2 The Weighted MBC Method

2.2.1 The Weighted MBC Method Based on the Extremum Optimization Angle

Here we define W_{kj} as the indicator that measures the reliability degree of the bilateral comparison between country k and country j, and if the Fisher volume index is more reliable, then the value of W_{kj} is higher. After introducing weights idea, the optimization issue in the WMBC method will be as follows:

$$\min_{\pi} (\max \mu_j) = \min \left\{ \max \left(w_{kj} \left| \frac{\sum_{i=1}^{I} \pi_i q_{ij}}{\sum_{i=1}^{I} \pi_i q_{ik}} / Q_{kj}^{F} - 1 \right| \right) \right\} \tag{6}$$

$$\pi_i > 0 (i = 1, 2, ..., I); \forall k, j = 1, 2, ..., J$$

2.2.2 The Weighted MBC Method Based on the Least-Squares Angle

In the regression model of the MBC method based on the least-squares angle, by assuming the distribution form of the random error term in equation (4), the variance of the Fisher volume index can be assumed to distinguish the reliability differences of different pairwise comparisons. This can be achieved using the following model:

$$Q_{kj}^{F}=\frac{\sum_{i=1}^{I}\pi_i q_{ij}}{\sum_{i=1}^{I}\pi_i q_{ik}}+u_{kj} \tag{7}$$

$$\text{with},\pi_i>0, i=1,2...,I; E(u_{kj})=0; v(u_{kj})=\frac{\sigma^2}{w_{kj}}; \forall k, j=1,2,...,J, j\neq k$$

In this case, the WMBC method is equivalent to solve the WLS estimation of equation (7). And similarly, we can also construct the MBC method based on the logarithmic least-squares angle.

3. Result

This paper uses the price data and expenditure data of the categories of the actual individual consumption in ICP2011and introduces seven different weight functions

Table 1 Actual Individual Consumption Volume Index（US=100）

Country	Fisher	GEKS	GK	MBC-EO	MBC-LS	MBC-LS-log
Congo, Dem. Rep	0.258	0.256	0.250	0.270	0.251	0.250
Central African	0.031	0.033	0.031	0.034	0.033	0.032
Tanzania	0.379	0.397	0.383	0.427	0.392	0.397
Madagascar	0.227	0.242	0.229	0.253	0.243	0.246
Djibouti	0.013	0.013	0.013	0.014	0.013	0.013
Mauritania	0.060	0.062	0.060	0.065	0.059	0.061
Cambodia	0.287	0.281	0.292	0.306	0.297	0.282
Angola	0.686	0.719	0.719	0.767	0.747	0.736
Philippines	3.472	3.621	3.580	3.794	3.592	3.626
Bhutan	0.025	0.025	0.025	0.026	0.025	0.025
Indonesia	10.077	9.996	10.076	10.659	9.985	9.966
Mauritius	0.148	0.149	0.148	0.156	0.150	0.150
Bulgaria	0.732	0.720	0.735	0.752	0.730	0.725
Croatia	0.599	0.602	0.597	0.617	0.607	0.597
Estonia	0.167	0.167	0.167	0.171	0.170	0.165
Malta	0.080	0.081	0.080	0.082	0.081	0.081
Italy	12.444	12.891	12.676	13.104	12.851	12.876

Taiwan, China	5.297	5.099	5.232	5.246	5.059	5.190
Australia	5.275	5.378	5.366	5.433	5.420	5.441
Switzerlan	1.984	2.019	2.035	2.068	2.065	2.069
China Mainland	54.025	51.435	52.999	53.725	50.957	51.562
India	29.744	31.165	31.191	32.643	31.276	31.234

Table 2 Actual Individual Consumption Volume Index（US=100）：The WMBC Method Based on Extremum Optimization

Country	WMBC EO LP	WMBC EO KHS	WMBC-EO HSA	WMBC EO VMTp	WMBC EO WRPDF	WMBC EO WRPDT	WMBC-EO VMTq	WMBC EO WAQD	Max-Min
Congo, Dem. Rep	0.249	0.266	0.269	0.268	0.249	0.250	0.272	0.245	0.027
Central African	0.032	0.033	0.034	0.034	0.032	0.032	0.034	0.031	0.003
Tanzania	0.393	0.418	0.425	0.426	0.410	0.406	0.430	0.392	0.038
Madagascar	0.245	0.242	0.252	0.252	0.242	0.241	0.258	0.231	0.027
Djibouti	0.013	0.014	0.014	0.014	0.012	0.013	0.015	0.013	0.002
Mauritania	0.061	0.064	0.065	0.065	0.062	0.061	0.065	0.059	0.006
Cambodia	0.292	0.299	0.305	0.303	0.287	0.287	0.300	0.277	0.028
Angola	0.738	0.749	0.765	0.763	0.722	0.726	0.765	0.726	0.043
Philippines	3.651	3.707	3.777	3.762	3.608	3.603	3.727	3.553	0.224
Bhutan	0.025	0.026	0.026	0.026	0.025	0.025	0.026	0.024	0.002
Indonesia	10.028	10.424	10.568	10.512	9.835	9.859	10.514	9.710	0.858
Mauritius	0.150	0.151	0.155	0.155	0.147	0.147	0.156	0.143	0.013
Bulgaria	0.734	0.745	0.751	0.747	0.713	0.713	0.755	0.694	0.061
Croatia	0.604	0.612	0.614	0.612	0.587	0.590	0.616	0.575	0.042
Estonia	0.169	0.168	0.171	0.170	0.162	0.163	0.169	0.160	0.011
Malta	0.082	0.081	0.081	0.081	0.080	0.081	0.081	0.079	0.003
Italy	12.903	12.914	13.055	13.055	12.693	12.697	13.160	12.536	0.625
Taiwan, China	5.230	5.235	5.230	5.240	5.210	5.183	5.189	5.053	0.188
Australia	5.429	5.374	5.423	5.422	5.394	5.397	5.403	5.351	0.078
Switzerlan	2.071	2.051	2.065	2.065	2.033	2.039	2.065	2.028	0.043
China Mainland	52.212	53.827	53.593	53.590	52.262	52.005	53.777	50.317	3.510
India	31.266	32.019	32.507	32.555	31.313	31.237	32.429	31.434	1.318

For the content limitation, we don't display all the tables.

4. Discussion and Conclusion

It shows that the MBC results are very close to the Fisher index and the official GEKS results, which shows that the MBC method better achieves the characteristics of the bilateral comparison. The MBC method theoretically can

achieve characteristicity to the largest extent, and the international average price vector in this method is not affected by the size of a country, so this method may reduce the substitution biases compared with the GK method. In addition, the MBC method also satisfies additivity, so the MBC method can be recommended for actual volume comparison and actual expenditure structure analysis.

The empirical illustration verifies the necessity of considering the reliability of bilateral comparison and the feasibility of introducing weight functions. It is worth noting that in practical applications, it may need to further study the constructive principles, advantages and disadvantages of various indexes that measure the reliability of bilateral comparison, and choose the corresponding weight function according to the focus of research. Or the geometric mean of the WMBC results using different weight functions can be used as a compromise. But it is more important that further improvement on the index that measure the reliability of bilateral comparison should be the focus of future researches.

References

1. Balk, B. M., "A Comparison of Ten Methods for Multilateral International Price and Volume Comparison", *Journal of Official Statistics*12（2）, 199-222, 1996.
2. Cuthbert, J. R. , "Categorisation of Additive Purchasing Power Parities", *Review of Income and Wealth*, 45（2）, 235-249, 1999.
3. Drechsler, L. , "Weighting of Index Numbers in Multilateral International Comparisons", *Review of Income and Wealth*, 1（1）, 17-34, 1973.
4. Dikhanov, Y. , "Sensitivity of PPP-Based Income Estimates to Choice of Aggregation Procedures", a paper presented at 23rd GeneralConference of the International Association for Research in Income and Wealth St. Andrews, Canada, 21-27, August 1994.
5. Diewert, W. E. , "Similarity Indexes and Criteria for Spatial Linking", in Rao, D. S. Prasada（eds）, *Purchasing Power Parities of Currencies : Recent Advances in Methods and Applications*, 2009.
6. Elteto, O. , and Koves, P. , "On an Index Computation Problem in International Comparisons", *Statiztikai Szemle*, （42）, 507-518, 1964 .
7. Geary, R. G. , "A Note on the Comparison of Exchange Rates and Purchasing Power Between Countries", *Journal of the Royal StatisticalSociety. Series A*（*General*）, 121（1）, 97-99, 1958.

8. Gerardi，D.，"Sul Problema della Comparazione dei Poteri d'Aquisto della Valute"，Istituto di Statistica，Universita degli Studi diPadova, 1974.
9. Gini，C.，"Quelques Considerations au Sujet de la Construction des Nombres Indices des Prix et des Questions Analogues"，*Metron*，（1），3-162, 1924.
10. Heston，A.，Summers，R.，and Aten，B.，"Price Structures，the Quality Factors，and Chaining"，*Statistical Journal of the United Nations ECE*，18（1），77-101, 2001.
11. Hill，R. J.，"A Taxonomy of Multilateral Methods for Making International Comparisons of Prices and Quantities"，*Review of Incomeand Wealth*，43（1），49-69,1997.
12. Hill，R. J.，"Comparing Price Levels Across Countries Using Minimum-Spanning Trees"，*The Review of Economics and Statistics*，81（1），135-142, 1999.
13. Hill，R. J.，"Measuring Subsitution Bias in International Comparisons Based on Additive Purchasing Power Parity Methods"，EuropeanEconomic Review，44（1），145-162, 2000.
14. Himmelblau, D.，*Applied Nonlinear Programming*, McGraw-Hill,1972.
15. Ikle，D. M.，"A New Approach to the Index Number Problem"*The Quarterly Journal of Economics*，86（2），188-211, 1972.
16. Kravis，I. B.，Kenessey，Z.，and Heston，A.，et al.，*A system of International Comparisons of Gross Product and Purchasing Power*，UNSO，World Bank，The University of Pennsylvania, 1975.
17. Kravis，I. B.，Heston，A.，and Summers，R.，*International Comparisons of Real Product and Purchasing*

Estimating measurement errors in mixed-mode surveys using a Multitrait-Multierror Model

Joseph W. Sakshaug[1,2], Alexandru Cernat[3]
[1] University of Mannheim, Germany
[2] Institute for Employment Research, Germany
[3] University of Manchester, UK

Abstract
Mixed mode designs are becoming standard in the collection of survey data. Despite this, there are still unknowns regarding how the mode (e.g., Web) or mode design (e.g., sequential mixed mode) impact measurement error. Previous research has been limited by the confounding of selection and measurement mode effects and the investigation of only one type of measurement error at a time. In this paper, we use three waves of the Understanding Society Innovation Panel to investigate whether single mode versus sequential mixed mode and Web versus face-to-face modes have different impacts on measurement error. We make use of a quasi-experimental design that randomly allocated respondents to either a unimode face-to-face interview or a sequential mixed-mode (Web and face-to-face) design. Through this design, we implement a new multitrait-multierror model that estimates social desirability, acquiescence, and method effects simultaneously. The results show no differences in measurement error between single modes and mode designs with respect to acquiescence and method effect but some differences are found for social desirability. We discuss the practical implications of these findings and their possible causes in conclusion.

Keywords
Measurement error; mixed-mode surveys; social desirability bias; method effect; acquiescence bias

1. Introduction

Using multiple modes to collect data has become an increasingly common practice in survey research. In particular, the practice of deploying multiple modes of data collection in sequence is widely implemented in several large, policy-relevant surveys. Prominent examples include the U.S. American Community Survey, a cross-sectional survey which uses relatively inexpensive self-completion modes (mail and Web) in the initial phase of data collection, followed by more expensive interviewer modes (telephone and face-to-face) during the nonresponse follow-up phase (U.S. Census Bureau 2014). Another high-profile example is Understanding Society – the UK Household

Longitudinal Study (UKHLS), which is gradually moving away from a single-mode, computer-assisted personal interviewing (CAPI) design towards a sequential Web-CAPI mixed-mode design (Lynn 2017).

Studying measurement effects in mixed-mode surveys faces several challenges. For example, one of the challenges faced in many studies is separating selection effects from measurement effects. Mode selection effects are caused by different subgroups of respondents who have different likelihoods of completing the survey in a given mode. These mode-specific response propensities can affect the composition of the respondents answering in a given mode and, in turn, confound the investigation of measurement effects. Although previous studies have tried to analyze measurement effects in mixed-mode surveys, it is rare that they account for selection effects in their analysis (e.g. Allum et al. 2018; Heerwegh and Loosveldt 2011; Gordoni, Schmidt, and Gordoni 2012). A further limitation of previous studies is that methods of assessing and correcting for measurement errors typically focus on a single type of measurement error (e.g. social desirability) independently of other measurement error types (e.g. acquiescence, method effects) (Couper 2011). Assessing multiple errors simultaneously in a multivariate context has the potential to improve estimation of said errors as well as determine their relative contributions to the measurement accuracy for a given item. Such an approach is particularly useful for practitioners as it provides revealing information about how one can better allocate resources for minimizing multiple sources of measurement error in surveys.

In this article, we address these research gaps by using a quasi-experimental design implemented in multiple waves of a longitudinal mixed-mode (CAPI and Web) survey. Further, we use an innovative multitrait-multierror (MTME; Cernat and Oberski 2018; Cernat and Oberski in press) modelling approach to assess the relative magnitude of multiple types of measurement error simultaneously while accounting for selection effects. With the results of this study we aim to provide researchers with a better understanding of the contributions of different measurement error sources that can arise in mixed-mode survey designs.

The following research questions are addressed:

1. To what degree does using a mixed-mode approach compared to a single-mode approach lead to different measurement errors?
2. To what degree does CAPI versus Web surveys lead to different measurement errors?

2. Methodology

Multitrait-Multierror (MTME) Modelling Approach

The MTME approach was recently developed to deal with some of the inherit limitations of the MTMM approach (Cernat and Oberski 2018; Cernat and Oberski in press). The strength of the MTMM is the ability to estimate random error and method effects using a within-experimental design. However, the MTMM model does make a strong assumption regarding the absence of any other type of measurement error. This assumption may not hold in some cases, such as when measuring attitudes towards immigrants, where other factors, such as social desirability or acquiescence, might be present. In the MTME approach used here multiple potential sources of error are experimentally manipulated and modelled. In a sense the MTMM is a special type of MTME in which only the method is manipulated. For more details on designing MTME, the reader is referred to Cernat and Oberski (2018) and Cernat and Oberski (in press). Next, we describe how the MTME was implemented in the UKHLS-IP.

Experimental Design

In waves 7, 8 and 9 the UKHLS-IP implemented a MTME experiment that manipulated the wording and response scales of six questions in order to estimate: social desirability, acquiescence and method effects. It is these estimates of measurement error which are the focus of this paper.

The MTME experiment was implemented using six questions regarding attitudes towards immigrants. The design of the MTME experiment starts from the decision regarding the types of systematic errors one wants to estimate. In this case, these were social desirability, acquiescence, and method effects. For each type of systematic error, two possible manipulations were implemented. To estimate method effects, either a 2-point scale or an 11-point scale were used. To manipulate the direction of acquiescence either an agree-disagree format of the response scale or a reversed scale format (disagree-agree) was used. Finally, to impact the direction of social desirability the wording of the question was either positive (e.g., we should allow more immigrants) or negative (e.g., we should allow fewer immigrants). By combining these three dimensions, eight different ways of asking about attitudes towards immigration was developed for a given item/trait.

The implementation of the MTME is similar to the one used in earlier MTMM split ballot experiments (Saris, Satorra, and Coenders 2004, Saris and Gallhofer 2007). Each respondent was randomly assigned to receive the six items/traits twice, once at the beginning of the survey and once at the end. The order of the forms was randomized. Overall, the average time between the two sets of questions in the survey was 30 minutes.

Model Estimation

To estimate the MTME model, we use Bayesian Structural Equation Modelling with non-informative priors as implemented in Mplus 8.2. Trait 1 (T1) is measured using eight different wordings (W1-W8). These are the observed variables which measure an unobserved/latent variable T1. Additionally, there are the three types of systematic measurement error. These can be estimated either using an effect coding approach, where 0 is the average of the two conditions (which is unobserved), or a dummy coding approach, where 0 is the reference condition. In our model, we use both approaches. For social desirability (S) and acquiescence (A) we use an effect coding approach while for the method effect (M) we use a dummy coding approach (similar to using an MTMM-1; Eid et al. 2003). In order to identify the latent variables, different constraints for the loadings are used. These depend on the experimental design. All of the observed variables measure trait 1 so all of them have their loadings fixed to "+1". Similarly, wordings 3, 4, 7 and 8 are measured using the 11-point scale so they have "+1" in relationship with the method effect. For social desirability and acquiescence, we use "+1" or "-1" depending on the type of wording used. Finally, we restrict all of the means/thresholds of the observed variables and estimate the means of the latent variables. The full model also includes the latent variables T2-T6 as well as the observed variables measuring those questions.

The measurement error estimates from the MTME are used to compare the effect of mode (design) in UKHLS-IP. The following steps are used. We start by using the simple MTME model in each wave and compare the posterior distributions of the three systematic measurement errors by mode (CAPI vs. Web) and mode design (CAPI vs. Web-CAPI). Secondly, we develop regression models where mode (design) explains the three types of systematic errors (Research Question 1). Thirdly, we develop a model that explains the measurement error in MTME using both control variables and the mode of interview (Research Question 2). The model is shown below:

$$Y = \beta_0 + \beta_1 * mode\ (design) + \boldsymbol{\beta_2} * controls + \epsilon$$

where Y represents the three types of measurement error, $\beta 0$ is the intercept term, β_1 is the effect of the mode (design), and $\boldsymbol{\beta_2}$ is the vector of coefficients for the control variables.

