62nd ISI World Statistics Congress

ISI 2019 Kuala Lumpur

18 - 23 August 2019

PROCEEDING

SPECIAL TOPIC SESSION

VOLUME 4

62nd ISI WORLD STATISTICS CONGRESS 2019

18 - 23 August 2019, Kuala Lumpur

Come | Connect | Create

www.isi2019.org

PROCEEDING

ISI WORLD STATISTICS CONGRESS 2019

SPECIAL TOPIC SESSION (VOLUME 4)

Published by:
Department of Statistics Malaysia
Block C6, Complex C
Federal Government Administrative Centre
62514 Putrajaya
MALAYSIA

Central Bank of Malaysia
Jalan Dato' Onn
P.O. Box 10922
50929 Kuala Lumpur
MALAYSIA

Malaysia Institute of Statistics
Department of Mathematical Sciences
Faculty of Sciences and Technology
43600 UKM Bangi, Selangor
MALAYSIA

Portal : https://www.isi2019.org
Email : lpc@isi2019.org

Published in February 2020

Suggested citation:
Department of Statistics Malaysia (DOSM). 2019. Proceeding of the 62nd ISI World Statistics Congress 2019: Special Topic Session: Volume 4, 2019. 419 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

Special Topic Session (STS): Volume 4

Preface i

Scientific Programme Committee ii

Local Programme Committee iii

STS556: Advanced Statistical modeling for data analysts

Recovery signal from noise of higher volatility 1

Trade credit – An empirical analysis of Indian firms 8

STS558: Development of End-To-End Integrated Solution for Statistical Reporting: Opportunities and Challenges

The integrated data submission/collection platform for regulator and statistical community 17

Handshakes corporate intelligence solution 25

STS560: Advancing to the Next Stage of Talent Analytics using Psychometric Assessments

Exploratory study of key traits for the fourth industrial revolution among employees of financial institutions in Malaysia 31

Using covert response biases in psychometric assessments to bolster job candidate interviews: an example with hospitality roles 39

STS563: Producing Population Estimates from Administrative Data Sources – How to Deal with Under- and Over-Coverage Error

Developing an integrated survey for admin-based population estimates and labour market statistics 48

Recent progress on implementing a Bayesian approach to population estimation from an administrative list subject to under and over-coverage 56

Population size estimation from incomplete multisource lists: A Bayesian perspective on latent class modelling 65

STS566: Forecasting for Currency Demand

Forecasting banknote demand at the Reserve Bank of Australia 73

Currency demand forecasting: The Philippine experience 82

Complex seasonal autoregressive model compared to machine learning methods for cash volume forecasting 89

STS570: Multinational Profit Shifting and Illicit Flows – Can We Measure Them?

The emergence of legal and organisational arrangements to minimise global tax burden, and its impact on monitoring domestic economic activities 99

Tracking the international footprints of global firms 108

STS571: Developing Mobile Positioning Data for Official Statistics: Experiences from Europe and Asia

Mobile phone and credit card data: Experience from 10 years of public private partnership 117

Combining mobile phone data and survey data for the best result: Experience from Indonesia 127

The Use of mobile phone data in Tourism Statistics 135

STS577: Recent Advances in Statistics and Computation

Improved robust rank-based test statistics in high-dimensional regression model 143

Perfection of volatility prediction with time scale information using wavelet transformation 150

STS579: Asymptotic Behavior and Numerical Simulation of Stochastic Evolutionary Systems

Stochastic evolutionary system on multidimensional lattices 156

Simulation of branching random walks with different intensity of branching sources 164

Survival analysis of particle populations in branching random walks 172

Civil registration and identity for all, a pathway to the Sustainable Development Goals (SDG's): Malaysia's perspective 180

STS580: Advances and Applications of Statistical Process Monitoring

An approach to monitoring multivariate time between events 194

Healthcare fraud detection using machine learning approaches 201

STS582: Statistical Analysis of Complex Data in Statistical Genetics and Bioinformatics

Proteogenomics: statistical issues in data integration and prediction 210

Mendelian randomization, causal relationship, and statistical approaches 219

STS583: Geospatial Information for International Statistics

Using GIS and Official Statistics to support assessments of risk to sustainable development from Environmental Degradation 228

Addressing the issue of missing or non-ideal sampling frames in household surveys in developing countries through remote sensing data 236

Implementing a geospatial data strategy in the European Statistical System 245

Cost-effectiveness of remote sensing for Agricultural Statistics in developing and transition countries 253

STS587: Measurement Issues in an Age of Digital Technologies

The accuracy and relevance of GDP measures in a digital economy 261

Measuring the structure of digital economy - The case of China 268

The impacts of digitalisation on China's economic system 279

STS637: National Statistical Offices in the Data Science Era

Using big data in monitoring indicators of sustainable development goals in Egypt 286

Statistics monopoly: No room for nostalgia 295

UN global platform as a data science collaboration environment for official statistics 304

STS700: Special STS - Using Big Data for Official Statistics – the Asia and Pacific experience

Measuring the rural access index in the Philippines 311

Big data utilization for Official Statistics in Thailand 317

Accommodating Big Data in Nepalese Statistical System: Challenges and opportunities 323

STS1080: Special STS: Data Science and Healthcare Informatics for the Developing Countries Special Session

Analysis of blockchain for healthcare applications 331

Data science and the Big Data framework for development, and to benefit from disruptive technology advances 342

Diet4You: A personal intelligent assistant for diets integrating data science 348

A survey of Machine Learning Algorithms for efficient biomarkers identification 357

STS2319: Asian Development Bank's support to innovative data collection and analytical methods

Measuring Rice Yield from Space: The Case of Thai Binh Province, Viet Nam 366

How much better is better? Quantifying the CAPI advantage using Viet Nam's Labor Force Survey 375

STS2320: Moving from Traditional Data to Big Data in Assessing Knowledge Societies

On the identification and handling of outliers in composite index data 384

Composite indices and traditional data – The global knowledge index 392

Moving from traditional data to big data in assessing knowledge societies 401

Index 409

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

Recovery signal from noise of higher volatility

Mohd Bakri Adam[2], Nurul Nisa' Khairol Azmi[1,2], Norhaslinda Ali[2], Mohd Shafie Mustafa[2]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Negeri Sembilan
[2]Institute for Mathematical Research, Universiti Putra Malaysia

Abstract

Compound smoother is a non-linear smoothing technique that has the ability to reduce heavy noise from signal and at the same time, is resistant to sudden changes and impulse in a data series. In this study, the compound smoother of 4253HT has been adjusted in the algorithm, specifically to estimate the middle point of running median for even span size by applying the following types of means; geometric, harmonic, quadratic and contraharmonic. The performance of smoothers in extracting the signal from heavy noise were assessed via simulation function of sinusoidal of high frequency plus trend with higher percentage contaminated normal noise added. The result found that modified 4253HT using geometric mean managed to smooth the data of high frequency with fluctuation effectively compared to the existing and others modification.