As mentioned in the introduction, one of the biggest challenges in researching mode effects is the confounding of mode selection and mode measurement effects. Here we deal with this problem in two different ways. Because mode design (CAPI vs. Web-CAPI) is randomized one can use a simple regression model to explain measurement error. The randomization should control for the majority of selection confounding between the two mode

designs.[16] Such a comparison is not possible between Web and CAPI as people can self-select into one mode within the mixed mode design. To partially account for this confounding, the following control variables are added to the model: age, gender, having a partner, being white British, living in rural area, having a degree, and being employed. The effects of mode are investigated only after controlling for these potential confounders.

3. Result

RQ1: To what degree does using a mixed-mode approach compared to a single-mode approach lead to different measurement errors?

Next, we investigate the differences in measurement error by mode design using the aforementioned regression approach. The regression slopes of the mode design on measurement error show the impact on their mean while the R2 indicates how much of the variation in measurement error is explained by the mode design. Similar to what was observed in the descriptive analysis, the main consistent difference is in social desirability (Table 1). The mixed-mode respondents have more extreme levels of social desirability compared to respondents allocated to the single-mode design. The mode design explains around 2% of the variation in social desirability, indicating that mode design is only a small part of the mechanisms behind this type of measurement error.

We also observe that the mixed-mode design has slightly lower means for acquiescence for the wave 9 respondents,[17] but that is not true for waves 7 and 8. Finally, we see that there are no differences in the method effect between mode designs at any point in time.

Table 1. Regression coefficient and R^2 of mixed mode vs. single mode by wave

Wave	ME	Est.	Lower C.I.	Upper C.I.	R^2
	Social desirability	**-0.22**	**-0.43**	**-0.08**	**2.5**
7	Method	-0.05	-0.22	0.11	0.1
	Acquiescence	0.01	-0.09	0.12	0.1
	Social desirability	**-0.18**	**-0.36**	**-0.06**	**1.4**
8	Method	-0.11	-0.26	0.04	0.4
	Acquiescence	-0.03	-0.17	0.11	0.1
	Social desirability	**-0.37**	**-0.66**	**-0.17**	**2.7**
9	Method	-0.02	-0.19	0.15	0.1
	Acquiescence	**-0.12**	**-0.25**	**0.00**	**0.8**

[16]As a sensitivity analysis we reran the models using control variables which resulted in similar findings.

[17]The effect of acquiescence is no longer statistically significant after adding the control variables to the model.

RQ2: To what degree does CAPI versus Web surveys lead to different measurement errors?

Lastly, we make a direct comparison of Web and CAPI responses in the three waves. As mentioned in the methods section, due to the confounding effect of selection, control variables are used. Here, we investigate the impact of mode of interview after controlling for: age, gender, having a partner, being white British, living in rural area, having a degree, and being employed (Table 2).

The results are similar to those seen previously. The main difference between modes is evident for social desirability. The expected mean for CAPI respondents on the social desirability variable is lower than for the Web respondents. This is consistent in all three waves with an R^2 ranging from around 1% to approximately 4%. No differences in acquiescence and method effect by mode of interview are present.

Table 2. Regression coefficient and R2 of CAPI vs Web by wave (with control variables)

Wave	ME	Est.	Lower C.I.	Upper C.I.	R^2 extra
	Social desirability	**0.48**	**0.29**	**0.69**	**3.8**
7	Method	0.00	-0.12	0.26	0.3
	Acquiescence	-0.01	-0.12	0.10	0.1
	Social desirability	**0.51**	**0.23**	**0.81**	**1.1**
8	Method	0.04	-0.15	0.22	0.3
	Acquiescence	0.01	-0.19	0.20	0
	Social desirability	**0.77**	**0.45**	**1.11**	**1.7**
9	Method	-0.05	-0.23	0.13	0.4
	Acquiescence	0.08	-0.08	0.24	0.8

4. Discussion and Conclusion

This paper investigated the impact of mode and mode design on measurement error using a combination of experimental designs and statistical modelling. We leveraged an experimental design that randomly allocated respondents to either a single-mode CAPI survey or a sequential mixed-mode Web-CAPI design, as well as the implementation of a multitrait-multierror model with three types of systematic measurement errors: social desirability, acquiescence, and method effect. The experimental mode design enables us to control for the confounding of measurement and selection while the MTME enables us to estimate multiple types of measurement error simultaneously. Finally, the use of three waves of longitudinal data allowed us to validate the findings and assess how stable they are in time.

Overall, we find small differences in measurement error across mode (designs). The descriptive analysis showed that the variance of the systematic

errors is similar across mode (designs) while some differences were present in the means for social desirability. These findings were supported using regression models. Social desirability was systematically different by mode (design). Although the results show that there are mode differences in measures of social desirability, this explains only a small amount of variance. Additionally, we found no mode (design) differences with respect to method effects and acquiescence, which goes against some of the previous research in this area (e.g., De Leeuw 1992; Revilla 2010), but is consistent with other studies (Fricker et al. 2005; Revilla 2012; Revilla 2015).

The most surprising finding was the direction of the mode effect on social desirability. While most of the research in this area has shown that social desirability effects are more prominent in interviewer modes (e.g. face-to-face) compared to self-administered modes (such as Web), we find that changing the wording of the item in question has a larger impact on the mean of the observed responses in Web than in CAPI. A post-hoc explanation could be that wording changes are more salient in self-administered modes and that the experimental manipulation does not only encompass social-desirability, but also other types of measurement error.

As all research studies do, this one has some limitations. The first limitation is our approach to estimating differences in measurement errors by mode makes the assumption that the measurement model is the same for the two mode groups. A way to free this assumption could be to run a multi-group model by mode (design). Unfortunately, this approach does not work with the available data due to sample size limitations. Additionally, it would be ideal to use such an approach in different countries, alternative mode designs, and different manipulations of the MTME. Investigating such variations is a topic for future work.

That being said, this paper contributes to the mixed-mode literature through the use of experimental designs both in dealing with the confounding of measurement and selection effects in mode (designs) as well as measurement confounding (i.e. looking at one measurement error at a time and ignoring the influence of the others) which is common in the survey literature. The research shows no differences between single-mode and mixed-mode designs or between Web and CAPI modes with respect to acquiescence and method effect (impact of response scale) – a promising result for survey practice where mixed-mode designs are increasingly common. On the other hand, practitioners should be aware of the potential that responses collected in a self-administered mode are more likely to be influenced by the wording of the question stem than in an interviewer mode, which may affect estimates of social desirability bias in multitrait experiments.

References

1. Allum, N., Conrad, F., and Wenz, A. (2018). Consequences of Midstream Mode Switching in a Panel Survey. *Survey Research Methods*, 12(1), 43-58.
2. Cernat, A., and Oberski, D. (2018). Estimating Stochastic Survey Response Errors Using the Multitrait-Multierror Model. NCRM Working Paper, National Centre for Research Methods. <http://eprints.ncrm.ac.uk/4156/>.
3. Cernat, A., and Oberski, D. L. (*In Press*). Extending the Within-Persons Experimental Design: The Multitrait-Multierror (MTME) Approach. In P. J. Lavrakas (Ed.), *Experimental Methods in Survey Research*. New York: John Wiley & Sons.
4. Couper, M.P. (2011). The Future of Modes of Data Collection. *Public Opinion Quarterly*, 75(5), 889-908.
5. De Leeuw, E. (1992). *Data Quality in Mail, Telephone and Face to Face Surveys*. Amsterdam: TT-Publ.
6. Eid, M., Lischetzke, T., Nussbeck, F.W., and Trierweiler, L.I. (2003). Separating Trait Effects from Trait-Specific Method Effects in Multitrait-Multimethod Models: A Multiple-Indicator CT-C (M-1) Model. *Psychological Methods*, *8*(1), 38-60.
7. Fricker, S., Galesic, M., Tourangeau, R., and Yan, T. (2005). An Experimental Comparison of Web and Telephone Surveys. *Public Opinion Quarterly*, 69(3), 370-392.
8. Gordoni, G., Schmidt, P., and Gordoni, Y. (2012). Measurement Invariance Across Face-to-Face and Telephone Modes: The Case of Minority-Status Collectivistic-Oriented Groups. *International Journal of Public Opinion Research*, 24, 185-207.
9. Heerwegh, D., and Loosveldt, G. (2011). Assessing Mode Effects in a National Crime Victimization Survey Using Structural Equation Models: Social Desirability Bias and Acquiescence. *Journal of Official Statistics*, 27, 49-63.
10. Lynn, P. (2017). Pushing Household Panel Survey Participant from CAPI to Web. Presented at the 28th International Workshop on Household Survey Nonresponse, Utrecht, August/September.
11. Revilla, M. (2012). Impact of the Mode of Data Collection on the Quality of Answers to Survey Questions Depending on Respondent Characteristics. *Bulletin de Methodologie Sociologique*, 116, 44-60.
12. Revilla, M. (2015). Comparison of the Quality Estimates in a Mixed-Mode and a Unimode Design: An Experiment from the European Social Survey. *Quality and Quantity*, 49, 1219-1238.
13. Saris, W.E., and Gallhofer, I.N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (Vol. 548). John Wiley & Sons.

14. Saris, W., Satorra, A., and Coenders, G. (2004). A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology*, 34(1), 311–347.

Improving statistical literacy in Albania, the role of the National Statistical Institute

Klajd Shuka, Alban Cela
Institute of Statistics Albania

Abstract

We are living in times where the statistical offices need to adapt to the changes in economy, society and technology, to keep pace with time and continue to play their role as provider of official statistics. In this respect NSOs face several challenges in providing up to date quality statistics that meet user needs. In the continuous work towards improving efficiency and effectiveness new technological and organizational approaches are needed. Staff and user satisfaction surveys are tools used by National Statistical Institute in Albania to monitor their needs and improve the statistical literacy among users and producer of official statistical. A desired level of statistical capabilities is needed not only for the staff directly involved in the statistical production process but also in increasing statistical literacy among public administration and general public. In this paper human resource effective management and training, in reaching the statistical offices goals towards modernization of statistical production, will be highlighted. Institute of Statistics work in improving statistical literacy among main actors using official statistics is explained. The cooperation essential for improving statistical literacy are pointed out. The initiatives mention in this paper are expected to improve the Statistical Literacy in Albania.

Keywords

Literacy; Users; Human resources; Capacity building

1. Introduction

Staff capacities of every statistical office are one of the crucial aspects that need to be taken into consideration from the high management as one of the most important assets of an institution. The statistical production process is related directly to the staff knowledge and this knowledge has a direct impact on the total quality of the official statistics produced. The successful knowledge management and training of INSTAT staff is a very important part of the training strategy. The main objectives of INSTAT training strategy are:

- Organizing trainings focusing on current priorities such as: Statistical Quality; Use of administrative resources; Basic knowledge on statistics for public administration;

- Organizing trainings in a unified system;
- Organizing vocational training in addition to compulsory and administrative training courses;
- Focusing on the work currently carried out by INSTAT, developing practical skills and adding to the theoretical knowledge;
- Organizing quality trainings both by the content of training topics and by the level of participants in the training.

The Training Strategy aims to reach a desired level of statistical abilities in implementing the European Code of Practice and achieving the production of high quality statistical product by following international standards and methodologies. Based on this strategy training programs have been developed and for three main groups: Producers of official statistics, public administration and general public.

Human resources development in INSTAT

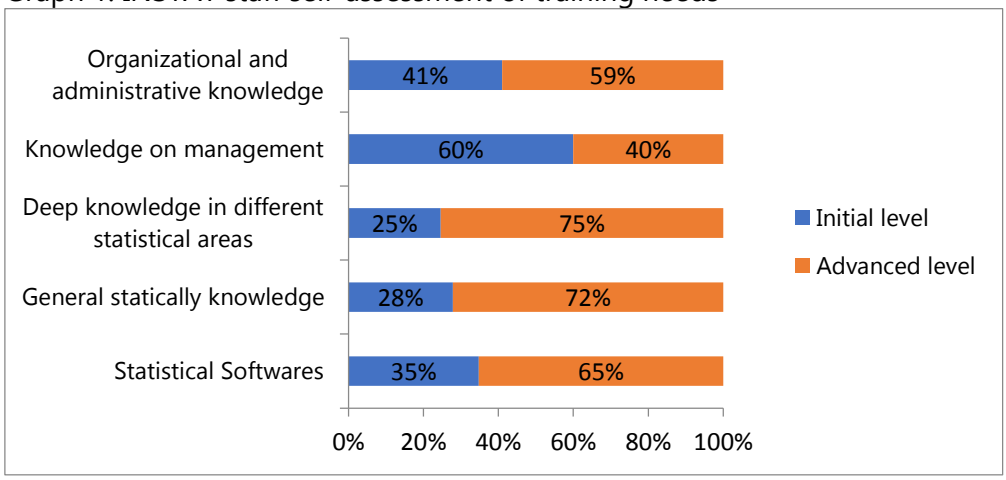
The first statistical office in Albania was opened in 1924, as part of the Ministry of Public Affairs and Agriculture. The statistical service was institutionalized for the first time in 1940. The National Statistics Institute of Albania (INSTAT) has been established in 1993. The activity of the first years of the statistical service has been very limited in statistical production and until the 90's under the communist regime. From 1993 until 2017 in INSTAT there were no dedicated human resources sector, but they were part of the Legal Issues, Procurements and Human Resources sector.

In 2017 INSTAT prepared the development strategy 2017-2030 with four main strategic objectives. One of the objectives of INSTAT's development strategy is directly related to the professional and organizational improvement of the National Statistical System by creating a motivational work environment and developing the professional capacity of employees. The newly created human resources sector started to develop a proactive human resources management approach by defining core competencies of INSTAT staff and assessing training needs. The human resources are a fundamental part of a statistical office and their role in the production of official statistics is unique.

INSTAT core competencies are considered: the general knowledge in statistics, the knowledge of methodologies in producing official statistics, specialised knowledge in different statistical areas and in statistical systems and the knowledge of the needs for official statistics in a changing environment.

Based on these competencies a training assessment has been conducted, where staff has self-evaluated the needs for training in the following areas: general statically knowledge, deep knowledge in different statistical areas, statistical software, knowledge on management and organisational and administrative knowledge.

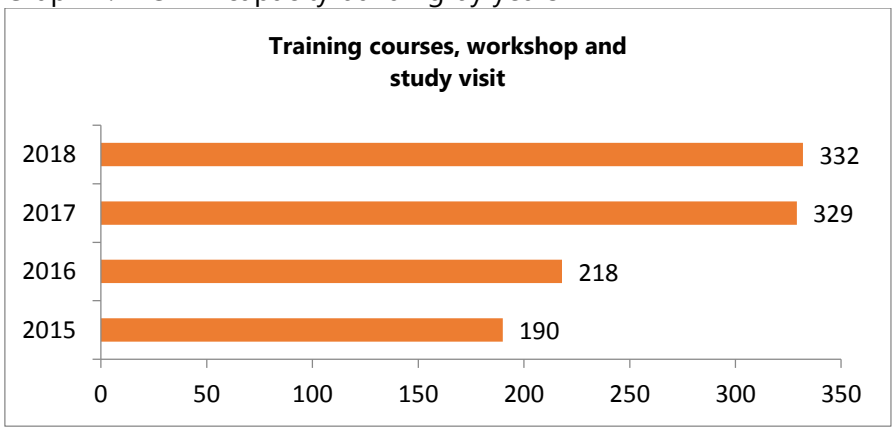
Graph 1: INSTAT staff self-assessment of training needs



The advanced training on deep knowledge in different statistical areas was the one which has been the most needed training, based on employee's replies.

A special attention is given to the training of the new staff. For this category the skills development is achieved based not only on training but also in mentoring and coaching. Training curricula is developed for the newcomers for giving knowledges on the overall institution, methodologies, standards and techniques. While the more in depth development is done through specialised training courses and mentoring. Based on the law On Public administration, for each newcomer, a mentor is appointed, usually one senior expert of the same category.

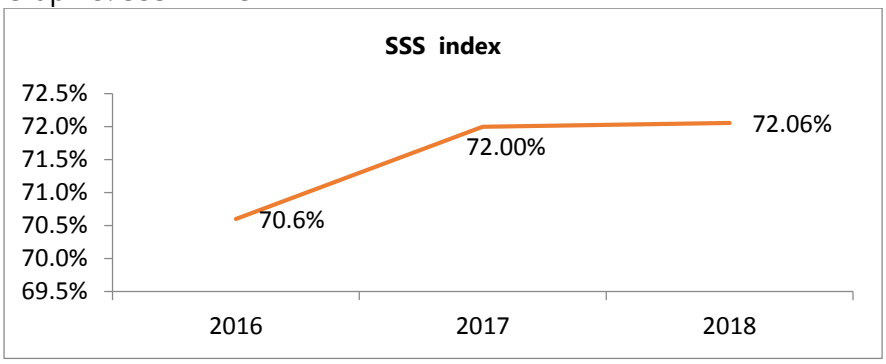
Graph 2: INSTAT capacity building by years



During years the capacity building of INSTAT staff has improved, in average one employee have participated at least in 3 training courses, workshops or study visits.

In December 2016 INSTAT conducted for the first time the Staff Satisfaction Survey (SSS), with the main objective to analyse attitudes and perceptions of the INSTAT staff and identify issues that require attention from managers. The measurement scale was from 1 to 5, and over the years it has been improved steadily. The graph below presents the staff satisfaction index for the last three years.

Graph 3: SSS in INSTAT

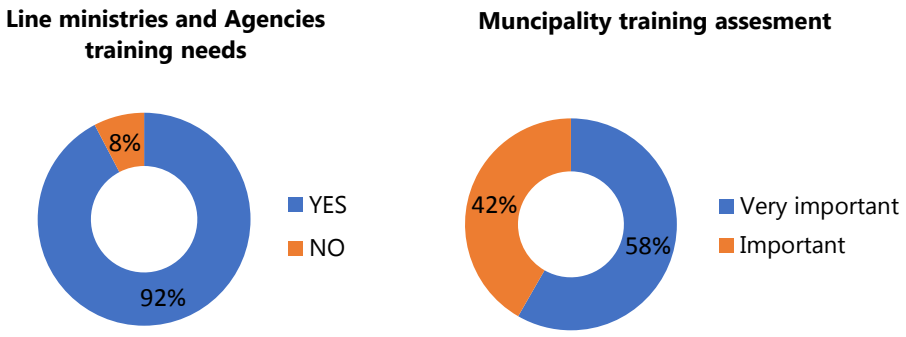


Statistical literacy in public administration

INSTAT is committed to communicate the overall statistical culture and to provide the development of the skills, knowledge and expertise needed to fulfil its mission in the whole group of users. These capabilities cannot be developed in Albanian education system due to the lack of relevant specialties such as statistics, econometric, etc. in universities. Currently, the education system provides general knowledge on statistics and economic statistics in particular, but does not provide specialized knowledge for different areas of statistics.

To assess the availability of administrative data and the needs for improving statistical literacy in public administration INSTAT implemented two surveys, one on Central public institutions, directed to line ministries and Agencies and another to Municipalities; to identify the data being produced by them, which can be used for producing qualitative official statistics. Part of the focus of the questionnaires was to identify training needs in capacity building and systems regarding official statistics.

Graph 4: Training needs in public administration



Approximately 92% of the central public institutions were positive to participate in INSTAT training programme. On the other hand, from the 61 municipalities in Albania, all were positive to participate in INSTAT training programmes and 58% of them, evaluate as very important these kinds of initiatives.

The new Law on Official Statistics No 17, dated 05/04/2018, has strengthened INSTAT role as a leader of the statistical system. The law stipulates the creation of the Statistical Training Centre (STC) as a key instrument to provide the necessary training infrastructure for different target groups. Taking into consideration all training needs assessed and the growing demand for official statistics, a training programme was set up. During 2018, nine training modules have been organised in the STC, six of them were directed to INSTAT staff and three directed to central public institutions.

Education of statistics for schools and general public

INSTAT conducted for the first time User Satisfaction Survey (USS) in 2017 with the main scope to understand user expectations and needs for official statistical data. On the indicator from the Code of Practice is the implementation of USS. This survey is conducted every year and its results are monitored regularly to fulfil the expectations of different type of users. The main indicator, the satisfaction index has increased by around 10.20% on the last two years.

In order to ensure the official statistics quality and their reliability, the official statistics procedures, methodologies and concepts must be understood by the general public. Statistical literacy can not only be left in the hands of schools and universities, it is also a responsibility of the statistical office to improve statistical literacy among users. Statistical literacy and reasoning as Lancaster (2011) concluded will one day be as necessary for efficient citizenship as the ability to read and write. For this reason, especially

in the last years NSOs, are dedicating time and effort for this area. INSTAT has been working in different directions to reach this goal such as: the cooperation with academia and universities, the Master on Official Statistics, statistical events in schools and e-learning. During last two years five Memoranda of Understanding (MoU) have been signed with different universities in Albania and abroad. One of the objectives of these MoUs is to collaborate in capacity building for the use and interpretation of official statistics. In this framework organised open lectures help students increase the understanding of the use of statistics in their researches for masters or PHDs.

The Institute of Statistics, in cooperation with the Faculty of Economics, University of Tirana, have established for the first time the opening of a scientific master program dedicated to official statistics. This study program is based on the best experiences of other universities offering similar study programs. This master addresses students who want to have specialized knowledge in the basics and principles of official statistics, as well as enable professional development in the field of statistics.

Over the years statistical education has been limited in primary and secondary schools with a limited focus on the measure of central tendency and statistical theory. To improve statistical literacy in obligatory education years and in secondary schools INSTAT in the Statistical Literacy Strategy 2018-2021 has dedicated a special part on events organised in schools:

- School Corners - Special corners at schools where small publications of INSTAT as well as leaflets are presented where pupils and students could increase their nowledge on official statistics;
- School Visits - Direct interaction with students in INSTAT with organised visits of different schools and different grades;
- E-learning tools - innovative tools have been developed, which speak to the new generation with the language of new technology, such as digital interactive publications with much more navigation power than physical books, interactive infographics for facts and figures on different areas, short videos explaining

2. Discussion and Conclusion

Know-how is one of the main resources for modernisation, efficiency and effectiveness, not only in the field of official statistics but also nation-wide. Several initiatives need to be taken in order to monitor and improve core competencies in the statistical offices.

Statistical Training Centre (STC) can be used as key instrument to provide the necessary training infrastructure for different target groups.
Based on the findings of SSS appropriate action plans have to be prepared on yearly bases, to improve staff satisfaction.

The level of education and knowledge, are crucial for official statistics production and understanding, this cannot be let only on the hands of the education system. National statistical offices need to play their role in order to generate statistical literacy not only for their staff but also for data producers and users.

In a changing world new partnerships with users are a must. Partnerships with academia, multimode data custodians and general public are important to ensure a proper level of statistical literacy, because statistics belong to users and are not a property of statistical offices.

References

1. Law on Official Statistics No 17, dated 05/04/2018 http://instat.gov.al/media/3972/law-no17-2018-on-official-statistics.pdf
2. Learning and Development in Challenging Times, Anne Kofoed and Mats Olsson, Eurostat, 2012
3. Lancaster, G.A. (2011): How statistical literacy, official statistics and self-directed learning shaped social enquiry in the 19th and early 20th centuries. Statistical Journal of the IAOS, 27, 99–111

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

Composite indicator of Food insufficiency

Amal Mansouri
High Commission for Planning, Rabat, Morocco

Abstract

In this paper, we constructed a new composite indicator for food insufficiency based on co-movement of process, state and outcome food indicators, mainly provided by the Food and Agricultural Organization and World Bank. We used the dynamic factor modelling approach to estimate the common component that capture the food insufficiency on a sample of four countries, over the period 2000-2016. The main contribution of this index is in providing insight about the evolution of food situation by country, taking into account the various aspect of food insufficiency. The methodology used has also enabled to establish forecasts of composite indicator of food insufficiency as well as its main determinants.

Keywords

Food insufficiency; unobserved component model; Kalman filte

JEL Classifications:

C43, O13, O57, Q18

1. Introduction

Eradicate hunger has been a focus of international concern because of the successive food crises that have affected many developing countries over the last 40 years. Hunger concept refers to a situation in which a person's food intake is insufficient to cover basic energy needs. When persisted over a long period, this insufficiency could lead to undernourishment, which is more sensitive to adult productivity and to cognitive impairment and stunting of children. Undernourishment is also a factor in triggering social crises and political conflicts.

Understanding of the hunger issue has led to much literature that has been engaged in quantifying the different dimensions of food security. The concept of food security includes a range of factors that affect the development of undernourishment or hunger, not just the supply side. FAO has identified four dimensions of food security: availability, accessibility, use and stability. Therefore, several studies aimed to summarize the four dimensions of food security, with an equal weight aggregation of variables or by using factor analysis techniques. Composite indicators dealing with other aspects of hunger were also produced and published, the most prominent of which is the

FAO indicator of prevalence of undernourishment and IFPRI's global hunger index.

Following this current approaches, we propose to construct a new composite indicator of food insufficiency that highlights the multidimensional aspect of hunger. Food insufficiency is closely linked to factors and effects related to the evolution of hunger in the world. The main contribution of this indicator is to smooth a great part of the volatility of process, state and outcome food indicators and to capture the co-movement that characterize their fluctuations. It will enriches the panoply of monitoring indicators of the food situation and could be used to inform public policies, thanks in particular to the end-of-series forecasts.

The remainder of this paper is organized as follows. The next section focuses on specifying the different aspects of food insufficiency. In the third section, the construction methodology of the composite indicator of food insufficiency is described. The empirical results and interpretation grids are presented in the final section.

2. Food insufficiency framework

In general, food insufficiency can be understood through three types of variables: process, state and outcome indicators.

For the first type of indicators, which describe the insufficiency of the underlying factors of hunger, various choices can be made from the list of food security dimensions proposed by FAO (2013). The simplest choice would be to use variables that capture the factors and conditions that influence food accessibility and supply, such as the average value of food production, per capita income, dependency ratio on grain imports, the rate of inflation and political stability. The second type of indicators describes the state of food insufficiency, referring to the situation during which a person does not consume over a prolonged period of time, a food energy intake sufficient to cover his minimum needs to lead a healthy life.

The third type of indicators is largely composed of anthropometric variables that trace the consequences of food insufficiency on the state of health. These measures, generally performed on children under 5 years, are considered as reliable approximations of the nutritional status of the entire population. For example, insufficient short-term dietary intake could result in wasting (children being too lean compared to their height), while stunting (children too small for their age) could often be caused by insufficient food intake over a prolonged period, by serial infections or repeated episodes of acute under nutrition.

A more consistent measure of food insufficiency may be based on the synthesis of the common information contained in the 3 types of indicators. The construction of a composite indicator imposes the choice of the type of

data normalization, the selection of the weighting and aggregation method. In the view of arbitrary weighting limits, we opted for a factor analysis that provides weighting coefficients at the end of a statistical process.

3. Empirical methodology

The methodology for developing a composite indicator of food insufficiency falls under the framework of the dynamic factor analysis technique implemented in numerous studies, in particular Stock and Watson (1989) and Doz and Lenglart (1999).

In dynamic factor analysis, each indicator is written as the sum of a component that follows a common dynamic to the other variables and a term specific to the variable considered. This common component represents the general state of food insufficiency and its estimated value constitutes the composite indicator of food insufficiency. Borrowing from Stock and Watson (1989), we consider the following bivariate dynamic factor model:

$$\begin{cases} Y_t = \tilde{\beta} + \tilde{\Theta}(L)F_t + \tilde{u}_t \\ \tilde{D}(L)F_t = \tilde{\varepsilon}_t \\ \tilde{\Psi}(L)u_t = \tilde{\eta}_t \end{cases} \qquad (I)$$

Where Yt the vector of the observed variables which will make up the composite indicator, Ft represents the composite indicator of food insufficiency to be estimated, L is a delay operator, Θ(L), Ψ (L) and D(L) are delay polynomials of dimension k, q and p.

Taking into account the problem of non-stationary series, the estimation of the system will made in first difference and its simplified formulation is as follows:

$$\begin{cases} \Delta y_{it} = \beta_i + \gamma_i(L)\Delta F_t + e_{it} \\ \Delta f_t = \phi_1 \Delta f_{t-1} + \phi_2 \Delta f_{t-2} + w_t \\ e_{it} = \psi_{i1} e_{it-1} + \psi_{i2} e_{it-2} + \varepsilon_{it} \end{cases} \qquad (II)$$

Where $\Delta y_{it} = \Delta Y_{it} - \Delta \bar{Y}_{it}$ and $\Delta f_t = \Delta F_t - \delta$

In model (II), Δft and the error terms are supposed to follow an autoregressive order process 2 (p=2). This model admits a representation called linear space-state model, which is written as follows:

$$\Delta y_t = C\Delta z_t \text{ space equation)}$$
$$z_{t+1} = Az_t + B\eta_t \ (state\ equation)$$

With : $\Delta y_t = (\Delta Y_{it})'$ and $Z_t = (\Delta F_t, \Delta F_{t-1}, \Delta e_{it}, \Delta e_{it-1})'$, A, B, C, D are matrix

The resolution of this system was conducted through the Kalman filter estimation, composed of two procedures. The first procedure is a recursive process that filter the optimal estimate for the state variables using information available in the recent past (t-1), by minimizing the forecast error

by maximum likelihood. The second procedure smoothed the obtained estimate based on the information available over the whole sample period.

4. Empirical analysis

4.1 Data sources

The study was conducted over the period 2000-2016. All series come from the FAO database with the exception of consumer prices and per capita income that were collected from the World Bank database. Figure 1 summarizes the set of indicators and specifies their sources and dimensions, referring to the classification used in the state of food insecurity report (FAO, 2013). Data collection concerned four countries, particularly Bangladesh, Chile, Jamaica and Vietnam. The choice of countries was essentially driven by the availability of data. Anthropometric variables are not published on continuous data ranges for most developing countries.

Table 1: Main indicators of food insufficiency

Indicators	Source	Dimensions	typology
Average value of food production	FAO	Availability	Process indicators
Density of the road network		Access	
income per capita	WB	Economic access	
GDP at current prices $	FAO		
Inflation	WB		
Access to improved water sources	FAO	use	
Dependency ratio on cereal imports Political stability		Vulnerability Shock	
Prevalence of undernourishment	FAO	Access	State indicators
Prevalence of food insufficiency		Access	
Percentage of children under 5 years of age who are stunted	FAO	Use	Outcome indicators
Percentage of children under 5 years affected by wasting			

Source: FAO, World Bank, author development

In order to capture the different aspects of food insufficiency and to ensure an interchanged reading of the results, we have opted, in a first step, for the synthesis of the process indicators through a principal components analysis. Of the eight process variables studied, two were excluded, especially the density of the road network and access to improved water sources while six process variables had a strong correlation with an MSA (Kaiser's measure of Sampling Adequacy) greater than 0.74. Hence, the idea of summarizing the common information contained in the six variables into a single composite index. As result, the first factor, explaining 74.7%, 76.9 %, 60.5% and 74.9% of the total inertia of process variables in Bangladesh, Chile, Jamaica and Vietnam, is thenconsidered as a composite indicator of food insufficiency factors.

In the second stage, we analyze the coincident and delayed correlation of the composite indicator of food insufficiency factors with the prevalence of undernourishment, percentages of wasting and stunting among children under 5 years. This exercise, carried out at the four countries, showed a strong correlation between three main variables: composite indicator of food insufficiency factors, the index of prevalence of undernourishment and the percentage of children stunted.

The results of the Dickey and Fuller (Augmented Dickey Fuller) stationary test conducted on the three variables selected attest to the presence of a unit root, particularly in the prevalence of undernourishment and percentage of stunted children (model II). Parameter estimates of the state-space are reported in Table 2

Table 2: Parameter estimates

Countries	Ft	Process index (Y1)	Undernourishment index (Y2)	% Stunted children (Y3)
Bangladesh	ϕ1= 0,21 (0,22) ϕ2= 0,39 (0,23)	Ψ1 = 0,25 (0,24) Ψ2 = -0,15(0,07) δ1 = -0,87 (0,5) δ2 = -0,42 (0,26)	Ψ1 = 0,78 (0,21) Ψ2 = 0,18(0,08) δ1 = 0,4 (0,24) δ2 = -0,57 (0,36)	Ψ1 = 1,73 (0,7) Ψ2 = -0,54 (0,26) δ1 = -0,53 (0,27) δ2 = 0,01 (0,16)
Chili	ϕ1= 0,94 (0,25) ϕ2=-0,16 (0,21)	Ψ1 = -0,42 (0,06) Ψ2 = -0,17 (0,1) δ1 = 0,18 (0,3) δ2 = -0,8 (0,31)	Ψ1 =1,03 (0,13) Ψ2 = 0,29 (0,08) δ1 = 0,79 (0,31) δ2 = -0,18 (0,33)	Ψ1 = 0,32 (0,15) Ψ2 = -0,14 (0,09) δ1 = - 0,03 (0,14) δ2 = -0,26 (0,25)
Jamaique	ϕ1=-0,13 (0,29) ϕ2= 0,16 (0,2)	Ψ1 = -0,8 (0,15) Ψ2 = -0,17 (0,09) δ1 = -0,05 (0,17) δ2 = 0,48 (0,24)	Ψ1 = 0,1(0,06) Ψ2 = 0,17 (0,07) δ1 = 0,65 (0,34) δ2 = -0,18 (0,7)	Ψ1 = 1,8 (0,2) δ1 = 0,24 (0,25) δ2 = 0,52 (0,28)
Veitnam	ϕ1= 0,33 (0,1) ϕ2= 0,19 (0,13)	Ψ1 = -1,08 (0,3) Ψ2 = 0,49 (0,2) δ1 = -0,4 (0,3) δ2 = 0,38 (0,2)	Ψ1 = 0,42 (0,14) Ψ2 = -0,12 (0,03) δ1 = 0,33 (0,22) δ2 =-0,37 (0,29)	Ψ1 = 1,6 (0,43) δ1 = -0,13 (0,26) δ2 = - 0,4(0,22)

The standard deviations of the estimated parameters are in parentheses.
Sources : author's calculation

The model seems to fit the data quite well, especially for Vietnam and Chile data. The significance of a large part of the parameters is considered satisfactory at 5% level. The estimated autoregressive coefficients are positive and significant for the first delay. This suggests a large persistence of fluctuations in the food insufficiency indicator. In order to check the adequacy of the model specification, we analyze the innovations disturbances. The results providing by the Ljung–Box tests for residual autocorrelation and the correlogram profile lead us to not reject at the 5% the hypothesis of uncorrelated distributed residuals for all countries.

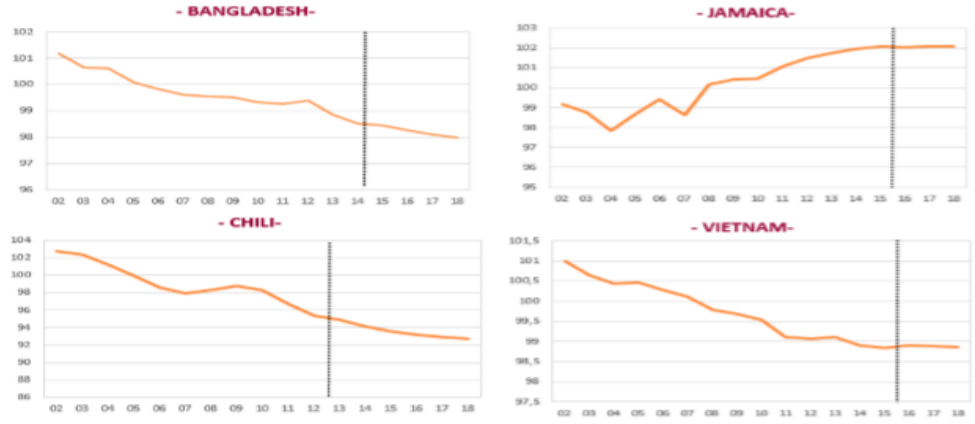
5. Relevance of interpretations

The composite indicator of food insufficiency corresponds to the value of Ft: a decrease of the indicator signifies an improvement of the food situation, while an increase reflects a deterioration of the food situation and a shift compared to the SDG targets. The indicator analysis provides a new look at food trends. In particular:

The comparison between the composite indicator of food insufficiency and its three determinants revealed a heterogeneity between the four countries. While the percentage of stunted children is extremely closed to the food insufficiency indicator in Bangladesh and Vietnam, the prevalence of undernourishment and supply factors punctuate more the composite indicator food insufficiency indicator in Chile and Jamaica.