Keywords

compound smoother; running median; non-linear; 4253HT; signal; noise

1. Introduction

It is has been recognized that linear smoother is optimal to eliminate Gaussian noise and track trends that are common in practice, Bernholt et al. (2006). However, noise of high volatility tends to mask the general picture of a data series. The existence of non well-behaved noise makes the assumptions of linear model violated. Usually, least square estimation which is well known for its poor performance in the presence of outliers or long-tailed distribution data is used. According to Venetsanopoulos and Pitas (1990), linear smoothers also have a high tendency to blur important features and lack of the ability to remove impulsive noise. Not only that, linear smoothers are highly vulnerable to outliers and could not deal well with nonlinearity in a data series. Blurry edge which leads to the lost of important information is actually due to the sudden changes in a series, Bernholt et al. (2006). Due to its ability to remove non-Gaussian noise from a data series, median smoother is usually the favored smoothing tools. Unfortunately, median smoother tends to over smoothed a data series since it eliminates Gaussian noise too. One of established types of median smoothers which have been widely employed in various area settings

is compound smoother. compound smoother is known as a powerful tool to smooth a data series without excessively disrupting the details of a data series. Despite this good traits, the compound smoother does not respond well to oscillated trend, Tothmeresz and Erdei (1995) and Jin and Xu (2013). The number observations of compound smoother should be at least seven, otherwise it will converge to constant root, Janosky et al. (1997). The Velleman's compound smoother as indicated by Sargent and Bedford (2010) has been revised when possible combinations of multiple step of running median, Hanning and resmooth the rough are tested out. Improvement on the existing compound smoother in comparison has yet been explored.

2. Methodology

The 4253HT is one of the non-linear smoothing techniques that combined running median, weighted moving average and re-smoothed the rough. This technique was first introduces by (Tukey, 1977) and described detail in different version by (Velleman et al., 1981). Let **X** be a doubly-infinite sequence of real data $\{X_{t-n},..., X_{t-1}, X_t, X_{t+1},..., X_{t+n}\}$. A smoother *M* is defined as an algorithm that works on **X** to generate a new series $M(\mathbf{X}_t)$, smoothed values. The algorithm of 4253HT is as follows:

Step 1: Perform running median of span size two

$$M_1(\mathbf{X}_t) = \text{median}\left(X_{t-2}, X_{t-1}, X_t, X_{t+1}\right) \tag{1}$$

Step 2: Re-centered the equation (1)

$$M_2(\mathbf{X}_t) = \text{median}\left(M_1(\mathbf{X}_t), M_1(\mathbf{X}_{t+1})\right) \tag{2}$$

Step 3: Next, equation (2) are smoothed again by applying running median of span size three

$$M_3(\mathbf{X}_t) = \text{median}\left(M_2(\mathbf{X}_{t-2}), M_2(\mathbf{X}_{t-1}), M_2(\mathbf{X}_t), M_2(\mathbf{X}_{t+1}), M_2(\mathbf{X}_{t+2})\right) \tag{3}$$

Step 4: Perform running median of span size three

$$M_4(\mathbf{X}_t) = \text{median}\left(M_3(\mathbf{X}_{t-1}), M_3(\mathbf{X}_t), M_3(\mathbf{X}_{t+1})\right) \tag{4}$$

Step 5: Apply weighted moving average or Hanning with coefficients $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$

$$M_5(\mathbf{X}_t) = \frac{1}{4}M_4(\mathbf{X}_{t-1}) + \frac{1}{2}M_4(\mathbf{X}_t) + \frac{1}{4}M_4(\mathbf{X}_{t+1}) \tag{5}$$

Step 6: Re-smooth the rough and added to the smoothed values in (5)

$$M_6(\mathbf{X}_t) = M_5(\mathbf{X}_t) + M_5\left[\mathbf{X}_t - M_5(\mathbf{X}_t)\right] \tag{6}$$

The running median of even span size is computed by taking the average of two subsequent points in the middle using arithmetic mean. This value is better than running median of odd span size in the sense that it preserves the significant spike in the data series. The smooth value produce by running median span size four and re-centered by running median of span size two is a combination of Equation (1) and (2) and can be expressed as follows:

$$\begin{aligned} M_2(\mathbf{X}_t) &= \frac{1}{4}\left[\text{median}\left(X_{t-2}, X_{t-1}, X_t, X_{t+1}\right) + \text{median}\left(X_{t-1}, X_t, X_{t+1}, X_{t+2}\right)\right] \\ &= \frac{1}{4}\left(X^*_{t-1} + X^*_t + X^{'}_t + X^{'}_{t+1}\right) \end{aligned} \tag{7}$$

where X^* is the ordered observation from window in $X_{t-2}, X_{t-1}, X_t, X_{t+1}$ and $X^{'}$ is the ordered observation from window in $X_{t-1}, X_t, X_{t+1}, X_{t+2}$.

Some adjustments are proposed by applying different type of mean. The type of means involved are geometric, quadratic, harmonic and contra harmonic. The modification of running median span size 42 are as follows:

Geometric Mean

$$M_2(\mathbf{X}_t) = \left(X^*_{t-1} \times X^*_t \times X^{'}_t \times X^{'}_{t+1}\right)^{\frac{1}{4}} \tag{8}$$

Quadratic Mean

$$M_2(\mathbf{X}_t) = \left(\frac{X^{*2}_{t-1} + X^{*2}_t + X^{'2}_t + X^{'2}_{t+1}}{4}\right)^{\frac{1}{2}} \tag{9}$$

Harmonic Mean

$$M_2(\mathbf{X}_t) = \frac{1}{4}\left(\frac{1}{X^*_{t-1}} + \frac{1}{X^*_t} + \frac{1}{X^{'}_t} + \frac{1}{X^{'}_{t+1}}\right) \tag{10}$$

Contra harmonic Mean

$$M_2(\mathbf{X}_t) = \left(\frac{X^{*2}_{t-1} + X^{*2}_t + X^{'2}_t + X^{'2}_{t+1}}{X^*_{t-1} + X^*_t + X^{'}_t + X^{'}_{t+1}}\right) \tag{11}$$

The types of mean that produce smaller value than arithmetic mean; geometric and harmonic are expected to be more resistant to negative impulse or block pulse. On the other hand, quadratic and contra harmonic are more responsive to positive changes in the data series. Some of the modifications are not working if the observations consist of zero or negative values. Hence, a constant point should be added to the data to ensure the smoothed value can be computed. Modification of 4253HT involves the running median of span size 42 only. For smoothing the rough part, original algorithm is maintained where the middle points for running median of span size four and two are computed by arithmetic mean. The evaluation process

of smoothing is done by simulation of signal and noise. The process of simulation is based on procedure from Conradie et al., 2009. Generally, data can be decomposed into the following components:

$$\text{Data}_t = \text{Signal}_t + \text{Noise}_t = X_t \tag{12}$$

The signal is a combination of sinusoidal function with linear curve:

$$\text{Signal}_t = \mu_t = \eta t + A\sin B(t-C) \tag{13}$$

with η is the slope of trend, t is the index, $|A|$ is an amplitude, $B=\frac{2\pi}{d}$ where d is the period and frequency is $\frac{1}{d}$, and C represent the displacement. Hence,

$$\begin{aligned} X_t &= \mu_t + D_t \\ &= \eta t + A\sin B(t-C) + D_t \end{aligned} \tag{14}$$

For the sine function, let $\eta = 0.7$, the amplitude |A|=3 and the displacement C=1. The parameter η, |A| and C were chosen according to Conradie et al., 2009. This parameter values will produce a smooth sine curve. Two hundred values from function $\mu_t = \eta t + A\sin B(t-C)$ were simulated for t between 0.542 and 19.6416 with increments of 0.2 at high frequency which is $\frac{13}{16}$. This high frequency is very difficult to be extract since the wavelength and noise tend to mixed up. Figure 1 shows a sinusoidal of frequency $\frac{13}{16}$ with linear curve.