Over the study period, we can read in particular the downward trend of indicators of food insufficiency in Bangladesh, Chile and Vietnam. This trend mainly reflects the progress made in terms of supply and its effective use by the population. However, while overall, the food situation has improved; non-negligible differences have been detected between countries. Chile is the country where the food deficit has been reduced the most, though with modest progress at the end of the period. Bangladesh and Vietnam have made similar progress. In Jamaica, the food situation has deteriorated with the reduction of supply, including food production and the increase in the share of undernourishment and stunted children. These results, which provide information on the evolution dynamics of food insufficiency, contrast with the conclusions reached by a univariate analysis of the indicators of undernourishment or undernutrition, which ranks Jamaica at a high level of food performance than Bangladesh and Vietnam.

Figure 1: composite indicators of food insufficiency



Notes: the composite indicator of food insufficiency has been normalized to 100 in 2001. Forecast after the dotted line

Another aspect of the analysis is the evolution of the specific components of the three variables used as a basis for the construction of the composite

indicator of food insufficiency. When the specific information is positive, this means that the information provided by the variable is considered more pessimistic than the composite indicator of food insufficiency suggests. This is notably the case of the indicator of stunted children, whose specific information was positive during the last 4 years of the study period, indicating the existence of a surplus of under nutrition, which is not related to undernourishment or process composite indicator (Chile and Jamaica).

The new Food Insufficiency Indicator can also be used to predict undernourishment and stunting among children under 5, published by FAO. The indicator specification, estimated from an autoregressive model, generates forecasts from 1 to 3 years that could be integrated to project undernourishment or stunted children (figure 1).

6. Conclusion

In this paper, we have developed a composite indicator of food insufficiency using the Stock and Watson methodology (1989). This indicator complements the arsenal of composite indicators of food situation, concentrating the common information provided by eight indicators tracing the shortcomings in supply, food and nutrition. It thus provides insight into the success of policies to fight hunger and makes also it possible to predict the evolution of traditional measures of undernourishment or stunted children, thanks to its autodynamic specification.

However, it should be noted that the length of the data sets did not allow for more appropriate stability test and diagnostic measures, in order to attest to the quality of the forecasts. The methodological approach adopted, although providing a multidimensional measure of food insufficiency, remains however specific to the country studied and not comparable territorially. This question could be avoided through an ascending hierarchical classification based on Ward's method. Another approach would consist in aggregating the three aspects of the food insufficiency, by retaining a matrix of three variables most representative of each aspect and the most correlated with the others. The aggregation of these variables will be determined once their weights are calculated in the structure of the composite indicator of food insufficiency indicator, after the exercise of simulating the effects of unit shocks on each of the variables from the Kalman filter. Such assessment should of course trigger further treatments, which could be the topic for future research.

References

1. Comité de la sécurité alimentaire mondiale (2011). Mesurer l'insécurité alimentaire : des concepts et des indicateurs pertinents pour l'élaboration de politiques fondées sur des données probantes, FAO.
2. Doz C. and Lenglart F. (1999). Analyse factorielle dynamique : test du nombre de facteurs, estimation et application à l'enquête de conjoncture dans l'industrie », Annales d'Économie et de Statistique, n° 54, pp. 91-127
3. FAO (2013, 2015, 2017). L'état de l'insécurité alimentaire dans le monde.
4. Lemoine M, Pelgrin F (2003). Introduction aux modèles Espace-état et au filtre de Kalman, OFCE.
5. John Hoddinott (1999). Choosing outcome indicators of household food security, International Food policy Research Institute.
6. Rokhaya Diagne (2014). Sécurité alimentaire et libéralisation agricole, Université de Nice Sophia Antipolis.
7. James H . Stock and Mark W. Watson (1989). New Indexes of Coincident and Leading Economic Indicators, NBER Macroeconomics Annual 1989, Volume 4.
8. Vincent Bodart, Bertrand Candelon (2000). Appréhender la conjoncture à l'aide de la méthode de Stock- Watson : une application à l'économie belge, Économie & prévision, n°146, pp. 141-153.

Monitoring Population Strategies in GCC: opportunities and challenges

Ali Sulaiman Al Flaiti
Gulf Cooperation council statistics

Abstract

Governments all around the world take an active interest in population policy. Demographic indicators are critical to help decision makers assess if these policies are on track. The six countries of the Gulf Cooperation Council (GCC) adopted a common population strategy in 1998 (GCC Secretariat, 1998). The policy was established to better balance population growth with the longer-term development aspirations. The region has undergone considerable development over the last fifty years. Along with the well known infrastructure developments such as the Burj Khalifia in Dubai, there have also been significant improvements in social development, including in health and education outcomes. Much of this development has been heavily dependent on non-GCC citizens. This paper reviews a selection of demographic indicators for the GCC region which provide an assessment of progress towards the GCC population strategy's objective of improving the balance between citizens and non-citizens in the GCC.

Keywords

GCC, population strategy; Demographic Indicators; Citizens; Non-Citizens

1. Introduction

The Gulf Cooperation Council (GCC) region is comprised of six oil and gas producing countries in the Middle East. The six countries (United Arab Emirates, Kingdom of Bahrain, Kingdom of Saudi Arabia, Sultanate of Oman, State of Qatar and State of Kuwait) have all undergone significant development in the last fifty years. The GCC was formed in 1981 in order to achieve a high level of institutional coordination in economic, social, political, defence and security fields. The GCC Secretariat is responsible for activities such as cooperation, coordination, planning and programming for common action.

In 1998, the Secretariat and member countries prepared the first GCC Population Strategy. This policy aimed to provide a framework for the integration of population and development across the GCC. It was subsequently updated in 2012.

While the GCC Population Strategy has seven of policy domains that cover different aspects demographic and social aspect. The strategy sets several objectives under each domain. This paper will highlights on the first objective,

which is the population structure. This objectives aims to improve the balance between citizens (people who have citizenship of one of the six GCC countries) and non-citizens (the other residents). With this goal in mind, the policy aims to increase the population growth of the citizen part of the population, while decreasing the growth of the non-citizen part of the population.

GCC-Stat, the Statistics Centre of the GCC, provides a range of statistics to support the monitoring and evaluation of this and other GCC policies. This paper presents a number of statistical indicators that provide an assessment of progress towards this area of the Population strategy. The paper concludes with some observations about the implications of the available regional level statistics for the GCC policy process. The paper begins with a brief overview of the population situation in the GCC and a summary of the GCC Population Strategy.

- **Population Situation in GCC**

Over past few years, the demographic structure of the GCC (citizens / non-citizens) has been discussed by many researchers (e.g. Pranav Naithani and A.N. Jha, 2009). In highlighting the contribution of foreigners (non-citizens) to the growth of the population, the Naithani and Jha study emphasised the need for governments in GCC countries to play a more effective role in attracting better quality expatriate worker, who in turn could make a better contribution to the economic development of the region. This is exactly what underlies the population strategy has pointed (The success of development efforts depends, to a large extent, on the strong link between population and development. This is because incorporation of population changes in the economic and population strategies leads to acceleration of the pace of sustainable development and contributes toward achieving population goals, which in turn lead to improvement in the quality of life).

GCC-Stat estimates that the population of the GCC in 2017 was 54.9 million (GCC-Stat, 2019), compared to the UN Population Division (UNPD) estimate of 28.0 million in 1998 (UNPD, 2017). This was an increase of 26.9 million or 96.4 percent over the 1998-2017 period, and represents an annual growth rate of 3.5 over the period.

In 2017, less than half of the GCC population were citizens, with the balance non-citizens. Each of these categories of population (citizens / non-citizens) have different demographic drivers. Among citizens, natural increase (the difference between births and deaths) is the main determinant of growth. Among non-citizens, the movement of people into/out of the region is the main driver of population change.

There is also a significant sex difference in the citizen and non-citizen populations. The non-citizen population still remains predominantly male – with a sex ratio (ratio of males to females) of 159 in 2017 (GCC-Stat, 2019).

This means that males made up 61.4% of the overall population, totalling 33.7 million. There were 21.2 million females, or 38.6% of the total population.

The working age population (15-64) is the majority of the citizens, with an average of 62.3%. In 2017 with media age of 30.7 years (GCC-Stat, 2019). The non-citizen population is also mainly of working age, with an average of 87.7% with the median age of 34.7 (GCC-Stat, 2019).

- **Population Strategy in GCC**

As in many countries, GCC countries have recognized that the success of development depends in part on the strong link between population and development. In 1998, the GCC countries, supported by the GCC Secretariat developed the GCC population strategy. (GCC Secretariat, 1998)The issuance of the strategy recognized that while a large part of the development of GCC countries had relied on the active participation of a foreign workforce, the time had now arrived for the GCC states to move to a pattern of development dependent on the Gulf workforce.

The measures and challenges with measuring the balance in the population structure are the focus of the rest of the paper.

2. Methodology

Indicators for Monitoring Population Rebalancing Objectives:

The first objective of the population domain of the strategy is to balance the population structure, through reducing the growth rate of non-citizens, compared with the growth rate of citizens. In this paper, a range of indicators were identified to measure the progress of the strategy. In some cases, due to data availability issues, it was not possible to use the preferred indicators and so proxy indicators were required. For example, population projections are not available at the GCC level, so proxy measure for future population composition- the doubling time indicator was used.

The following indicators were used in this study:

I. Composition and Growth rate of citizens and non-citizens.
II. Doubling time of local and expatriates population.
III. Distribution by age and sex.
IV. Sex Ratio.
V. Labour Force Characteristics

Data Sources

For the purposes of this study, GCC level indicators were firstly sourced from the GCC-Stat portal and are shown as (GCC-Stat). Where the required indicators were not available, data was sourced from the United Nations Population Division database and are shown as (UNPD). This means that the indicators presented in the next section come from GCC-Stat directly, or a combination of GCC member countries and UNPD data. National data are not

always available for all member states by disaggregation (e.g. by citizen/non-citizens). For example, disaggregation by citizen/non-citizen for UAE and Qatar were sourced from UNPD data.

3. Results and Discussion

The key results for the five selected indicators were as follows:

Growth of Population

The total population of the GCC was 54.9 million in 2017, compared with 44.3in 2010, an increase of 24.0 percent. (GCC-Stat) The total citizen population in 2010 was 23.7 compared with 26.9in 2017, an increase of 13.7% or 3.2 million, between 2010 and 2017 (UNPD). In the same period, the non-citizen population increased from 20.8 million in 2010 to reach 28.1 million in 2017. (UNPD) This represented an increase of 7.3 million or 35.6%.

During 2010-2017, the annual growth rate for non-citizen was 4.3%, a rate far higher than the annual growth rate for citizens, which averaged 1.8% during the same period.

While there was considerable growth of the non-citizen population during the 2010-2017 period, this should be seen in the context of the longer-term growth. Over the 1990-2017 period, the number of non-citizens increased by 20.1 million, an annual increase of 743,000 non-citizens per year. Over this long period, the number of citizens only increased by 11.8 million, or on average around 437,000 per year

The demographic effect of these differences in growth rates was that by 2017, citizens made up 48.8% of the total population of GCC countries, a decline from 2010 where they were 53.4% of the total population.

Doubling indicator

The doubling indicator – the number of years that it will take for a population to double in size, based on current growth patterns, provides a picture of the possible future population structure. If the population components continued to grow at the average population growth rate, experienced over the 2010-2017 period, then it is estimated that the non-citizen population would double in about 16 years that is reaching approximately 56.0 million people in 2033.

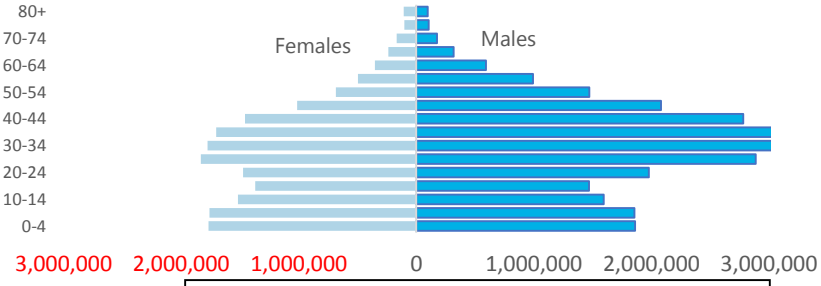
In the other hand, it is expected if the current growth rate for citizens was maintained, the GCC citizen population would double in 38 years, to reach 53.8 million in 2056. This slower increase is because the annual growth rate for citizens has been much smaller.

Age and Sex Structure

To understand some of the demographic drivers for the population imbalance, it is useful to examine the age and sex structure.

As Figure 1 shows, the total population of the GCC is dominated by the working age population, which accounts for 73.7 % of the population. People under 15 years of age represent nearly a quarter (23.3%) of the total population. The remaining 3% of the population are all aged 65 years and over.

Figure 1: Population Pyramid for GCC Total Population



Source: Derived from GCC-Stat and UNPD

While non-citizens only comprise about half of the total population, they dominate the age/sex structure. This is because the non-citizen population contains a large proportion of men of working age. In 2017, approximately 7.76 million of non-citizens were male.

Table 1 shows the population distribution by nationality (citizen/non-citizen) and gender in the GCC countries in 2017. The number of non-citizens in the GCC was 28.0 million, with 72.3% of them male. In fact non-citizen males account for 60 percent of the total population. In comparison, the number of citizens totalled 27 million.

Table 1: Population by Nationality(Citizen/Non-citizen) and Gender in GCC, 2017

Item	GCC	Kuwait	Qatar	Oman	KSA	Bahrain	*UAE
Total	54,929,523	4,226,920	2,724,606	4,559,963	32,612,641	1,501,116	9,304,277
Male	33,731,279	2,587,728	2,046,047	2,984,404	18,745,846	951,312	6,415,942
Female	21,198,244	1,639,192	678,559	1,575,559	13,866,795	549,804	2,888,335
Citizen	26,908,445	1,303,246	1,003,214	2,505,369	20,427,357	677,506	991,753
Male	13,469,057	647,694	601,359	1,263,764	10,404,282	343,340	208,618
Female	13,439,388	655,552	401,855	1,241,605	10,023,075	334,166	783,135
Non-Citizen	28,021,078	2,923,674	1,721,392	2,054,594	12,185,284	823,610	8,312,524
Male	20,262,222	1,940,034	1,444,688	1,720,640	8,341,564	607,972	6,207,324
Female	7,758,856	983,640	276,704	333,954	3,843,720	215,638	2,105,200

Source: Derived from GCC-Stat and UNPD for Qatar and UAE data by nationality

Sex Ratio

The sex ratio - the number of males for every 100 females provides another way of assessing the population structure. While the sex ratio for the GCC citizen population tends to replicate standard demographic patterns, the sex ratio for non-citizens shows the high proportion of male workers. In 2017, the sex ratio was 262.7 (UNPD), an increase from the 206.6 recorded in 1990

(UNPD). As Figure 2 shows, the sex ration increases with age. In The sex ratios among the working age of non-citizens population rapidly increase – peaking among people in the 50-54 age group, to reach a level of 252 (UNPD), before declining very rapidly.

Labour Force Characteristics

As noted earlier, most of the non-citizen population is male and of working age. It is therefore important to understand their role in the labour force, compared with citizens.

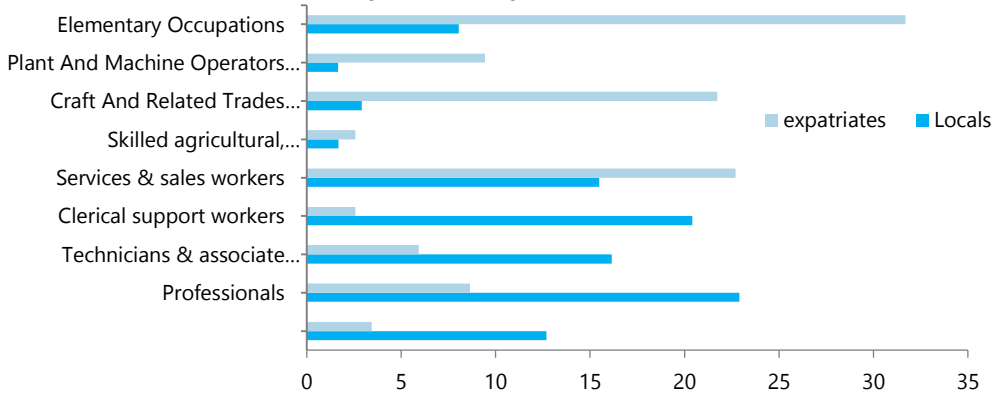
The total number of employed people in the GCC (excluding the UAE) was estimated to be 19.4 million in 2016. (GCC-Stat) Of these, 13.3 million were non-citizens and 6.1 million were citizens.

Consistent with the population growth discussed earlier, the non-citizen labour force in the GCC (excluding the UAE) increased by 9.5 million, growing on average 661,500 per year. At the same time, the citizen labour force in these five countries grew by 1.0 million, an increase of 20.5 %/ Reflecting these differences in growth rates, the non-citizen share of the labour force increased from 66.4% to 79.4% in this period, with a corresponding decline for the proportion of citizens among the labour force.

In 2017, the majority of non-citizens were concentrated in non-specialized or unskilled occupations such as Craft and Related Trades Workers, Plant and Machine Operators and Assemblers and Elementary Occupations. In comparison, citizens were heavily concentrated among the specialized technical professions, including the higher administrative roles, which require scientific qualifications. (See Figure 3)

Figure 2: Sex Ratio for non-citizens, 2017

Figure 3: Percentage Distribution of Employed (15 Years and Above) by Nationality and Occupation



Source: Derived from GCC-Stat

4. Conclusion

Five indicators were selected to assess the objective of the Population strategy to reduce the population imbalance. Analysis of these indicators at the GCC level, shows the following

- The growth rate of the non-citizen population is more than twice the growth rate for the citizen population.
- The working age population (15-64) accounts for 73.7% of the total population - most of them are non-citizens.
- The sex ratio is higher than international patterns, due to the large number of Non-Citizen Males, peaking in the 50-54 age group
- The majority of the Non-citizen labour force are concentrated in non-specialized or unskilled occupations, while citizens are concentrated in specialized technical professions, including the higher administrative roles.

Over the last decades, the reliance of the gulf cooperation council (GCC) countries economy on expatriate's workforce has increased incessantly. This increasing in the demand of expatriates, due to rapid growth of development for different aspects.

This growing in demand for this category of population has in influence in the implementing of population strategy that founded for better balance the growth rate of population between citizens and non-citizens.

The issue of rebalancing the population structure will continue for coming years, unless there is no legalization and reducing the rates expatriate workers. It is likely the proportion of expatriates to be more than the percentage of locals in the GCC for coming years.

Through some of the solutions have been implemented in the recent past to eliminate the expatriate moving to the GCC countries, yet a lot more is still required to be done by the governments and private sector organizations in the GCC countries. Gulf countries still remain an attractive destination for the expatriate workforce.

In conclusion, this paper has presented a selected number of demographic indicators that can be used to assess the Population Strategy in the GCC. While there were some data availability issues, the paper found that there are a number of indicators that decision makers to use to measure the strategy's progress.

References

1. GCC-Stat Center, 2014, Population statistics in GCC countries for 2010-2014. Retrieved https://www.gccstat.org/ar/statistic/publications/population-statistics-in-gcc-countries-2010%E2%80%932014
2. GCC-Stat Center, 2014, Population statistics in GCC stat data portal. Retrieved http://dp.gccstat.org
3. Pranav Naithani and A.N. Jha, 2009, Demographic Transitions and Imbalances in the GCC: Security Risks, Constraints and Policy Challenges. Retrieved http://ndrd.org/Demographic_Transitions_and_Imbalances_in_the_GCC.pdf
4. The Secretariat General of Cooperation Council for the Arab States,2012, The General Framework of the Population Strategy for the GCC Member States http://www.gcc-sg.org/en-us/CognitiveSources/DigitalLibrary/Lists/DigitalLibrary/Statistics/1274263573.pdf
5. GCC-Stat Center, 2016, Labour Statistics in GCC Countries for 2012-2016. Retrieved https://www.gccstat.org/ar/statistic/publications/labor-statistics
6. GCC-Stat Center, 2015, Labour Statistics in GCC Countries for 2012-2016. Retrieved https://www.gccstat.org/ar/statistic/publications/labor-statistics
7. United Nations Population Division, 2017 - Population data by age and sex from https://population.un.org/wpp/Download/Standard/Population/
8. ESCWA, 2016, Demographic Profile of the Arab Region Realizing the Demographic Dividend. Retrieved https://www.unescwa.org/file/53215/download?token=kvpcB95T

Robust estimation of multi - input transfer function model with structural change

Angelita P. Tobias, Erniel B. Barrios, Joseph Ryan G. Lansangan
University of the Philippines Diliman

Abstract

A nonparametric regression model to estimate multi-input transfer function model is proposed. Issues and limitations of the parametric transfer function such as linearity, correlated inputs, misspecification errors, short time series and presence of structural change are addressed by the nonparametric model. Three modelling approaches were compared - parametric transfer function(ARMAX), nonparametric regression generalized additive model (GAM), and forward search and nonparametric bootstrap (FSNB) method. Simulation results show that GAM performs best under short time series. Moreover, GAM is robust under the presence of misspecification error and structural change, on the number of inputs, correlated inputs, and length of time series. ARMAX on the other hand performed better on longer time series and exponentially decaying form. Forward search and nonparametric bootstrap method performed the least among the three approach but the mean absolute percent error (MAPE) is stable under different conditions of structural change such as location and length of structural change. Overall, the nonparametric approach is superior and most efficient in fitting different form of transfer function especially when there is misspecification error and correlated inputs.

Keywords

ARMAX; Generalized additive model; Forward search; Nonparametric bootstrap

1. Introduction

Suppose that an output time series $\{y_t\}$ is related to an input time series $\{x_t\}$, then we could use current and past realizations of $\{x_t\}$ to understand the structure of $\{y_t\}$. These type of models are called Box Jenkins transfer function models. Wei (2006) discussed a function describing the dynamics that exists between the output series $\{y_t\}$and the input series $\{x_t\}$ through a linear filter as follows:

$$y_t = \mathbf{v}(B)\mathbf{x}_t + \mathbf{n}_t$$

Transfer function models have been well studied and proven to be successful in many areas (e.g. Liu, L.M. & Hanssens, D. (1982); Pukkila (1982); Rahiala (1986); Nogales, F. & Conejo, A. (2006)). However, its linearity assumption limits its capability to approximate nonlinear behaviors. Moreover, it requires that the input and output series must be bivariate stationary and

that they must be cointegrated, indicating the presence of long-run equilibrium between two series (Baghestani, 1991). When there are unexpected shifts that impacts the output series such as misspecification error or structural change, the methods that we expect to work under stationary assumption are no longer appropriate. In reality, misspecification errors and structural changes happen. Moreover, structural change comes in different forms. It can be sudden or gradual, single or multiple and permanent or temporary.

Fitting a multi-input parametric transfer function requires more work specially when the inputs are correlated. Hence, there is a need to explore other options to fit a multi-input transfer function model. Transfer function with structural change is another area that was not given much attention up to date. Campano, W. & Barrios, E. (2011) proposed a robust procedure in estimating time series models with structural change by combining the power of forward search algorithm and nonparametric bootstrap (FSNB) method.

This paper proposes a nonparametric procedure in estimating transfer function with multiple inputs that will address nonlinear relationship between the input and output series even with the presence of misspecification error, correlated inputs, and presence of structural change. The proposed approach is easier to execute while delivering desirable results without going through the tedious process of parametric transfer function modelling. Also, this compared the predictive ability of three estimation methods under different scenarios using simulated data. These methods are parametric transfer function (ARMAX), nonparametric transfer function using generalized additive model (GAM), and forward search algorithm with nonparametric bootstrap (FSNB) approach.

2. Methodology

A total of 420 scenarios are considered. For each scenario, there are 100 replicates for a total of 42,000 simulated data. Each data were fed into the three estimation procedures.

Table 3.1 Simulation Settings

Setting	Levels
Form of transfer function	Finite lag, Exponentially decaying lag
Time	60, 360, 1200 time points
Points Number of Inputs	2 and 3 input variables
Levels of Correlation	All input series are uncorrelated, uncorrelated or only two inputs are correlated.
Presence of misspecification error	With/Without misspecification error
Presence of structural change	With/Without structural change
Location of structural change	Start, Middle and End
Length of structural change	5% affected, 10% affected

Input series were simulated as follows: $x_{1,t} = 20 + 0.5x_{1,t-1} + e_{1,t}$, $x_{2,t} = 0.99x_{1,t} + e_{2,t}$ and lastly $x_{3,t} = 0.5x_{1,t} + 0.5x_{2,t} + e_{3,t}$ where error terms are assumed to follow standard normal distribution. Output series were simulated as shown in Table 3.2.

Table 3.2 Output series simulation settings

Without SC	$y_{tF} = \sum_{i=1}^{3}\left(0.25 + 0.5B^1 + 0.25B^2\right)x_{i,t-1} + n_t$	$y_{tD} = \sum_{i=1}^{3}\frac{(0.25+0.5B^1+0.25B^2)}{(1-0.5B)}x_{i,t-1} + n_t$
With SC at the **start** of the series	$y_{tF} = \sum_{i=1}^{3}\left(0.1 + 0.9B^1 + 0.1B^2\right)x_{i,t-1} + n_t$	$y_{tD} = \sum_{i=1}^{3}\frac{(0.1+0.9B^1+0.1B^2)}{(1-0.9B)}x_{i,t-1} + n_t$
With SC at the **middle** of the series	$y_{tF} = \sum_{i=1}^{3}\left(0.1 + 0.8B^1 + 0.1B^2\right)x_{i,t-1} + n_t$	$y_{tD} = \sum_{i=1}^{3}\frac{(0.1+0.8B^1+0.1B^2)}{(1-0.8B)}x_{i,t-1} + n_t$
With SC at the **end** of the series	$y_{tF} = \sum_{i=1}^{3}\left(0.8 + 0.1B^1 + 0.1B^2\right)x_{i,t-1} + n_t$	$y_{tD} = \sum_{i=1}^{3}\frac{(0.8+0.1B^1+0.1B^2)}{(1-0.8B)}x_{i,t-1} + n_t$

Embedding structural change was done by simply replacing the selected segments (start, middle, and end) of the uncontaminated series with observations generated from a supposed structural change model as described below. For example, to incorporate structural change at the start of the series, generate $\{Y_1^*,\ldots Y_5^*\}$ using the appropriate structural change model, then replace $\{Y_1,\ldots,Y_5\}$, in the original series with $\{Y_1^*,\ldots,Y_5^*\}$. Thus, the new series with structural change at the start of the series is $\{Y_1^*,\ldots Y_5^*, Y_6,\ldots, Y_{95},\ldots, Y_{96},\ldots, Y_{100}\}$. The same concept applies when structural change is implanted at the middle or end of the series. After estimating models using the three methods, the models were assessed based on Mean Absolute Percent Error(MAPE).

3. Result

Table 3.1 MAPE of the models for the different scenarios by structural change (SC)

Settings	Levels	Without SC			With SC		
		ARMAX	GAM	FSNB	ARMAX	GAM	FSNB
Location of SC	Start				9.63	14.99	16.07
	Middle				8.97	10.69	15.42
	End				7.71	10.12	13.34
Length of SC	5%				8.46	10.19	14.78
	10%				9.07	13.67	15.11
Mis-specification	Without	2.21	3.12	7.37	3.68	8.79	10.22
	With	12.92	10.96	18.93	13.85	15.07	19.67
Form	Exp. Decaying	5.34	5.19	11.48	8.24	15.37	15.58
	Finite	9.79	8.89	14.82	9.29	8.50	14.30
Time points	T = 60	7.22	4.76	14.29	10.32	9.32	16.24
	T = 360	7.64	7.87	10.49	8.43	13.33	11.62
	T = 1200	7.84	8.50	14.67	7.56	13.15	16.97
Number of input variables	2 – inputs	9.68	8.70	14.49	10.44	13.01	15.58
	3 – inputs	6.16	5.93	12.25	7.65	11.21	14.52

Degree of correlation	All uncorrelated	7.70	6.83	10.92	9.17	12.07	12.34
	Partial	6.14	5.83	12.11	7.62	11.13	15.34
	All correlated	8.15	7.86	15.89	8.94	12.20	17.35

Table 3.2a MAPE of the models without structural change by misspecification error

Settings	Levels	Without misspecification error			With misspecification error		
		ARMAX	GAM	FSNB	ARMAX	GAM	FSNB
Form	Exp. Decaying	2.37	3.09	9.43	8.32	7.28	13.53
	Finite	2.06	3.15	5.31	17.52	14.63	24.33
Time points	T = 60	2.14	2.11	5.82	12.30	7.41	22.75
	T = 360	2.17	3.49	5.38	13.11	12.24	15.60
	T = 1200	2.33	3.77	10.90	13.35	13.22	18.43
Number of input variables	2 – inputs	2.49	3.37	7.61	16.87	14.03	21.38
	3 – inputs	2.03	2.96	7.21	10.29	8.91	17.29
Degree of correlation	All uncorrelated	1.98	2.49	4.34	13.43	11.16	17.50
	Partial	2.02	2.85	7.72	10.27	8.80	16.50
	All correlated	2.55	3.89	10.22	13.74	11.83	21.57

Table 3.2b MAPE of the models with structural change by misspecification error

Settings	Levels	Without SC			With SC		
		ARMAX	GAM	FSNB	ARMAX	GAM	FSNB
Location of SC	Start	4.43	11.77	11.77	14.84	18.20	20.37
	Middle	3.58	7.22	10.29	14.36	14.17	20.54
	End	3.05	7.38	8.59	12.36	12.85	18.09
Length of SC	5%	3.62	7.17	10.28	13.30	13.22	19.27
	10%	3.74	10.41	10.16	14.40	16.93	20.06
Form	Exp. Decaying	5.20	14.28	14.95	11.28	16.45	16.21
	Finite	2.17	3.30	5.49	16.42	13.70	23.12
Time points	T = 60	5.69	7.12	7.99	14.96	11.52	24.49
	T = 360	2.78	9.57	6.69	14.07	17.09	16.55
	T = 1200	2.58	9.68	15.97	12.54	16.61	17.97
Number of input variables	2 – inputs	3.93	8.92	9.70	16.96	17.10	21.46
	3 – inputs	3.52	8.70	10.57	11.78	13.73	18.47
Degree of correlation	All uncorrelated	3.48	8.41	6.04	14.86	15.72	18.63
	Partial	3.48	8.58	12.63	11.77	13.67	18.04
	All correlated	4.00	9.27	13.19	13.89	15.13	21.51

4. Discussion and Conclusion

Overall, MAPE rises in the presence of structural change. Significant increases in MAPE is more obvious with the GAM approach as compared to ARMAX and FSNB. However FSNB approach have the smallest jump in MAPE in most cases. This implies it is robust under the occurrence of perturbations in the data. In the absence of structural change, GAM performs better than ARMAX except when there is no misspecification error and the time series is

sufficiently long. Specifically, GAM is superior on short time series and becomes robust on longer time series.

To investigate the resulting detail, consider the case of no structural change in Table 3.2a. When there is no misspecification error in the output series, ARMAX provided better but sometimes comparable MAPE across different scenarios as compared with GAM and FSNB. However, in the presence of misspecification error, GAM showed superiority over ARMAX regardless of form of transfer function, length of time points, number of inputs and degree of correlation between the input variables. Specifically, GAM is significantly superior over the other two method under short time series and finite lag form of transfer function.

When there is structural change but no misspecification error, ARMAX outperforms the other two methods even with the presence of structural change. This time, GAM is comparable with FSNB. When there is both structural change and misspecification error in the output series, the MAPE of the three methods grew significantly specially the parametric approach. However in this instance, GAM has the lowest MAPE specially in fitting a finite lag transfer function and short time series while ARMAX has the lowest MAPE in fitting exponentially decaying form and in cases of sufficiently long time series. The parametric transfer function seems to perform better when there is structural change but this is due to the fact that the order and form of transfer function is known in advance. In reality, fitting a parametric transfer function with structural change is much more difficult when the form and order is yet to be determined based on cross correlation function (CCF), autocorrelation function (ACF) and partial autocorrelation function (PACF).

In summary, ARMAX performs well on sufficiently long time series in fitting exponentially decaying form of transfer function while GAM delivered significantly low MAPE for short time series and in fitting finite lag form of transfer function. On the other hand FSNB had higher MAPE compared with ARMAX and GAM but it was shown to be robust under different circumstances. In conclusion, the proposed approach which is GAM has proven to be more efficient than the traditional transfer function because of computing time consideration.

References

1. Ansley, C. and Kohn, R., (1994), "Convergence of the backfitting algorithm for additive models", Journal of Australia Mathematical Society (Series A), 57, 316-329.
2. Atkinson, A. and Riani, M., (2007), "Building regression models with forward search", Journal of Computing and Information Technology – CIT, 15, 287–294
3. Baghestani, H., (1991), "Cointegration Analysis of the Advertising-Sales Relationship", The Journal of Industrial Economics,39-6, 671-681.
4. Box,G., Jenkins, G. and Reinsell G., (2008), Time series analysis: forecasting and control,4 th edition, John Wiley & Sons, Inc.: Hoboken, New Jersey.
5. Buja, A., Hastie, T. and Tibshirani, R., (1989), "Linear Smoothers and Additive Models", The Annals of Statistics, 17-2, 453-510
6. Campano, W. and Barrios, E., (2011), "Robust Estimation of a time series model with structural change", Journal of Statistical Computation and Simulation, 81-7, 909-927.
7. Chen, R. and Tsay, R., (1993), "Nonlinear Additive ARX Models", Journal of the American
8. Statistical Association, 88-423, 955-967.
9. Chen, J., Gao, J. and Li, D., (2010), "Estimation in Semiparametric Time Series Regression", The University of Adelaide School of Economics, Research Paper No. 2010-27.
10. Daquis, J., (2010), "Nonparametric Transfer Function Models with Localized Temporal Effect", University of the Philippines School of Statistics.
11. Davison A. C. and Hinkley, D. V., (1997). "Bootstrap methods and their application", Cambridge U.K.: Cambridge University Press. Dominici, F., McDermott, A., Zeger, S. and Samet, J., (2002), "On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health", American Journal of Epidemiology, 156-3, 193-203.
12. Hastie, T. and Tibshirani, R., (1986), "Generalized Additive Models", Statistical Science, 1-3, 297-318.
13. Liu, L.M. and Hanssens, D., (1982), "Identification of Multiple-Input Transfer Function Models", Communications in Statistics, 11-3, 297-314.
14. Liu, J., Chen, R. and Yao, Q., (2010), "Nonparametric transfer function models", Journal of Econometrics, 157, 151-164.
15. Marx, B. and Eilers, P., (1998), "Direct generalized additive modelling with penalized likelihood", Computational Statistics and Data Analysis , 28, 193-209.
16. Montgomery, D., Johnson, L. and Gardiner J., (1990), Forecasting and Time Series Analysis, 2nd edition, McGraw-Hill.:Singapore.