The noise, $\{D_t\}$ were generated as identically and independently random variables from contaminated normal distribution as

$$D_t = \begin{cases} \alpha Z_t & \text{if } Y_t = 1, \\ \beta Z_t & \text{if } Y_t = 0 \end{cases} \tag{15}$$

with $\{Y_t\}$ i.i.d Bernoulli(p) and independent of the $\{Z_t\}$. Thus $P(Y=1)=p$ and $P(Y=0)=1-p$ so that

$$\begin{aligned} P(D_t \le d) &= P(\alpha Z_t \le d \mid Y_t = 1)P(Y_t = 1) + P(\beta Z_t \le d \mid Y_t = 0)P(Y_t = 0) \\ &= p\Phi\left(\frac{d}{\alpha}\right) + (1-p)\Phi\left(\frac{d}{\beta}\right). \end{aligned} \tag{16}$$

with $\{Z_t\}$ i.i.d $N(0,1)$. To generate noise with high volatility, let $\alpha = 5.06$ and p=0.75, so that Var(X)=(0.75)(5.06)^2 + 0.25 = 23.29. In the simulation of generating high volatility noise, approximately 75% of the values come from a $N(0,5.06^2)$ distribution and approximately 25% from a $N(0,1)$ distribution.

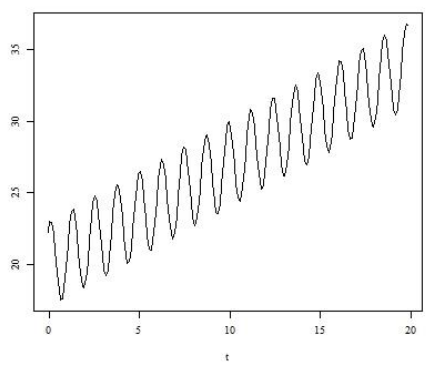


Figure 1 Sinusoidal function of frequency 13/16 with linear trend

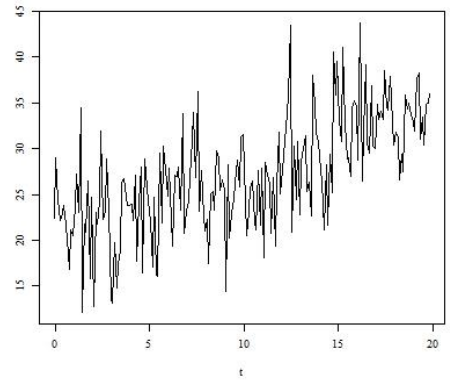


Figure 2 Sinusoidal function of frequency 13/16 plus linear trend with 75% contaminated normal noise added

Figure 2 depicts the sinusoidal of frequency plus trend with 75% contaminated normal noise added. It is hardly to capture the general trend and existence of seasonal oscillation with 75% contaminated normal noise added. Two hundred signals plus the generated noise were simulated and applied the existing and modified 4253HT smoother. The performances of these smoothers are evaluated by regression coefficient. Consider the following linear regression model with one independent variable:

$$Y_{ij} = \beta^* + \beta_i \mu_j + \varepsilon_{ij}, \qquad i = 1,..., N;\ \ j = 1,..., k \tag{17}$$

The closer the regression coefficient to one indicates that the signal has been extracted from noise very well. If the value of the regression coefficient is close to zero, a smoother performs poorly in recovery the signal from noise.

3. Results and Discussion

Table 1 shows the performance of smoothers measured via regression coefficient. The modified smoother using geometric mean was found to be the best avenue to extract sinusoidal signal of frequency $\frac{13}{16}$ from heavy noise. This was vouched by the value of regression coefficient that closest to 1.

Table 1 Performance of existing and modified smoother measured by regression coefficient

Type of modification	Regression Coefficient
Arithmetic	0.9605
Geometric	**0.9622**
Quadratic	0.9601
Harmonic	0.9615
Contra harmonic	0.9610

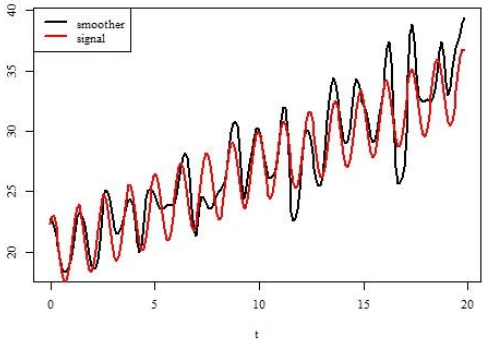


Figure 3 Plot signal versus modified compound smoother using geometric mean

The results, supported by the graphical analysis in Figure 3, demonstrates that the smoother has the capability to successfully recover the signal from noise of high volatility. Therefore, the main features of the signal were maintained, resulting in the further analysis such as model estimation, to be less complicated.

4. Conclusion

This study is mainly to assess the performance of modified 4253HT in capturing sinusoidal plus linear trend signal with heavy noise added. Noise with high volatility was added to the signal and the performances were measured by recruiting regression coefficient. The results show that modified 4253HT using geometric mean performed the best in extracting signal from heavy noise. For future works, the performance of proposed adjustment to compound smoother will be assessed with the inclusion of different types of signals and noise.

References

1. Bernholt, T., Fried, R., Gather, U. and Wegener, I. 2006. Modified repeated median filters. *Statistics and Computing* 16 (2): 177-192.
2. Conradie, W., De Wet, T. and Jankowitz, M. D. (2009). Performance of nonlinear smoothers in signal recovery. *Applied Stochastic Models in Business and Industry* 25 (4): 425-444.
3. Janosky, J., Pellitieri, T. and Al-Shboul, Q. (1997). The need for a revised lower limit for the 4253H, Twice nonparametric smoother. *Statistics & Probability Letters* 32 (3): 269-272.
4. Jin, Z. and Xu, B. (2013). A novel compound smoother RMMEH to reconstruct MODIS NDVI time series. IEEE Geoscience and Remote Sensing Letters 10 (4): 942-946.
5. Sargent, J. and Bedford, A. (2010). Improving Australian Football League player performance forecasts using optimized nonlinear smoothing. *International Journal of Forecasting* 26 (3): 489-497.
6. Tothmeresz, B. and Erdei, Z. (1995). New features of MULTIPATTERN 1.10: Robust Nonlinear Smoothing. *Tiscia* 29: 33-36.
7. Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Readings.
8. Velleman, P. F. and Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis.* Duxbury Press, Boston, Massachusetts.
9. Venetsanopoulos, A. and Pitas, I. (1990). *Nonlinear Digital Filters*. 1st edn. New York: Springer US.

Trade credit – An empirical analysis of Indian firms

Nitin Kumar, Arvind Shrivastava, Purnendu Kumar
Reserve Bank of India, Mumbai, India

Abstract

The paper examines determinants of trade credit for Indian firms. The dataset builds on annual firm specific variables and macro indicators spanning from 2004 to 2017 to understand their impact on extending and usage of trade credit amongst Indian firms. Applying dynamic panel framework it is found that inventory management and macro indicators are significant in determining trade credit for Indian firms. Larger firms are found to be leading suppliers of trade credit. Pecking order theory is clearly validated with net earnings being preferred over trade credit that is a more expensive source of finance. Significant decline is trade credit business is evidenced post crisis. Analysis of trade credit ratio has also been performed across constrained/ unconstrained firms that show variation. Trade credit behavior for distressed firms has also been examined. Distressed firms are evidenced to be involved in greater risk behavior having higher accounts payables liability.

Keywords

bank borrowing, firm distress, macro economy, panel data, generalized method of moments

JEL Classification: G3, G21, E4, C23

1. Introduction

Trade credit liabilities and assets in the form of account payables and account receivables are critical financing instruments depicting sources and uses of funds for firms. Trade credit is a short-term cash management tool wherein a firm can be seen both as a supplier and customer together. Literature supports the view that suppliers lend more liberally to borrowers especially during downturns due to availability of superior information about credit worthiness of borrowers overcoming moral hazard problem of lending (Smith, 1987; Biais and Golliers, 1997; Petersen and Rajan, 1997; Burkart and Ellingsen, 2004). In inefficient and less developed financial systems exhibiting tightened credit and monetary policies, firms may have to seek alternative sources of funds. Despite being costly post discount period, trade credit is vital option of finance. However, limited attention is provided to trade credit size and analysis. Employing global firm level datasets for later part of 1980s, Petersen and Rajan (1997) reported that accounts receivable to sales ratio for

small firms in US stood at 7.3 per cent versus 18.5 per cent for larger firms. Likewise, there existed predominance of trade credit amongst bigger firms, with accounts payables to sales ratio being only 4.4 per cent for small firms compared to 11.6 per cent for larger firms. For firms operating in UK's manufacturing sector over 1993-2003, Bougheas et al. (2009) reported account receivables to sales of 17 per cent whereas account payables to sales stood at 10 per cent. Using Euro area firm-level data Casey and O'Toole (2014) found highest usage of trade credit in Ireland, a country which has simultaneously experienced a severe banking crisis and sovereign debt funding crisis with 75 per cent of firms opting for alternative finance. As per Ghosh (2015), trade credit usage as percentage of total funding was roughly 16 per cent for Indian manufacturing firms during 1993–2012. The study discusses empirical methodology in Section 2. Analysis and results are provided in Section 3 and Section 4 concludes the findings.

2. Methodology

The study is based on information of non-government and non-financial public limited companies collated from their annual reports/balance sheets as carried out in Company Finance Division, Reserve Bank of India database on corporate sector. The state-owned and financial sector firms have been excluded from analysis due to their varied social objectives and separate regulatory structure. Our database is a balanced panel of 979 firms from 2003-04 to 2016-17 i.e. fourteen years annual figures. Account receivables (AR) and account payables (AP) normalized by assets constitute our dependent variables. The common set of explanatory variables are stock of inventory (INV) that is critical parameter that may drive trade credit in either direction and act as collateral to obtain trade credit and has been found to have positive relation with accounts payables by Cunat (2007) and negative effect on account receivables Bougheas et al. (2009). Size of firm (SIZE) is captured by natural logarithm of real sales, is a proxy for creditworthiness and reputation of a firm (Petersen and Rajan, 1997). Impact of profitability is measured by return on assets (ROA). Leverage is a vital financial indicator measured as debt to asset ratio (DEBT) that captures financial soundness/riskiness of firm. Bank borrowing (BORR) is defined as bank borrowings to total borrowings of a firm. Current assets to total assets (CATA) has been included to measure liquidity. GDP growth rate (GR_RATE), GDP deflator (INF) and weighted average call money rate (INT_RATE) are taken as macroeconomic indicators for macro growth, inflation and interest rate respectively. The lagged dependent variable and possible endogeneity of regressors renders ordinary least square estimation producing biased estimates due to correlation between lagged dependent variable and errors. So, Generalized Method of Moments (GMM) estimator developed for dynamic panel data, introduced by Arellano and Bond

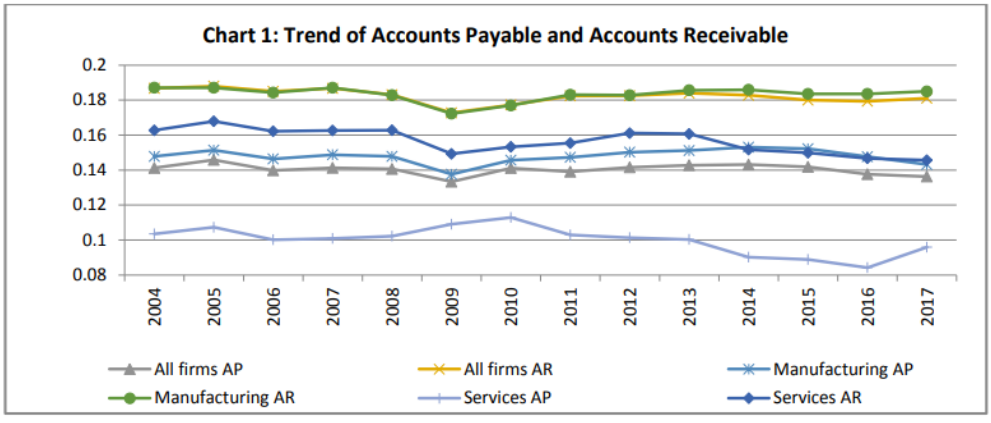
(1991) and Arellano and Bover (1995) has been employed to estimate casual effect of regressor on dependent variable, formulated as follows.

$$y_{i,t} = \rho y_{i,t-1} + X_{i,t}\beta + \epsilon_{i,t} \quad \text{where } \epsilon_{i,t} \sim N(0, \tau^{-1}) \qquad \ldots (1)$$

The lagged endogenous variable is represented by $y_{i,\,t-1}$ with $X_{i,\,t}$ being matrix of other exogenous variables as explained in previous section.

3. Result

As per summary results, at gross level average accounts receivables are observed to be 18 per cent that is higher compared to average accounts payables at 14 per cent for Indian firms during the study period. It signifies that at gross level the firms in sample have applied financial strategy to sell their goods. Temporal behavior of accounts payables and account receivables shows that Indian firms have been net accounts receiver throughout the period (Chart 1). It is observed that there has been slight dip in trade credit post financial crisis with AP falling from 14.0 per cent in 2008 to 13.3 per cent in 2009. Similarly, AR registered decline from 18.3 per cent in 2008 to 17.2 per cent in 2009.



The dynamic panel regression model applying GMM estimator with accounts payables as dependent variable is reported in Table 1. Separate regressions were performed for all firms, major sectors viz., manufacturing, services. Within manufacturing separate regression are performed for chemical product firms, textiles to detect possible variations, if any. Finally, separate regression has been done for small and large firms also. Firms existing in the top/bottom 25 percentile as per total assets are chosen for small/large classification. Most of estimates are significant signifying satisfactory fit.