17. Mooney, C.Z. and Duval, R.D., (1993). "Bootstrapping: A Nonparametric Approach to Statistical Inference. Sage Publications.
18. Niu, X.F., (1996), "Nonlinear Additive Models for Environmental Time Series, With Applications to Ground-Level Ozone Data Analysis", Journal of the American Statistical Association,91-435, 1310-1321
19. Nogales, F. and Conejo, A., (2006), "Electricity Price Forecasting through Transfer Function Model", The Journal of the Operational Research Society, 57-4, 350-356.
20. Opsomer, J., (2000), "Asymptotic Properties of Backfitting Estimators", Journal of Multivariate Analysis,73, 166-179.
21. Pukkila, T., (1982), "On the Identification of Transfer Function Noise Models with Several Correlated Inputs", Scandinavian Journal of Statistics,9-3, 139-146.
22. Rahiala, M., (1986), "Identification and Preliminary Estimation in Linear Transfer Function Models", Scandinavian Journal of Statistics,13-4, 239-255.
23. Riani, M. (2004). "Extensions of the Forward Search to Time Series", Linear and Nonlinear Dynamics in Time Series 8(Article 2).
24. Shafik, N. and Tutz, G., (2009), "Boosting nonlinear additive autoregressive time series", Computational Statistics and Data Analysis, 53, 2453-2464.
25. Wei, W.S., (2006), Time Series Analysis: Univariate and Multivariate Models, 2 nd edition, Pearson Education.:USA.

Competing risk analysis of lifetime data using Inverse Maxwell Distribution

M S Panwar
Department of Statistics, Banaras Hindu University,
Varanasi-221005, India

Abstract

The concept of competing risk arises in studies where failure of a system occurs due to one among several mutually exclusive causes. In this article, we consider the case when the lifetime of an individual or a component follows an inverse Maxwell distribution (an upside down bathtub hazard model). In classical approach, we give the point, asymptotic confidence interval and boot-p interval estimates of the parameters of inverse Maxwell distribution. We also applied Bayesian approach under square error loss function and give point and highest posterior density interval estimates. For illustration purpose, simulation results are established.

Keywords

Bayesian Inference; Boot-t Confidence Interval; Competing Risk Analysis; Highest Posterior Density Interval; Inverse Maxwell Distribution

1. Introduction

Maxwell distribution has broad application in many _elds such as in statistical physics, physical chemistry, and their related areas. Besides all these it has good number of applications in reliability theory also. The Maxwell distribution was first used as lifetime distribution by Tyagi and Bhattacharya(1989). Inferences based on generalized Maxwell distribution has been discussed by Chaturvedi and Rani(1998). Estimation of reliability characteristics under for Maxwell distribution under Bayes paradigm was discussed by Bekker and Roux (2005). Radha and Vekatesan(2005) discuss the prior selection procedure in case of Maxwell probability distribution. Krishna and Malik(2009) obtained the Bayes estimators of parameters and reliability functions of Maxwell distribution under progressive type-II censoring scheme. Day and Maiti(2010) proposed the Bayesian estimation of the parameter for the Maxwell distribution. Tomer and Panwar(2015) discussed the estimation procedure for the parameter of Maxwell distribution in the presence of progressive type-I hybrid censored data. Later, Modi (2015), Saghir and Khadim (2016), proposed lengths biased Maxwell distribution and discussed its various proper ties. Furthermore, several generalizations based on Maxwell distribution are advocated and statistically justi_ed. Recently, two more extensions of Maxwell distribution has been introduced by Sharma et

al.(2017a), (2017b) and discussed the classical as well as Bayesian estimation of the parameter along with the real-life application.

A random variable X follows Maxwell distribution (MWD) if its probability density function
is given by:

$$f(x,\theta) = \frac{4}{\sqrt{\pi}} \frac{x^2}{\theta^{\frac{3}{2}}} e^{\frac{-x^2}{\theta}}; \qquad x > 0, \theta > 0 \tag{1}$$

where, θ is the scale parameter.

a. Inverse Maxwell Distribution

If X has a Maxwell distribution then the random variable $Y = \frac{1}{X}$ is said to follow inverse Maxwell distribution. The pdf of inverse Maxwell distribution may be obtained by using the transformation. We have

$$f(y,\theta) = \frac{4}{\sqrt{\pi}} \frac{1}{y^4 \theta^{\frac{3}{2}}} e^{\frac{-1}{\theta y^2}}; \qquad y > 0, \theta > 0 \tag{2}$$

The survival function is given by

$$\bar{F}(t) = 1 - F(t) = \frac{2}{\sqrt{\pi}} \gamma\left(\frac{3}{2}, \frac{1}{\theta t^2}\right) \tag{3}$$

where $\gamma(a,z) = \int_0^z u^{a-1} e^{-u}\, du$ is lower incomplete gamma function. The hazard function, $h(y,\theta) = \frac{f(y,\theta)}{F(y,\theta)}$, of IMD is upside down bathtub in nature i.e. it increases sharply in initial phase then after reaching a peak point it deeps gradually and tending to zero. This means IMD represents the lifetime of such individuals/items which have a increasing chance of failing in early age of life span after survival upto a specific age, the rate of failure start decreasing as age increases.

2. Methodology

Suppose we have n such systems that each having k-components in series attachment. Here $Y_1, Y_2, \ldots, Y_n$ are the failure of all such systems where $Y_i = \min(Y_{i1}, Y_{i2}, \ldots, Y_{ik})$; i = 1, 2,..., n, k = 1, 2,...,K and Y_{ik} represents the failure time of k_{th} component of i^{th} system. Then the competing risk is represented by $\{Y_i; S_i\}$; i = 1, 2,...,n, where S_i is the cause of failure of i^{th} system. Then the likelihood is given by

$$L(\theta|\underline{y}) = \prod_{k=1}^{K} \prod_{\substack{i=1 \\ i \in S_i^{(k)}}}^{n_k} f(y_i, \theta_k) \prod_{\substack{l=1 \\ l \neq k}}^{K} \bar{F}(y_i, \theta_l) \tag{4}$$

where $S_i^{(k)}; k = 1,2, \ldots, K$ indicates the system failed due to k^{th} component and n_k denotes the number of components failed due to k^{th} component. The likelihood comes out to be

$$L(\theta|\underline{y}) = \prod_{k=1}^{K} \prod_{\substack{i=1 \\ i \in S_i^{(k)}}}^{n_k} \frac{4}{\sqrt{\pi}} \frac{1}{y_i^4 \theta_k^{\frac{3}{2}}} \exp\left(\frac{-1}{y_i^2 \theta_k}\right) \prod_{\substack{l=1 \\ l \neq k}}^{K} \frac{2}{\sqrt{\pi}} \gamma\left(\frac{3}{2}, \frac{1}{\theta_l y_i^2}\right) \tag{5}$$

Here for the ease of the problem we are considering K = 3. Then, taking logarithm of (5), we get the log likelihood as

$$l(\theta|\underline{d}) \propto -\frac{3n_1}{2}log\theta_1 - \frac{3n_2}{2}log\theta_2 - \frac{3n_3}{2}log\theta_3 - 4\sum_{i=1}^{n} logy_i - \frac{1}{\theta_1}\sum_{\substack{i=1 \\ i\in S_i^{(1)}}}^{n_1}\frac{1}{y_i^2} - \frac{1}{\theta_2}\sum_{\substack{i=1 \\ i\in S_i^{(2)}}}^{n_2}\frac{1}{y_i^2}$$
$$-\frac{1}{\theta_3}\sum_{\substack{i=1 \\ i\in S_i^{(3)}}}^{n_3}\frac{1}{y_i^2} + \sum_{\substack{i=1 \\ i\in S_i^{(1)}}}^{n_1} log\gamma\left(\frac{3}{2}, \frac{1}{\theta_2 y_i^2}\right) + \sum_{\substack{i=1 \\ i\in S_i^{(1)}}}^{n_1} log\gamma\left(\frac{3}{2}, \frac{1}{\theta_3 y_i^2}\right) + \sum_{\substack{i=1 \\ i\in S_i^{(2)}}}^{n_2} log\gamma\left(\frac{3}{2}, \frac{1}{\theta_3 y_i^2}\right)$$
$$+\sum_{\substack{i=1 \\ i\in S_i^{(2)}}}^{n_2} log\gamma\left(\frac{3}{2}, \frac{1}{\theta_1 y_i^2}\right) + \sum_{\substack{i=1 \\ i\in S_i^{(3)}}}^{n_3} log\gamma\left(\frac{3}{2}, \frac{1}{\theta_1 y_i^2}\right) + \sum_{\substack{i=1 \\ i\in S_i^{(3)}}}^{n_3} log\gamma\left(\frac{3}{2}, \frac{1}{\theta_2 y_i^2}\right) \tag{6}$$

Differentiating the log-likelihood partially with respect to θ_k; $(k = 1,2,3)$, and equating them to zero, we get the following expression which are used to obtain the MLEs of θ_k; $(k = 1,2,3)$, through numerical procedure.

$$\hat{\theta}_k = \frac{2}{3n_k}\left\{\sum_{\substack{i=1 \\ i\in S_i^{(k)}}}^{n_k}\frac{1}{{y_i}^2} - \frac{1}{\sqrt{\hat{\theta}_k}}\left\{\sum_{l=1}^{K^*}\sum_{l\neq k}^{n_l}\frac{\frac{1}{{y_l}^3}e^{\frac{-1}{\theta_k {y_l}^2}}}{\gamma\left(\frac{3}{2}, \frac{1}{\theta_k {y_l}^2}\right)}\right\}\right\} \tag{7}$$

where $K^* \in \{1,2,3\}$.

a. Asymptotic Confidence Intervals (ACIs)

The approximate (observed) asymptotic variance-covariance matrix for the MLE of parameters θ_1, θ_2 and θ_3 can be found by inverting $I(\underline{\hat{\theta}})$ as

$$I^{(-1)}(\underline{\hat{\lambda}}) = \begin{bmatrix} Var(\hat{\theta}_1) & Cov(\hat{\theta}_1, \hat{\theta}_2) & Cov(\hat{\theta}_1, \hat{\theta}_3) \\ Cov(\hat{\theta}_2, \hat{\theta}_1) & Var(\hat{\theta}_2) & Cov(\hat{\theta}_2, \hat{\theta}_3) \\ Cov(\hat{\theta}_3, \hat{\theta}_1) & Cov(\hat{\theta}_3, \hat{\theta}_2) & Var(\hat{\theta}_3) \end{bmatrix}_{\underline{(\theta)}=\underline{\hat{\theta}}}$$

Thus, using above equation we get the $100(1-\alpha)\%$ confidence limits for $\hat{\theta}_k$; $k = 1,2,3$ are given by $\hat{\theta}_k \pm Z_{\alpha/2}\sqrt{var(\hat{\theta}k)}$ where $Z(\alpha/2)$ is upper $100(\frac{\alpha}{2})^{th}$ percentile of standard normal variate. The required calculations are given below

$$\frac{\partial^2 l(\underline{\theta}|\underline{y})}{\partial {\theta_k}^2}|_{\theta_k=\hat{\theta_k}} = \frac{3n_k}{2{\hat{\theta}_k}^2} - \frac{2}{{\hat{\theta}_k}^3}\sum_{i=1_{i\in S_i^{(k)}}}^{n_k}\frac{1}{{y_i}^2} - \sum_{l=1}^{K*}\sum_{l\neq k}^{n_l}\psi(y_l, \hat{\theta}_k); K^* \in \{1,2,3\} \tag{8}$$

where,

$$\psi(y,\theta) = \xi(y,\theta)\left[\frac{1}{\theta^2 y^2} - \frac{5}{2\theta} + \xi(y,\theta)\right] \tag{9}$$

and

$$\xi(y,\theta) = \frac{1}{y^2}e^{\frac{-1}{\theta y^2}}\left[\theta^{\frac{5}{2}}\gamma\left(\frac{3}{2}, \frac{1}{\theta y^2}\right)\right]^{-1} \tag{10}$$

All non-diagonal elements, $\frac{\partial^2 l(\theta|d)}{\partial \theta_j \theta_k}$; j ≠ k,(j,k = 1,2,3) are zero.

b. Boot-p Intervals

In the case the number of sample observations in experiment is not very large ACIs mentioned above are not suitable. So we provide another procedure to obtain bootstrap CIs for θ advocated by Efron and Tibshirani (1986). The steps for applying parametric bootstrap method are given below:

1. Based on the original sample y = (y_1,y_2,...,y_n), obtain the MLE of $\hat{\theta}$.
2. Under the same conditions generate sample, say (x_1,x_2,...,x_m), from the underlying distribution IMD($\hat{\theta}$).
3. Compute the MLE of $\hat{\theta}$ based on (x_1,x_2,...,x_m), say $\hat{\theta}^*$.
4. Repeat step (2) and (3) B times and obtain $\hat{\theta}^*_1$, $\hat{\theta}^*_2$, ..., $\hat{\theta}^*_B$.
5. Arrange $\hat{\theta}^*_1$, $\hat{\theta}^*_2$, ..., $\hat{\theta}^*_B$. in ascending order.
6. A two-sided 100(1−α)% percentile bootstrap confidence interval of Θ, say $\left[\hat{\theta}^*_L, \hat{\theta}^*_U\right]$ is given by

$$\left[\hat{\Theta}^*_L, \hat{\Theta}^*_U\right] = \left[\hat{\Theta}^{*\left(\frac{\alpha}{2}\right)}_B, \hat{\Theta}^{*(1-\frac{\alpha}{2})}_B\right]$$

3. Bayesian Estimation

Under the Bayesian paradigm θ is considered a random variable. Let us consider inverted gamma distribution as prior distribution $IG(\mu_k,\nu_k)$ of θ_k where k = 1,2,3 given by the pdf

$$\pi_k(\theta_k \propto \frac{1}{\theta_k^{\nu_k+1}} \exp\left\{-\frac{\mu_k}{\theta_k}\right\}; \quad \mu_k, \nu_k > 0, k = 1,2,3. \tag{11}$$

Merging the joint prior density with the likelihood function, the required joint posterior, up to proportionality, comes out to be

$$\pi(\underline{\Theta}|\underline{y}) \propto \frac{1}{\theta_1^{\frac{3n_1}{2}+\nu_1+1}} \frac{1}{\theta_2^{\frac{3n_2}{2}+\nu_2+1}} \frac{1}{\theta_3^{\frac{3n_3}{2}+\nu_3+1}} \prod_{\substack{i=1 \\ i\in S_i^{(1)}}}^{n_1} \frac{1}{y_i^4} e^{-\frac{1}{\theta_1}\left(\mu_1+\frac{1}{y_i^2}\right)} \gamma\left(\frac{3}{2}, \frac{1}{\theta_2 y_i^2}\right) \gamma\left(\frac{3}{2}, \frac{1}{\theta_3 y_i^2}\right)$$

$$\prod_{\substack{i=1 \\ i\in S_i^{(2)}}}^{n_2} \frac{1}{y_i^4} e^{-\frac{1}{\theta_2}\left(\mu_2+\frac{1}{y_i^2}\right)} \gamma\left(\frac{3}{2}, \frac{1}{\theta_3 y_i^2}\right) \gamma\left(\frac{3}{2}, \frac{1}{\theta_1 y_i^2}\right)$$

$$\pi_k(\theta_k|\underline{d}) \propto \frac{1}{\theta_k^{\frac{3n_k}{2}+\nu_k+1}} exp\left\{-\frac{1}{\theta_k}\left(\mu_k + \sum_{\substack{i=1 \\ i\in S_i^{(k)}}}^{n_k} \frac{1}{y_i^2}\right)\right\} \prod_{\substack{l=1 \\ l\neq k}}^{n_k} \gamma\left(\frac{3}{2}, \frac{1}{\theta_k y_l^2}\right) \tag{13}$$

From the above expression we observe that the marginal distributions of θ_k;(k = 1,2,3), cannot be obtained in the closed form, which is essential in order to obtain the Bayes estimates of individual parameters or obtain the parametric functions. Therefore, for further analysis, we proceed to Gibbs Sampler. For this we have to obtain the full condition distributions for θ_k;(k = 1,2,3), which are given below as:

a. MCMC Method

We use Metropolis-Hastings algorithm method with proposal density normal distribution to generate sample observations from $\pi_k(\theta_k|d)$; (k = 1,2,3) given by equation (14). The Metropolis-Hastings Algorithm has the following steps:

1. Set t=1 and take $\theta_1^0 = \hat{\theta}_1$, $\theta_2^0 = \hat{\theta}_2$ and $\theta_3^0 = \hat{\theta}_3$.
2. Generate a candidate point θ_k^* from proposal density $q_k \sim N(\hat{\theta}_k, v(\hat{\theta}_k))$ and take a point u from a uniform distribution $U(0,1)$. Then compute an acceptance ratio

$$r_k = \frac{\pi_1(\theta_k^*|\underline{d})q_k(\theta_k^{(t-1)})}{\pi_1(\theta_k^{(t-1)}|\underline{d})q_k(\theta_k^*)}$$

3. Let $p(\theta_k^{(t-1)}, \theta_k^*) = min(r_k, 1)$, then set $\theta_k^{(t)} = \theta_k^*$ if $u \leq p(\theta_k^{(t-1)}, \theta_k^*)$ and otherwise $\theta_k^{(t)} = \theta_k^{(t-1)}$
4. set t=t+1.
5. Repeat steps (2)-(4) N times to get the chain $\theta_1^1, \theta_1^2, \ldots, \theta_1^N$, $\theta_2^1, \theta_2^2, \ldots, \theta_2^N$ and $\theta_3^1, \theta_3^2, \ldots, \theta_3^N$, where N is very large number.

After the convergence of chain, we obtain $N*$ out of N samples, say $\theta_1^1, \theta_1^2, \ldots, \theta_1^{N*}$, $\theta_2^1, \theta_2^2, \ldots, \theta_2^{N*}$ and $\theta_3^1, \theta_3^2, \ldots, \theta_3^{N*}$ and with the help of these sample observations, we obtain the Bayes estimates for the parameters. we also obtain Bayesian credible and HPD intervals for θ by using algorithm given by Chen and Shao.