Table 1: Accounts payables							
Variables	All	Manufacturing	Services	Chemical Product	Textiles	Small	Large
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	0.025*	0.015	0.048	-0.014	0.072**	-0.014	0.057
	(0.014)	(0.016)	(0.034)	(0.037)	(0.032)	(0.028)	(0.036)
L1.Dep	0.564***	0.602***	0.514***	0.437***	0.414***	0.479***	0.467***
	(0.017)	(0.019)	(0.044)	(0.042)	(0.038)	(0.041)	(0.034)
INV	0.003**	0.004*	0.002	6.8E-04	0.012***	6.8E-04	0.004
	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)	(0.002)	(0.003)
SIZE	-0.002	-0.001	-0.005	0.013*	-0.022***	0.004	-0.002
	(0.002)	(0.003)	(0.005)	(0.007)	(0.006)	(0.003)	(0.005)
ROA	-0.034***	-0.023**	-0.081***	-0.059**	-0.038*	-0.041***	-8.3E-04
	(0.01)	(0.011)	(0.022)	(0.026)	(0.021)	(0.015)	(0.019)
DEBT	-0.025***	-0.031***	0.014	-0.055**	-0.004	-0.012	-0.082***
	(0.007)	(0.008)	(0.019)	(0.023)	(0.014)	(0.012)	(0.014)
BORR	0.002	0.003	2.6E-04	0.007	-0.003	0.01**	-0.003
	(0.003)	(0.003)	(0.006)	(0.007)	(0.007)	(0.005)	(0.005)
CATA	0.111***	0.117***	0.046***	0.106***	0.075***	0.119***	0.06***
	(0.009)	(0.01)	(0.016)	(0.02)	(0.02)	(0.017)	(0.013)
L1.INF	-0.008***	-0.007**	-0.011	-0.002	0.014**	0.006	-0.008
	(0.003)	(0.003)	(0.008)	(0.008)	(0.007)	(0.01)	(0.005)
L1.INT_RATE	8.2E-04*	4.9E-04	0.001	-6.9E-06	-7.2E-04	-2.9E-04	7.6E-04
	(4.3E-04)	(4.9E-04)	(0.001)	(0.001)	(0.001)	(0.001)	(6.7E-04)
L1.GR_RATE	-0.001***	-0.002***	0.002	-0.002*	3.3E-04	-8.7E-04	-5.2E-04
	(4.8E-04)	(5.4E-04)	(0.001)	(0.001)	(0.001)	(0.001)	(7.3E-04)
Wald Statistics	1480***	1333***	206***	201***	192***	213***	263***
All specifications are estimated using GMM first-difference specification. *, **, *** indicate significance at 10%, 5%, and 1% level, respectively.							

Initiating with column (1), estimation results for entire sample shows positive and significant lag coefficient reflecting strong persistence. The significant and positive effect of inventories on accounts payables indicates higher usage of trade credit towards accumulating inventories validating Vaidya (2011). The finding is in consonance with ROA is having strong inverse impact on accounts payables implying profitable firms immediately repaying trade credit to rid from this expensive form of credit. The coefficient of debt to asset ratio is negative and significant. A higher debt financing is leading to reduced trade credit borrowing indicating debt financing used as substitute vis-a-vis trade credit. Both ROA and DEBT are in harmony with pecking order theory that postulates internally generated funds to be higher in order

compared to more expensive forms of credit (Myers and Majluf, 1984). Liquidity of firm as captured by current ratio is significant and positive. Higher liquidity affords higher trade credit liability to a firm. The finding is consistent as per Bougheas et al. (2009) and Vaidya (2011). Amongst macro indicators higher interest rate is having a positive influence on account payables. Higher rates narrow down the rate gap between formal sources of credit and trade credit leading to greater usage of trade credit that is relatively convenient. However, the role of both inflation and growth rate is negative dependent variable.

Continuing with other columns of Table 1, it is found that most of the results obtained for all the firms hold for other classifications also. Size of inventory although insignificant for service sector is positive and significant for manufacturing sector. The service sector like software firms, trade, communications, and transportation predominantly comprises of intangible goods where the role of physical inventory is limited. SIZE variable is positive and significant for chemical product firms, implying larger firms receiving more trade credit due to their creditworthiness and reputational considerations with potential buyers (Petersen and Rajan, 1997). Bougheas et al. (2009) found positive relation of size with both forms of trade credit although Vaidya (2011) found it significant only for account receivable. Debt to asset ratio is negative but insignificant for service sector turns to be negative and significant for entire sample and manufacturing sector also. Bank borrowing is recorded to be insignificant for most of firm classifications leading to inconclusive outcome. Inflation is having strong negative impact for both manufacturing and entire sample. Higher inflation is leading to lesser trade credit liability due to decline in real value of outstanding credit. Significant Wald statistics indicate rejection of null of parameter values being zero.

Table 2: Accounts receivable

Variables	All	Manufacturing	Services	Chemical Product	Textiles	Small firms	Large firms
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	-0.026**	-0.031**	0.016	-0.03	-0.061**	-0.009	-0.047
	(0.012)	(0.013)	(0.034)	(0.032)	(0.026)	(0.026)	(0.034)
L1.Dep	0.554***	0.535***	0.507***	0.371***	0.431***	0.397***	0.643***
	(0.018)	(0.02)	(0.052)	(0.04)	(0.042)	(0.041)	(0.034)
INV	-0.014***	-0.016***	-0.009***	-0.016***	-0.021***	-0.014***	-0.005**
	(0.001)	(0.002)	(0.003)	(0.004)	(0.003)	(0.002)	(0.003)
SIZE	0.021***	0.023***	0.014***	0.032***	0.03***	0.019***	0.016***
	(0.002)	(0.003)	(0.005)	(0.006)	(0.005)	(0.003)	(0.005)
ROA	0.028***	0.03***	0.008	-0.033	0.1***	0.005	0.049***

	(0.009)	(0.01)	(0.021)	(0.023)	(0.019)	(0.015)	(0.017)
DEBT	-0.022***	-0.021***	-0.039**	-0.035*	0.002	-0.02*	-0.003
	(0.007)	(0.007)	(0.02)	(0.021)	(0.012)	(0.012)	(0.012)
BORR	0.004*	0.004	-0.003	7.7E-04	0.012**	0.001	-6.7E-04
	(0.002)	(0.003)	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)
CATA	0.15***	0.164***	0.092***	0.197***	0.117***	0.255***	0.073***
	(0.008)	(0.009)	(0.016)	(0.019)	(0.018)	(0.016)	(0.012)
L1.INF	-0.005*	-0.006**	-0.008	-0.025***	0.003	-0.015	0.003
	(0.003)	(0.003)	(0.008)	(0.007)	(0.006)	(0.009)	(0.005)
L1.INT_RATE	-0.001***	-0.001**	-0.001	7.3E-04	6.0E-04	-0.004***	-0.002***
	(3.9E-04)	(4.2E-04)	(0.001)	(0.001)	(9.0E-04)	(0.001)	(6.1E-04)
L1.GR_RATE	-0.002***	-0.002***	-6.9E-04	-0.002*	0.002**	-2.0E-04	-0.001**
	(4.2E-04)	(4.6E-04)	(0.001)	(0.001)	(0.001)	(0.001)	(6.7E-04)
Wald Statistics	1992***	1633***	184***	323***	365***	428***	533***