4. Result

In this section, we perform the simulation study to verify the theoretical results numerically and check the performance of estimators. Since we have considered the three-component series systems, so actual values taken in computation are $\theta_1 = 1.2$, $\theta_2 = 1.25$ and $\theta_3 = 1.3$. The simulation study is carried out for sample size n = 20, 30 and 50. In this section, we perform the simulation study to verify the theoretical results numerically and check the performance of estimators. Based on this, the MLE, ACI and boot-p CIs are calculated. We repeat each generation and estimation procedures 1000 times and give average values of point estimates and corresponding mean square errors (MSEs) in Table 1. We also provide average length and coverage probabilities (CP) given in Table 2. Bayes estimates and HPD intervals are obtained using Gibbs Sampler in which samples are drawn from full conditional given in the section (2) by using Metropolis-Hasting algorithm. We run the three different chains by generating 50000 observations. For diagnosis of the convergence of chains, we draw cumuplot at 5%, 50% and 95% quantiles. Outcomes of cumuplot function applied to MCMC samples are given in Figures 1 for θ_1, θ_2 and θ_3, respectively. For the arbitrarily chosen values of prior parameters $\mu_1 = 5, \mu_1 = 5, \mu_1 = 5, n_1 = 2, n_2 = 2$ and $n_3 = 2$, we present the average values of Bayes

estimates along with their MSEs for 1000 repeated samples. The Bayes estimates and their MSEs are given in Table 1 and HPD with coverage probabilities are presented in Table 2.

Table 1: Average values of point estimate of θ1, θ2 and θ3 and their MSEs(in Bracket) for n = 50, 30 and 20.

	MLE's			Bayes Estimates		
n	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\tilde{\theta}_1$	$\tilde{\theta}_2$	$\tilde{\theta}_3$
50	1.221	1.275	1.3233	1.2660	1.3019	1.3381
	(0.048)	(0.0568)	(0.0523)	(0.0248)	(0.0245)	(0.0211)
30	1.222	1.2640	1.3076	1.2778	1.3078	1.3323
	(0.074)	(0.0718)	(0.0786)	(0.0289)	(0.0251)	(0.0230)
20	1.211	1.2323	1.2852	1.2813	1.3032	1.3295
	(0.099)	(0.0912)	(0.0992)	(0.0300)	(0.0243)	(0.0220)

Table 2: Average values of point estimate of θ1, θ2 and θ3 and their MSEs(in Bracket) for n = 50, 30 and 20.

	ACI			Boot-t			HPD		
n	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\tilde{\theta}_1$	$\tilde{\theta}_2$	$\tilde{\theta}_3$
50	0.8224	0.8105	0.9085	0.7524	0.7405	0.8185	0.7107	0.7275	0.7448
	93.75	87.71	92.58	91.75	89.71	93.58	98.50	97.67	96.25
30	1.1081	1.0840	1.2382	0.9008	0.9849	0.9381	0.8193	0.8308	0.8410
	93.13	89.12	93.91	95.13	87.12	96.90	98.83	98.58	98.67
20	1.4640	1.4189	1.6654	1.3516	1.3156	1.4521	0.8942	0.9007	0.9110
	93.86	86.99	95.98	97.12	91.12	95.45	99.33	99.67	99.42

5. Discussion and Conclusion

From the simulation study, it can be easily seen that as sample size increases the MSEs of the respective estimates decrease rapidly. As sample size increases average length of intervals also decreases with increasing coverage probability which is obvious and certify all theoretical derivation. ACI's are always symmetric, but boot-p and HPD CIs shows the actual nature of the parameter estimator. It can also be observed that the HPDs have shorter lengths then that of ACIs and Boot-t such that it provide us results in more precise manner. And we can recommend Bayesian inferential in place of classical estimation procedure.

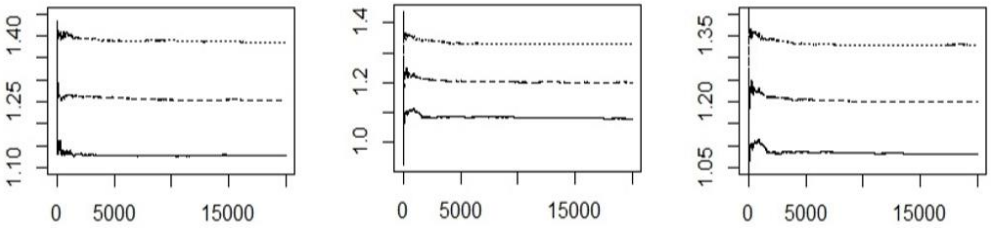


Figure 1: Cumsum plot for quantiles at 5%, 50% and 95% for θ1,θ2 and θ3.

Since the distribution with upside down bathtub hazard rate model are rarely used for competing risk analysis. So all such experiments where individuals/items in competing cause and follows upside down bathtub hazard

rate, this study can be utilise easily. We can guaranty more accurate results rather than using any other model can achieve which is not appropriate for the system.

References

1. Bekker, A. and Roux, J. (2005). Reliability characteristics of the Maxwell distribution: A Bayes estimation study. Communications in Statistics-Theory and Methods, 34(11), 2169-2178.
2. Chaturvedi, A.,& Rani, U.(1998). Classical and Bayesian reliability estimation of the generalized Maxwell failure distribution. Journal of Statistical Research, 32(1), 113-120.
3. Crowder MJ. Classical Competing Risks. Boca Raton, Florida: Hall, 2001.
4. Effron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Vol. 57. Monographs on Statistics and Appplied Probability: Chapman& Hall.
5. Mann, N. R., Singpurwalla, N. D., & Schafer, R. E. (1974). Methods for statistical analysis of reliability and life data.
6. Modi, K., & Gill, V. (2015). Length-biased weighted maxwell distribution. Pakistan Journal of Statistics and operation research, 11(4), 465-472.
7. Prentice, R. L., Kalbeisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., & Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. Biometrics, 541-554.
8. Sharma, V. K., Bakouch, H. S., & Suthar, K. (2017). An extended Maxwell distribution: Properties and applications. Communications in Statistics-Simulation and Computation, 46(9), 6982-7007.
9. Sinha, S. K. (1986). Bayes estimation of the reliability function and hazard rate of a weibull failure time distribution. Trabajos de Estadistica, 1(2), 47-56.
10. Tomer, S. K., and Panwar, M. S. 2015. Estimation procedures for Maxwell distribution under type I progressive hybrid censoring scheme. Journal of Statistical Computation and Simulation 85:33956.

Outlines of SCAD's pilot test for the 2020 register based census

Aisha Turki, Dr Salah Qudairi, Badreyya Al Shehhi, Thumna Alrashdi
Statistics Centre Abu Dhabi, UAE

Abstract

Countries, including developed countries, tend to use administrative registers to identify their population count, and to identify its characteristics, so that many countries use administrative registers data as an alternative to a field enumeration census, and this saves time, effort and money. As well as using administrative registers data contributes to periodically identifying the population count without waiting for a long period of time to conduct the field enumeration. This is useful for communities that undergo rapid population changes, as is the case in the UAE in general, and the Emirate of Abu Dhabi in particular. Based on the Abu Dhabi Government's interest in development, and its role in the community's wellbeing, using administrative registers is important in deriving statistical indicators that monitor and analyse population and social changes, and providing them to decision-makers, and this will contribute to developing programs and policies that will benefit the community. This requires building a database and maintaining its sustainability and populating it with data on a continuous basis. For this purpose, coordination has been established with the entities owning the administrative registers, and with support from senior management in the Statistics Centre – Abu Dhabi (the Centre) and senior management in those data providing agencies, the efforts have resulted in the signing of service level agreements, and data have begun to flow to the Centre by virtue of these agreements, and the Centre has established specialized registers such as, population register, employees register, unemployed register, education register, and people with disabilities register.

Establishing administrative registers at the Centre aims to:

- Conducting register-based census
- Creating a framework for designing samples for demographic and social surveys, etc.
- Providing statistical information and data that support planning and decision makers
- Producing statistical indicators
- Predicting future needs of social care services for the population.
- Preparing detailed analytical reports to assist decision makers in developing policies.

Keywords
Enumeration; Population; Housing; Administrative.

1. Introduction

1. Overview of Register-Based Census Project

1.1 Description

Register-based census project is a census project based on utilizing administrative registers data that are collected by government agencies for their own purposes, as an alternative to a traditional census based on collecting data from field.

a. Objective

Creating a detailed database on Abu Dhabi population's characteristics to support policy makers and decision makers in developing and monitoring social and economic policies and infrastructure programs in the Emirate.

b. Project Framework

The Centre's scope of responsibility is the Emirate of Abu Dhabi, and the project is linked to the Centre's strategic objective related to increasing reliance on administrative registers data and reducing field surveys.

2. Project Phases

2.1 Preparatory Phase

It was conducted from the second quarter of 2015 until the third quarter of 2017.

2.2 Pilot Test 2017

A pilot test on administrative registers data has been conducted, and this work paper will include a presentation of the challenges, and how to overcome them.

2.3 Pilot Census 2019

It will be carried out in the second quarter of 2019 to measure the extent to which the challenges of the 2017 pilot test experience have been overcome, and the effectiveness of the solutions provided.

2.4 Actual Register-based Census 2020

All lessons learned from pilot test and pilot census will be applied in this phase.

2. Methodology

A methodology has been developed to identify the concept of Abu Dhabi residents, and they have been defined as individuals who are habitual residents of the Emirate and are individuals who usually reside in the Emirate of Abu Dhabi, whether UAE nationals or non-UAE nationals. The definition covers citizens, who are habitual residents of the Emirate, as well as those who are from outside of the Emirate or the United Arab Emirates and habitually live in the Emirate of Abu Dhabi. It also covers all non- UAE nationals who intend to permanently and continuously reside in the Emirate, or for at least six months, regardless of visa status, or spent six consecutive months in the Emirate prior the census benchmark. It should be noted that periods of temporary absence due to annual leave or work
assignments are not considered to be an interruption within the consecutive six-month period.

Following are the most important points upon which Abu Dhabi residents are identified:

- The concept of Abu Dhabi residents includes all citizens holding family book issued from the Emirate of Abu Dhabi, regardless of their habitual residence.
- The concept of Abu Dhabi residents includes all GCC nationals who habitually reside in the Emirate of Abu Dhabi.
- The concept of Abu Dhabi residents includes all foreign residents who habitually reside in the Emirate of Abu Dhabi.
- The concept of Abu Dhabi residents includes all newly born children for Abu Dhabi UAE nationals as well as newly born children for non-UAE nationals who habitually reside in the Emirate of Abu Dhabi.
- The concept of Abu Dhabi residents excludes UAE nationals or non-UAE nationals who passed away prior to the benchmark of administrative register data.
- The concept of Abu Dhabi residents excludes foreign residents, whose residence permits have been cancelled, and who left the Emirate of Abu Dhabi prior to the benchmark of administrative register data.
- The individual shall be counted once in the population count regardless of the number of times of renewing residency or identity card.

3. Result

Technical achievements

Following are some of the most important technical achievements:

- ✓ At the level of variables: The following have been achieved:
  - Identifying variables and evaluating their coverage from administrative registers
  - Preparing itemized card for each variable

- Identifying and designing output tables
- At the level of concepts: The following have been achieved:
- Adopting methodology of Abu Dhabi population count
- Defining concepts associated with methodology
- Sharing methodology with key data sources
- Defining register-based census variables

✓ At the level of classifications: The following have been achieved:
- Developing classifications of each variable according to the classifications applied by the Centre
- Sharing classifications with data sources

✓ At the level of Data: The following have been achieved:
- Receiving, processing and analysing data from some key sources
- Designing the population register database populating the register-based census
- Preparing a methodology of collecting data from other administrative registers
- Developing the rules to match the variables

Technical Procedures

Statistics Centre - Abu Dhabi (SCAD) conducted the statistical processing necessary on the data. The data was reviewed and validated at the Centre according to several stages and processes to prepare and process data to derive the statistical indicators. The procedures included the following:

3.1 Corresponding Variables' Names in Different Data Sources (variables mapping)

The variables' names received from each source have been reviewed, corresponded and matched with the variables' names in the approved database, so that the same data feed into one variable regardless of the column name in the entity owning data. For example, identity card number variable has taken several different names in the sources, but these names have been corresponded with the variable name approved in the Centre's database, and the names given to this variable include, but are not limited to, SPM_NATIONAL_ID, Emirates ID, and PUPIL_EMIRATES_ID.

The name of identity card number variable has been unified in all registers and it has been named as NATIONAL_ID, and it is the identification number through which the various registers' data have been linked.

3.2 Coding

The Centre gave specific codes for data and replaced text answers (sentences, words or codes) with specific codes and with specific connotations,

giving each case a brief code for that case. For example, the nationality variable has contained several words denoting UAE nationals such as UAE, 101, and United Arab Emirates. All words and codes denoting UAE nationals have been unified into the United Arab Emirates and they have been given the code 784 according to Country Code Classification issued by Statistics Division in United Nations 2014.

3.3 Classifications

The Centre classified the raw data received from the entities in accordance with the international classifications.

The most important classifications used are as follows:

- Classification of Education: International Standard Classification of Education has
- been used (ISCED 2013)
- Classification of Occupations: International Standard Classification of Occupations
- 2008 has been used (ISCO-08)
- Classification of Nationalities: Classification of Nationalities (countries) 2014

3.4 Data Validation

Data have been checked to detect some problems, errors, or inconsistencies between the data.

- Example: Age Variable

Problem (1): There are no available birth dates which hinder the process of age calculation

Problem (2): Ages presented in large numbers exceeding 300 years (For example, birth year 1365)

Problem (3): Ages presented in negative numbers (For example, birth year 2070)

Accordingly, the cause of this problem is due to data entry error and in this case, the birth date has been entered from registers data from other available sources. If the required data is not available, the birth year birth is entered from identity card number, where the digits from 4 – 7 in identity card number denote the birth year.

3.5 Updating Data

The Centre has updated the database through its administrative registers data, in accordance with clear procedures that identify the data to be updated, and identify the priority main source for each variable, and then the sub-source. For example, students' data have been used in the Department of Education and Knowledge's administrative registers to reflect grades that

students have attained with an educational level variable, according to International Standard Classification of Education (ISCED 2013). Then the data from the remaining available registers have been used. For example, the educational levels of employees working for Abu Dhabi government have been updated using the Department of Finance register. For individuals, whose educational levels are not available in other registers, their educational levels have been adopted as indicated in Ministry of Interior's data.

3.6 Removing Duplicates

The data have been verified according to specific procedures to ensure duplicated cases are detected and verifying that the same individual has been only registered once in the database, that has included checking duplicates in the same register, and checking duplicates between registers. This is an important issue in dealing with administrative registers data. For example, it is normal for a person to be found duplicated in Department of Health's registers, because he visits the hospitals more than once. Each time the patient may be suffering from a different disease, other than the disease diagnosed during the first visit, and/or the patient may be following up with a different specialist doctor (Internist, Ophthalmologist, …..). Therefore, no matter how many visitations made by the patient to hospitals, we shall ensure that this person is not included in the population count more than once.

3.7 Data Validation

The Centre provides an integrated system of procedures to ensure quality. This system ensures continuity of providing the required data to the Centre, continuity of updating the register data, consistency of data with statistical definitions and classifications, and a specific mechanism for periodically measuring data quality and revealing shortcomings or lack of coverage in the data. In addition to placing the same codes for similar characteristics that share a particular feature, such as place of residence variable, and place of work variable.

4. Discussion and Conclusion

Pilot test for register based census has revealed several strengths that we are proud of in the United Arab Emirates, where administrative registers have been characterized by a large number of variables, as well as covering the UAE nationals through family books, and covering non- UAE nationals through residence permits issued to them. The data also has showed flexibility and ability to be classified according to international standard classifications, in addition to the possibility of deriving variables that are not available directly.

The population's addresses are considered one of the most important challenges faced by a register-based census 2020, due to lack of the accurate

addresses with a particular entity for the entire population. Although this challenge has been addressed by the Centre according to the register data available by the Centre, such as documentation of leases, Owners' details, electricity and water details and school students' addresses, but the addresses issue remains a major challenge to the register-based census, which we hope to overcome, through enacting a legislation or administrative decision binding to register the address in detail according to the addressing project (Onwani) in Abu Dhabi, as well as binding to updating the address after a move The person, who fails to comply with such legislation or administrative decision, shall bear legal consequences.

References

1. Federal Authority for Identity and Citizenship (ICA) UAE (2017) Abu Dhabi emirates population https://www.ica.gov.ae/en/open-data.aspx
2. Department of Urban Planning and Municipalities (DPM) Abu Dhabi UAE (2017) Address https://www.dpm.gov.abudhabi/en/News-and-Media/Statistics
3. Information & eGovernment Authority)IGA) Bahrain (2016) showcase Bahrain's advanced experience in the field of administrative records and its statistical uses in the Authority's General Directorate of Statistics & Population Registry. http://www.iga.gov.bh/en/article/abu-dhabi-statisticscenter-visits-iga-bahrain
4. Statistical Centre for the Cooperation Council for the Arab Countries of the Gulf (GCC-Stat) Sultanate of Oman (2016) Recommended 2020 Population and Housing Census Data Basket https://gccstat.org/en/statistic/projects/the-development-of-the-business-registers-in-the-gcc-countriesby-the-end-of-2017
5. ILO. (2008). The International Standard Classification of Occupations (ISCO-08)
6. Statistics Centre - Abu Dhabi. (2014). Nationalities (Countries) Classifications.
7. UNESCO. (2013). International Standard Classification of Education.

Advances in maintenance of critical plant machinery equipment, frequency optimization and minimization of breakdowns perspective

Rahul Chattopadhyay
Data Science & Analytics Manager Kolkata, India

Abstract

A steelwork comprises of various plant facilities (SMS, HSM, CRM) operating with different equipment at certain time intervals. Proper functioning of such equipment is vital for flow of production. So, keeping a track on their breakdown as well as timely maintenance becomes necessary factor to run manufacturing facility business. Predictive maintenance is one of the finest techniques to determine the condition of equipment employed, failure occurrence and predicting the timeline of maintenance. Once the process of maintenance is done on equipment, they pose a higher risk of breakdown. However, equipment after maintenance shows breakdown pattern similar to any new equipment. With ample data points, we can observe ideal bath tub curve measuring breakdown probability. This paper discusses the reduction in frequency of maintenance in respect to reduction in breakdown probability. Equipment maintenance frequency is optimized using survival analysis technique to determine the probability of maintenance frequency with respect to time in days.

Keywords

Predictive, Optimization, Survival, Equipment, Breakdown

1. Introduction

Today the business environment is very much competitive. Firms cannot afford to wait for equipments to fail before making any repairing work decision. It would not be ideal to depend on time based preventive maintenance to bring down the machine downtime. So it is best to devise a predictive approach that is ideal for running the operations.

Maintenance costs, consumes a major operating cost for all manufacturing plants. The cost of maintenance can go up to 60% of cost of goods produced depending upon the industry. For example, in agro related industries the average maintenance cost can be 20%, whereas for steel and paper industries it can go beyond 50% of total production cost. Such percentages vary and can be over forecasted. There are many costs for plants which are related to operations and market sentiments such as buying new product. These expenses are actually non-maintenance costs. But a true maintenance cost puts burden on the plant profitability. Past surveys of maintenance

management shows that one third of all maintenance costs are wasted as a result of inappropriate or unnecessary maintenance. Now we will go through the detailed study to see how survival analysis can become fruitful towards minimizing equipment breakdown frequency and unnecessary maintenance.

Objective and Assumptions

We can see that there is always some scope of reduction in equipment breakdown with optimal frequency of maintenance.

In this study analysis is done to see the overall breakdown events for the equipments and identify planned maintenance (Z2, Z4 & Zm06) events for the equipments.

To carry out this research study the following assumptions has been considered:

1. It is assumed daily process of tool will estimate the breakdown for next three days only.
2. It is assumed that notification data is based to Z2 & Z4 and order data based on Zm06 types.
3. Maintenance will be carried out for the current year only and will be shown in graph as average number of days.
4. First parametric estimation will be done followed by non parametric estimation by the survival tool.
5. Kaplan Mayer estimator is a high end estimator accounted for survival time of equipments and to work on non parametric data for this study.

Closely Related Research Work

In the past we have witnessed many closely related research works that has been carried out related to the study based on survival analysis and machinery maintenance methodology. One of the research work carried out by ZHIGUO LI1, SHIYU ZHOU, SURESH CHOUBEY and CRISPIAN SIEVENPIPER Jan(2006) based on Cox proportional hazard model to diagnose machine failure events. They brought in new methods which are related to design and data based. Design based methods are commonly found in automated manufacturing systems. In this method after obtaining the desired event sequence it is compared with the observed event sequence. Srinivas and Jafari(1993) proposed a control-monitoring maintenance architecture (CMM) for (flexible manufacturing system)FMS based on Petri nets with objects (PNO), where stochastic rates are associated to the modeling of maintenance planning. Since failures during operating time can degrade FMS performance or its availability so minimizing the unplanned costly breakdowns has become a necessity. They presented the study organized in different CMM modules. In each module the interaction of a control module, process's reference models (allowing the detection of failures) and a maintenance module (consisting of

a maintenance model and a statistics module) was observed. Their Modeling approach was based on Petri nets only allowing a modular optimization of the maintenance procedure. They didn't address the case in which several event sequences collected from different pieces of equipment. So Predicting failure event by aggregating all equipment the data together is to be explored. Silvia Madeira, Paulo Infante, Filipe Didelet(March 2013) applied a Cox model to a particular critical equipment in order to find process variables that cause its vibration and apply well known distributions towards baseline hazard rate. This approach was more related to chemical composition of variables and their influence on increasing risk of high vibration values, though these variables are difficult to control as they depend on reactor temperatures. But their research was not based on mechanical equipment in order to optimize detailing predictive maintenance scenarios.

So this paper will address all the new possibilities that can be used on real time and helps the manufacturing unit by cutting down the breakdown costs substantially. So using survival analysis technique a proper model has been implemented to visualize the outcome of time to failure events and minimize breakdowns.

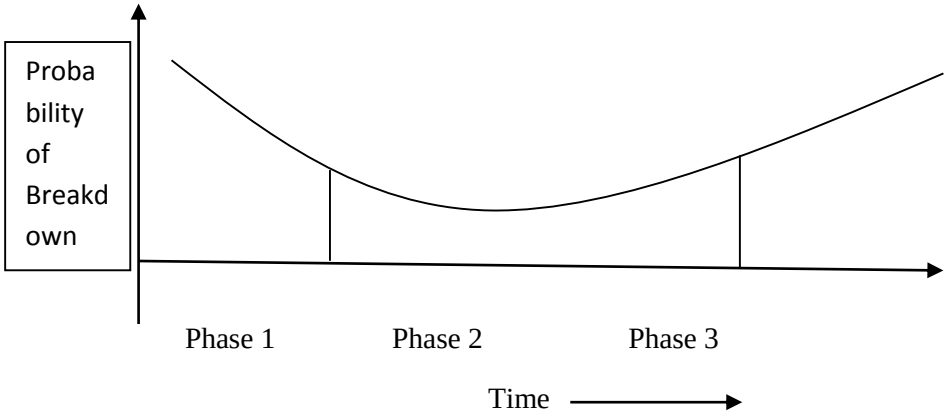
Description of the Data

In this section we will see the structure of data related to this research. The real-world data analyzed is based on process parameters as well as product parameters.

A Large amount of real time data is generally recorded over a long period of monitoring. Two types of data files are primarily used- notification data file and order data file. These data files are generated from three stations - steel melting shop(SMS), hot strip mill(HSM), cold rolling mill(CRM) ordered from left to right. Apart from these data files another aggregated data file is used in the process which includes aggregated equipment data for all the three stations. All the three data files are separately generated for three stations. The major insights are generated that can be used by the management of maintenance department and managers of mechanical department. Total 48 months of data from July ’2014 to March'2018 has been recorded. Order data file has 8 variables, notification has 10 and aggregated equipment data file has 23 variables. There are few overlapping variables in the data files e.g. equipment id, department, function. From the tool pane blank templates for notification and order data can be downloaded in a pre determined folder. It is that the final curve will be able to predict the timeline of breakdown for the specified equipment.

2. Methodology

The methodology being followed here to predict the time to failure of equipments is based on survival analysis. Though survival analysis has applications in clinical trials for treatments but now widely used in many industries including manufacturing. The need of survival analysis is to take preventive measures that are time driven. The life cycle of a machine is observed by bathtub curve which is also known as mean time to failure. Machinery equipment is more probable towards breakdown because of any installation issue during 1 months of operation (Phase 1). But after this period the probability of breakdown goes down for a certain period of time also known as normal life span of equipments (Phase 2). Again the probability of breakdown occurrence increases sharply once the normal period gets over (Phase 3). So equipment maintenance is normally based on MTTF figures as shown below.



Here breakdown frequency is optimized by keeping two underlying principles in mind-

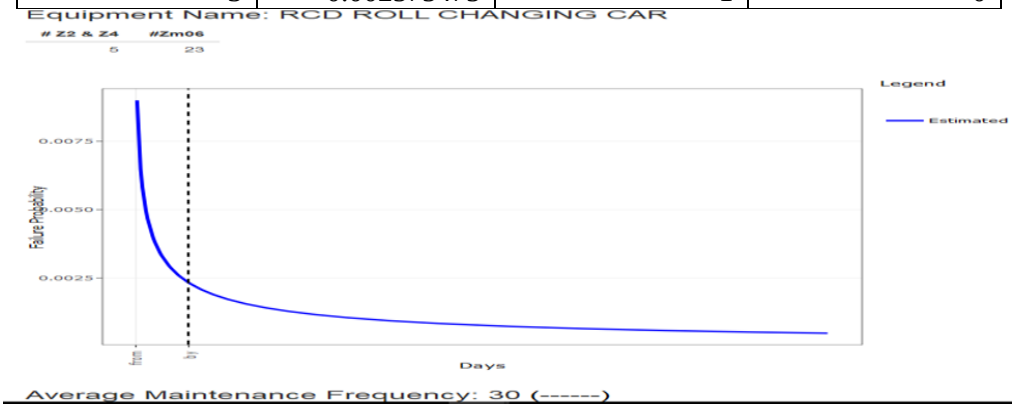
1. Equipment poses higher risk of breakdown soon after maintenance.
2. Equipment after maintenance display breakdown pattern similar to that of a new equipment. With enough data points, ideal bath tub curve is observed.

Next we will discuss results of the study.

3. Result and Discussion

Now we will discuss the results of the study. After running the survival analysis we get time for breakdown in days (time), predicted average probability (Est_probability), lower control limit (LCL) and upper control limit (UCL). Below is a sample of data generated for roll changing car equipment after the analysis.

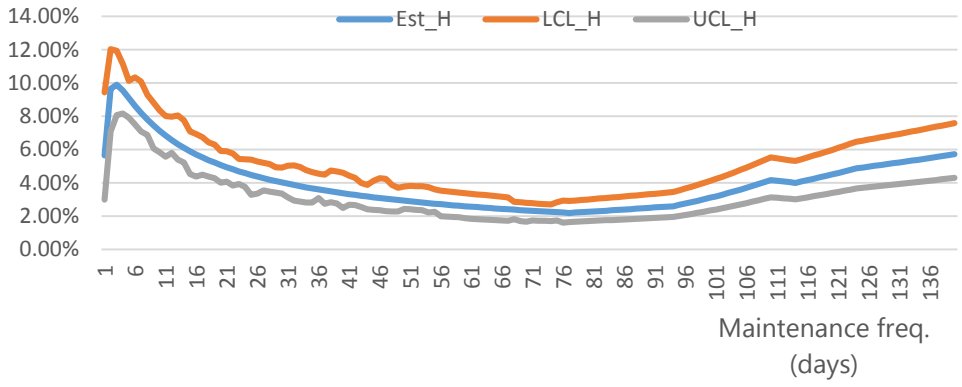
time	est	lcl	ucl
1	0	0.994921607	0
2	0	0.283044991	0
3	0.002575475	1	0



Est_H probability: Predicted average probability of breakdown for the given gap in maintenance activities on X axis. Higher probability means higher chances of breakdown at that probability.

LCL_H and UCL_H: For the given time period in X axis, 95% of predicted values lie between respective LCL and UCL values. Higher LCL value means, breakdown probability could be experienced to be higher.

As the study proceeded further, there are few important findings.



To summarize the above graphical output-

- Current maintenance frequency: 50 days
- Maintenance frequency till which probability breakdown trend is decreasing: 80 days
- Maintenance frequency between 50 and 80 days till which LCL is a decreasing trend: 74 days 74 days less than 100 % increment over original frequency
- New test frequency: 74 days

4. Conclusion

As per the research, study and impact on the machinery equipments breakdown was done taking four years data and meaningful insights were generated towards frequency optimization of critical equipments. Concluding survival method is one of the fittest methods to minimize the breakdown frequency of equipments.

Below are few important points that remain major findings for the study.

- Map the current maintenance frequency on the graph, and note the corresponding probability of breakdown.
- If the breakdown probability trend is majorly increasing both towards the right and the left side of the current maintenance frequency, it means the current maintenance frequency is already optimized.
- If the breakdown probability is majorly staying constant with change in maintenance frequency, it shows minimal impact of maintenance activities on equipment breakdown.
- If the breakdown probability is decreasing in right (left) side of the current maintenance frequency, go in the direction of decreasing probability and shortlist frequency basis following rules.

References

1. Qiao Dong, Baoshan Huang: (Dec 2015). Survival Analysis of the Failure Probability of Resurfaced Preventive Maintenance Treatments in the Long-Term Pavement Performance Program. Transportation Research Record Journal.
2. Chen, C., R. C. Williams, and G. M. Mervyn (2014). Survival Analysis for Com-posite Pavement Performance in Iowa. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C.
3. ZHIGUO LI1, SHIYU ZHOU1, SURESH CHOUBEY2 and CRISPIAN SIEVENPIPER2 (January 2006). Failure event prediction using the Cox proportional hazard model driven by frequent failure signatures, 303–315.
4. Chen, Y.-L., Provan G. (1997). Modeling and diagnosis of timed discrete event systems—a factory automation example. Presented at the American Control Conference, Albuquerque, NM.
5. Cox, D.R. (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society B, 34, 187–220.
6. George, L. (2003). Biomedical survival analysis vs. reliability: comparison, crossover, and advances. The Journal of the RAC, 11(4), 1–5.
7. Klein, J.P. and Moeschberger, M.L. (2003). Survival Analysis: Techniques for Censored and Truncated Data, Springer-Verlag, New York, NY.

8. Hougaard, P. (1999). Fundamentals of survival data. Biometrics, 55, 13–21.
9. Leffondre, K., Abrahamowicz, M. and Siemiatycki, J. (2003). Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: a simulation study. Statistics in Medicine, 22, 3781–3794.

A
Adhitya Ronnie Effendhie, 216
Aisha Turki, 419
Akram Musallam Alshawawreh, 308
Alan Huang, 282
Alban Cela, 382
Alexandru Cernat, 374
Ali Hebishi Kamel Abdelhamid, 190, 200
Ali Sulaiman Al Flaiti, 397
Amal Mansouri, 389
Ana Costa-Veiga, 225
Angelita P. Tobias, 405
Anisha P, 1
Anthony Davison, 290
Antonella Peruffo, 125
Areti Boulieri, 104
Arman Bidarbakhtnia, 8
Atikur R. Khan, 16
Atina Ahdika, 216
Aya Alwan, 282
B
Benjamin Ng, 159
Bruno Cozzi, 125
Bruno de Sousa, 225
C
Carla Nunes, 225
Carlos Pires, 225
Carlos Trucíos, 251
Chang Xie, 367
Chin Tsung Rern, 111
Chong Ning, 159
Christopher Ryan, 8
D
Damon Eisen, 46
Daniel J. Weiss, 80
Daniel Kilchmann, 350
David Harris, 65
David Johnson, 97
Dedi Rosadi, 216
Deemat C Mathew, 1
Dharini Pathmanathan, 111, 118
Dmitri Jdanov, 150
Domantas Jasilionis, 150
Dulce Gomes, 225
E
Elizabeth van der Merwe, 327
Emma McBryde, 46
Enrico Grisan, 125
Erniel B. Barrios, 405
Erwan Koc, 290
Ewan Cameron, 80
F
Francisco N. de los Reyes, 142
G
Gary Sharp, 327
Gerrit Grobler, 342
Gunardi, 216
H
Haakon Bakka, 314
Harry S. Gibson, 80
Havard Rue, 314
Hsein Kew, 65
I
Ibrahim Mohamed, 118, 258
Ivan Mizera, 134
J
Jamaludin Suhaila, 274
Janette Larney, 342
Jang Schiltz, 25
Jarod Y. L. Lee, 89
Jean-Marie Graïc, 125
Jennifer Rozier, 80
Jeremy Heng, 159
Jiefei Yang, 209
Jiti Gao, 65
Jittima Dummee, 56
João Mazzeu, 251
Jonathan Koh, 290
Joseph Ryan G. Lansangan, 405
Joseph W. Sakshaug, 374
Justin Wishart, 282
K
K. Hitomi[2], J. Tao, 335
K. Nagai, 335
Katherine E. Battle, 80
Katri Soinne, 299
Khaddouj Abu Baker Abdulla, 319
Klajd Shuka, 382
Kuntip Trongthamakit, 56
L
Lach Lachemot Tassadit, 173
Liu Huifeng, 72
Livio Corain, 125
Louise M. Ryan, 89
Ludovica Montanucci, 125
Luigi Salmaso, 125
Luiz Hotta, 251
M
M S Panwar, 412
M. Towhidul Islam, 16
Marc Hallin, 251
Markus Zwick, 165

Marta Blangiardo, 104
Matúš Maciak, 134
Mauricette Andriamananjara Nambinisoa, 80
Maurício Zevallos, 251
Michele Nguyen, 80
Mikhail Zhelonkin, 359
Muhammad Fauzee Hamdan, 274
N
Nandish Chattopadhyay, 34
Nugroho Puspito Yudho, 266
Nur Fatihah Mohd Ali, 258
Nurmitra Sari Purba, 266
Nurul Aityqah Yaacob, 111, 118
O
Ourbih Tari Megdouda, 173
Oyelola Adegboye, 46
P
Patrícia Filipe, 225
Pedro Valls, 251
Peter W. Gething, 80
Prajamitra Bhuyan, 34
Pranesh Kumar, 209
Prasada Rao, 367
R
Rahul Chattopadhyay, 426
Rani Nooraeni, 266
Rawia Wagih Abd ElMagid ElSayed Ragab, 190, 200
Robert Kohn, 89
Rodrigo Lovatón, 234
Rosalind E. Howes, 80
Rossita Mohamad Yunus, 258
Ruben Carvajal-Schiaffino, 125
S
Sadeg Ines Ines, 173
Saggou Hafida Hafida, 173
Salah Qudairi, Badreyya Al Shehhi, 419
Saleh Almansouri, 319
Saleheen Khan, 16
Scott A. Sisson, 89
Shafiqah Azman, 111
Sharita Serrao, 8
Sisa Pazi, 327
Siti Haslinda Mohd Din, 118
Solange Correa Onel, 97
Sudheesh K Kattumannil, 1
Sula Sarkar, 234
Sun Yunjie, 72
Suzanne Keddie, 80
T
Tabin Hasan, 16
Thomas Fung, 282
Thumna Al Rashedi, 319
Thumna Alrashdi, 419
Timothy C. D. Lucas, 80
V
Vincent Chin, 89
W
Wadeema Mohamed Alkhoori, 308
Weilun Zhou, 65
Wu Da, 72
X
Xuan Che, 181
Y
Y. Nishiyama, 335
Z
Zhongshan Yang, 367
Zhu Yingchun, 72

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 978-967-2000-69-3

9 789672 000693

#ISIWSC2019