Table 2 reports estimation results with accounts receivables as dependent variable. As earlier, columns 1, 2, 3 denote full sample, manufacturing, services respectively. A significant positive estimate for lagged dependent is evident for all firm classifications. Unlike for account payables size of inventory is having strong negative effect on account receivable. The outcomes suggest that firms having reasonable stock have less incentive to offer credit for obtaining additional stock leading to inverse relationship also supported by Bougheas et al. (2009) and Vaidya (2011). A direct positive impact of SIZE on account receivables are distinctly visible. The finding confirms that bigger firms are also the biggest lenders of trade credit in line with Petersen and Rajan (1997), Bougheas et al. (2009) and Vaidya (2011). Profitability as captured by ROA is having significant positive influence on trade credit receivable. It represents net earnings are channelized towards extending more credit. The finding is contrary to Vaidya (2011) that found significant negative impact of net profits, however in line with Petersen and Rajan (1997), Bougheas et al. (2009). Bank borrowings' coefficient is positive and significant implying firms borrowing more are also extending more trade credit. It points that bank borrowing and account receivables are complement in existing scenario. Liquidity (CATA) has strong positive impact on account receivable. It corroborates usage of additional financial resources towards extending credit to potential clients that is in tandem with Vaidya (2011). The coefficient of macroeconomic growth rate is having negative sign. The result portrays reduction in accounts receivables with general increase in income level. Inflation also has negative and significant coefficient for complete sample. It portrays decline in accounts receivables with increase in inflation due to decline in the real value of outstanding dues. Most of results obtained for entire sample are in harmony with other classifications.

Next, we examine if there exist any differential behaviour across financially constrained firms wherein constrained firms are assumed for whom external financing is either not available or available at high premium. As per the underlying idea of Acharya et al. (2007), we segregate top and bottom 25 per cent percentile firms according to BORR (bank borrowings/total borrowings) as a measure of tendency towards external financing. The top 25 percentile firms are defined as constrained firms whereas the bottom 25 percentile firms are categorized as unconstrained firms. An advantage of this approach is that it obviates any type of summarization as generally required in capital expenditure technique. The AP ratio varies marginally from 13 per cent to 14 per cent for constrained and unconstrained firms respectively (Table 3). However, AR ratio is observed 16 per cent for constrained and 20 per cent for unconstrained firms.

Table 3: Constrained v/s Unconstrained					Distress v/s Non-distress	
	Accounts Payables		Accounts Receivable		Account Payables	Account Receivable
Variables	Constrained	Unconstrained	Constrained	Unconstrained	Distressed Firms	Distressed Firms
Intercept	0.018	0.03	-0.011	-0.023	0.022	-0.022*
	(0.031)	(0.03)	(0.03)	(0.027)	(0.015)	(0.012)
L1.Dep	0.358***	0.435***	0.449***	0.36***	0.63***	0.555***
	(0.039)	(0.056)	(0.052)	(0.054)	(0.015)	(0.018)
INV	0.002	0.006*	-0.008***	-0.023***	0.002	-0.014***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.001)
SIZE	0.005	0.004	0.014***	0.032***	-0.002	0.022***
	(0.004)	(0.005)	(0.004)	(0.005)	(0.003)	(0.002)
ROA	-0.035**	-0.037*	0.005	0.063***	-0.064***	0.03***
	(0.016)	(0.019)	(0.016)	(0.017)	(0.01)	(0.009)
DEBT	0.028**	-0.083***	-0.021*	-0.026**	-0.024***	-0.023***
	(0.011)	(0.015)	(0.011)	(0.013)	(0.008)	(0.007)
BORR	0.002	0.01**	0.01***	-0.008*	0.001	0.004*
	(0.004)	(0.005)	(0.004)	(0.004)	(0.003)	(0.002)
CATA	0.097***	0.077***	0.125***	0.171***	0.112***	0.15***
	(0.014)	(0.015)	(0.014)	(0.014)	(0.009)	(0.008)
L1.INF	-0.014	-0.032***	0.018	-3.3E-04	0.017**	-0.01
	(0.012)	(0.01)	(0.012)	(0.008)	(0.007)	(0.006)
DISTRESS					-0.014***	-0.005*
					(0.003)	(0.003)
L1.INT_RAT	0.003***	0.001*	-0.003***	-8.7E-04	0.001***	-0.001***
	(0.001)	(7.4E-04)	(1.0E-03)	(6.4E-04)	(4.7E-04)	(3.9E-04)

L1.GR_RATE	-0.001	-2.9E-04	-0.002*	-0.002**	-0.001**	-0.002***
	(0.001)	(8.6E-04)	(0.001)	(7.6E-04)	(5.1E-04)	(4.3E-04)
Wald	173***	149***	213***	352***	2212***	2009***

Separate regressions have been repeated to evaluate if the trade credit ratios depict significant variation for constrained vis-à-vis unconstrained firms (Table 3). It is found that for unconstrained firms higher bank borrowings is leading to higher accounts payables. For accounts receivables higher borrowings translates to higher accounts receivables for constrained firms whereas the relationship is inverse for unconstrained firms. Last but not least, we analyze the impact of firms' financial distress on trade credit behavior broadly. In this regard, foremost, firms are classified into two groups viz., non-distress or distress using the criteria adopted by Shriavstava. et. al. (2008). The predicted probability values so obtained using logit regression included as an additional explanatory variable in model to assess its impact on trade credit operations. The regression results exhibiting the impact of distress indicator reveals a significant positive effect of financial distress on accounts payables (Table 3). The finding indicates that higher possibility of bankruptcy may lead a firm to greater risk taking by increasing its trade payables liability. However, as regards accounts receivable, although the relationship is negative, it is insignificant implying inconclusive effect of distress in case of accounts receivable.

4. Discussion and Conclusion

Trade credit has been a convenient source of financing especially for firms constrained by formal sources of borrowing. Applying dynamic panel generalized method of moment methodology it is found that pecking order theory is strongly validated with net earnings being preferred source of financing compared to trade credit that is a more expensive source of financing. Bank borrowing is complementing trade credit receivables for the sample of firms. Macro indicators like growth rate, inflation and interest rate are significant for trade credit decisions. It is found that for unconstrained firms higher bank borrowings is leading to higher accounts payables. For accounts receivables higher borrowings translates to higher accounts receivables for constrained firms whereas the relationship is inverse for unconstrained firms. The finding further indicates that higher possibility of bankruptcy may lead a firm to greater risk taking by increasing its trade payables liability. However, as regards accounts receivable, although the relationship is negative, it is insignificant implying inconclusive effect of distress in case of accounts receivable.

References

1. Acharya. V.V., Almeida. H. & Campello. M. (2007). Is cash negative debt? A hedging perspective on corporate financial policies. Journal of Financial Intermediation, 16(4), pp.515-554.
2. Arellano. M.. Bond. S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies, 58, 277-97.
3. Arellano. M.. Bover. O. (1995). Another look at the instrumental-variable estimation of error-components models. Journal of Econometrics, 68, 29-52.
4. Biais. B.. Gollier. C. (1997). Trade credit and credit rationing. Review of Financial Studies, 10(4), 903-937.
5. Bougheas, S., Mateut. S. & Mizen. P. (2009). Corporate trade credit and inventories: New evidence of a trade-off from accounts payable and receivable. Journal of Banking & Finance, 33(2), 300-307.
6. Burkart. M.. Ellingsen. T. (2004). In-Kind Finance. A Theory of Trade Credit, American Economic Review, 94(3), 569-590.
7. Casey. E.. O'Toole. C.M. (2014). Bank lending constraints, trade credit and alternative financing during the financial crisis: Evidence from European SMEs. Journal of Corporate Finance, 27, 173193.
8. Cunat. V. (2007). Trade credit: Suppliers as debt collectors and insurance providers. Review of Financial Studies 20, 491–527.
9. Ghosh.S. (2015). Trade Credit, Bank Credit and Crisis: Some Empirical Evidence for India. Margin: The Journal of Applied Economic Research, 9(4), pp.333-361.
10. Myers. S.C.. Majluf N.S. (1984). Corporate financing and investment decisions when firms have information that investors do not have. Journal of Financial Economics, 13(2), 187–221.
11. Petersen. M.A.. Rajan, R.G. (1997). Trade credit: theories and evidence. Review of financial studies, 10(3), 661-691.
12. Rajan. R. G.. Zingales. L. (1998). Financial Dependence and Growth. American Economic Review, 88(3), 559–586.
13. Shrivastava. A., Kumar. N. & Kumar. P. (2017). Bayesian analysis of working capital management on corporate profitability: evidence from India. Journal of Economic Studies, 44(4), pp.568-584.
14. Shrivastava. A., Kumar. K. & Kumar. N. (2018). Business distress prediction using Bayesian Logistic model of Indian Firms. Risks.
15. Smith. J. (1987). Trade Credit and Information Asymmetry. The Journal of Finance, 4,863-869.
16. Vaidya. R.R. (2011). The determinants of trade credit: Evidence from Indian manufacturing firms. Modern Economy, 2(05), p.707.

The integrated data submission/collection platform for regulator and statistical community

Alex Khor, Chan Foo Keong
Ocean Bridge Sdn Bhd

Abstract

Ocean Bridge has been positioned itself in the market as one of the leading software development company specialised in the development and implementation of Credit Bureau Applications, Data Submission Systems and Data Management Services. We also provide IT Advisory and Professional Services, System Integration Services and Custom Turn-Key Solutions and Software Development Outsourcing Services for our prestige customers. In this paper, Ocean Bridge will share its experience implementing Data Submission Systems and share its view on what an ideal Data Submission System should have and as considerations in future implementation for regulators and statistical bodies.

At present, many regulators and statistical bodies still use multiple platforms for collecting various data from respective Reporting Entities (REs)/filers and data providers. This approach is not only costly and inflexible but is hindering the adoption and implementation of a model-driven data submission or reporting solution which has proven to be capable of simplifying the preparation of submission template, data submission process and the consumption of data and sharing of information by the regulators and statistical bodies.

The ideal Data Submission solution is a credible, secured and highly scalable system which can fit well as an integral part of a holistic Integrated Information Management Platform (end-to-end data management platform covering data capturing, transformation and dissemination) to overcome the challenges of using multiple-platforms to cater for the data of diverse subject areas in the data collection process.

Keywords

data submission; submission platform; submission framework; submission history; content processor

1. Introduction

There are various challenges which are faced by regulators and statistical bodies today for collecting data in various data formats and from disperse type of reporting entities/filers. To implement a unified "One-for-All" Data Submission System, a model-driven submission framework is required empower the data submission process and cope with the ongoing data

requirement change and is capable to fit or integrate with the existing applications and infrastructure of the regulator and statistical bodies.

The Data Submission Platform is an integral part of the end-to-end Integrated Information Management Platform which provides the inevitable functionalities for data collection, data processing, data sourcing, data processing and data staging before the data is transported/transformed and loaded into the back-end Data Lake/Data Warehouse for generation of useful information and reports.

Being a single platform for all type of submission data, the system must be inclusive to cater for all data submission needs of the regulators and statistical bodies. With all its operational functions standardised, it is independent from any specific data / content of various subject areas, so as to enable future scaling to other data / content subject areas. The intention is to move away from silo'ed submission systems by standardising and consolidating generic submission functions across all subject areas.

The platform adopts a proven Submission Framework and Methodology, and it is build based on the cutting-edge technology and carefully-engineered algorithms which will overcome the following challenges:

1. **Submission Operational Efficiency**

   With the Standardised and unified Submission Framework as the base layer, the submission obligation is configurable to cater for the diversification of all type of data submission or reporting needs. Reduced reporting overhead for system owner and Reporting Entity (RE) via enabling submission through a single reporting or submission platform and via a single point of entry.

2. **Submission Processing Efficiency**

   The platform adopts Content-free concept to support multiple data submission or reporting channels and can easily integrate with the Master Data Management System (Taxonomy Management, Reference Registry, Entity Database and Metadata Repository etc) and respective content processors. It supports two-way integration with various Content Processors to enable end-to-end submission status tracking and facilitates the complete loop of submission flow.

3. **Streamlined Submission Process**

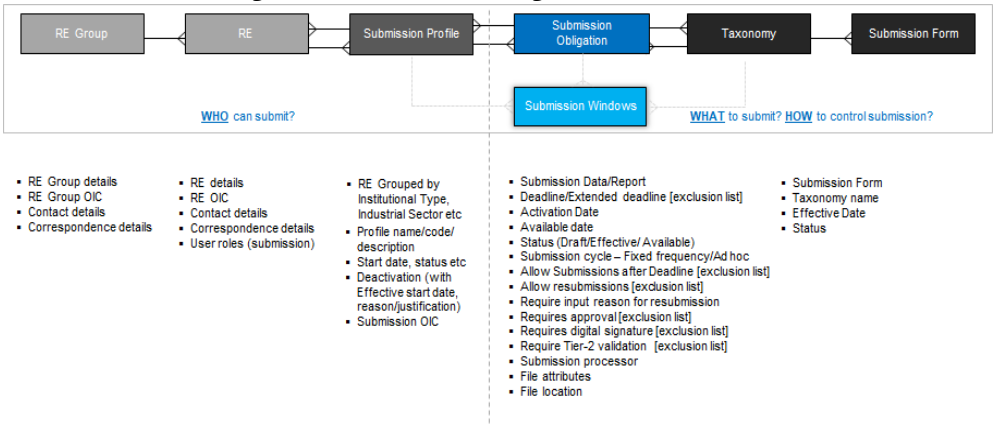
   The platform's Functional Model is a thinking model used to manage and coordinate the solutions around people, process and technology. It focuses in collecting data of various formats (by subject areas) from respective REs or filers but can integrate with various internal business support systems to facilitate master data sourcing, security authentication, workflow and the Content Processors for processing the submitted data.

2. Methodology

The platform provides the capabilities to support diversified and ongoing data consumption requirements of regulators and statistical bodies. This is made possible by adopting a proven Submission Obligation Framework which generalise all requirements of Data Submission process for most, if not all type of subject areas.

The platforms provide a self-service capability for the system owners to configure what and when to submit, who can submit and the controlling attributes of submission process for each type of submission. After the platform receives data submitted in various permissible formats, it will perform object-level validation before handling over to respective Content Processor(s) for data shredding, validation and storing into the Submission Database. The platform monitors the status of submitted data and its processing statuses on real time basis. The REs or filers not only can check the status of the submitted data online, but they can also access to the exception and error reports from the platform as acknowledgement or for rectification and resubmission.

Figure 1 – Submission Obligation Framework



The platform provides a Portal which is equipped with the following self-service functions for various type of users to operate:

No	Component	Functionalities	Remark
1.	Submission Dashboard	Personalised landing page of the Submission Portal which provides the summary of submission statuses, pending/overdue submission windows and tasks assigned to the user. Serves as the main page for accessinf to the published submission forms, submission guidelines, announcements, alerts and messages.	Example of submission status information: o Count of submission window overdue o Count of submission window pending o Count of submission window pending for approvals o Count of successful submitted submission windows in current quarter o Count of submissions windows that have been rejected
2.	Administration and Configuration	A group of functions associated with administering REs and Submission Obligations.	Example of submission administration functions: o Defines Subject Area Profile (grouping of entities) and Submission Template o Links Subject Area Profile to Submission Template o Define attributes of submission obligation such as submission cycle (frequency), submitter type information, submitter profile (RE profile or Subject Area Profile) and other submission process control attributes (for e.g. deadline, reminder trigger, requiring approval and/or digital signature, etc) o Define Content Processor for each type of submission data/form.
3.	Security and Access Management	A group of functions associated with administering internal and external user accounts and access privileges. Provides user registration, authentication and authorisation (user access control) services.	All external or RE users will be self-registered and managed by the respective Security Administrator of the same RE. The platform can be customised to integrate with internal security systems to enable Single-Sign-On (SSO).

No	Component	Functionalities	Remark
4.	Submission Operation	A group of functions to facilitate RE users to perform the data submission based on the submission/ resubmission windows generated and assigned to them. Embedded workflow to support submission approval and digital signing. Once submission data/file are submitted, the *Rules Engine* processes the file by sending them to respective Content Processors The close-loop integration with Content Processor(s) enable the status of validation (by Content Processor) and acknowledgement/ validation/ exception/ error reports to be accessible via the Submission Portal. Upon submission, the validation process will be executed by the platform and the respective Content Processor. The status and results of the validation will be made available for further action.	Submission Windows are a means to inform the RE of a report that is ready or due to submit. Support the following modes of submission: o Web forms o File Upload o Straight-Through or Business-to-Business (B2B) Interface in eXtensible Markup Language (XML) or eXtensible Business Reporting Language (XBRL) format o Bulk Data Submission (Transactional) Creation of submission window is automated based on the submission obligations using the scheduled process which creates the submission window automatically on the scheduled date provided. The platform can be customised to integrate with existing workflow system and email services to streamline the business processes.
5.	Submission Monitoring	Provide the status of submissions performed and statistics by submission windows based on user specified criteria (different version for internal and external user)	

No	Component	Functionalities	Remark
		Access to the submission histories and activities log for each form submitted by user.	
6.	Integration and Interfaces	For inclusiveness and heightening of business processes, the platform can integrate with the respective internal systems (subject to availability).	Example of internal systems can be integrated: o Master Data Management System (Reference Registry & Entity Database) o Workflow or Business Process Management System o Security systems (ActiveDirectory/LDAP/ePKI/Digital Signature) o Email Server/Services
7.	Back-end and non-functional functions	Embedded audit trails features and functions to support the data submission functionalities and system management purposes.	Embedded with logging of the user actions and system activities (for e.g. login/logout, change user details, perform submission, etc), integration status, exception/error. Perform data archiving by transferring data from active database schema to archive database schema base on Retention Period. Maintain email template, generating and sending of submission alert and notification based on submission obligation setting (e.g. allow submission notification, late submission, etc)

Table 1 – Submission Platform Functions

Some of these solutions are developed based on proprietary specifications and needs, while there are also solutions integrating from end-to-end components available to provide a credible, cutting-edge and holistic solution. The platform can be customise and leverage on/ integrate with the existing internal databases and data management system (for e.g. sourcing of standard or reference data managed in Reference Registry), security and business process application (for e.g. for Single-Sign-On and workflow).

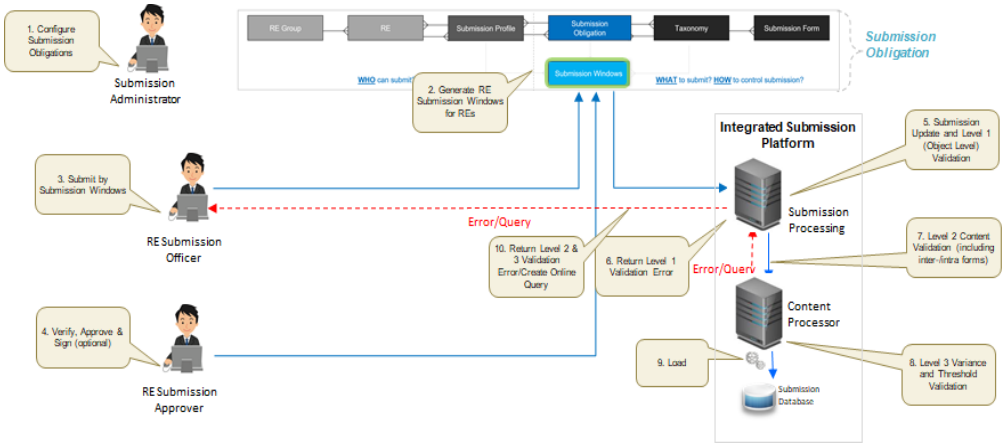


Figure 2 – Configurable Submission Process

The platform adopt role-based access for users and support Delegated User Account Management where the appointed Security Administrator of RE will be responsible to manage (approve, reject and assigning to respective user group with pre-assigned access privileges) the users registered under the same RE organisation. Up to the Security Administrator, each user can be assigned to multiple roles with different access privileges which can be summarised as follow:

System Owner (Regulator/Statistical Body)

User Role	How do they use the system?
Submission Manager	To administer submission obligation – RE Group, RE Profile information, submission template/forms, submission obligation attributes and assigning Support Officer to respective RE or submission forms.
Security Administrator	To administer related user roles for internal and external (RE) users associated with permissible functions.
System Administrator	To manage system parameters and perform system housekeeping.

Data Submitter/Reporting Entity/Filer

User Role	How do they use the system?
Security Administrator	To approve, assign or remove submission related user roles to RE, such as RE Submitter, RE Approver and RE Officer-In-Charge (OIC).
Submitter	To perform form(s) submission for the RE.
Approver	To approve or reject the submitter form(s) by RE Submitter.

Submitter	Perform submission on behalf of RE(s).
Support and Processing Officer	To monitor the submission progress of REs, and to approve or reject access requests to specific data. To manage and monitor the Operational Data Management and Maintenance functions.

Table 2 – Stakeholders and Roles

The platform support Non-repudiation of submitted data. This is made possible with embedded workflow (if activated) to allow the submitted data to be approved and digitally signed after submission.

3. Results

The platform is a suite of ready-to-deploy applications providing dedicated functionalities to meet most if not all of the data submission/collection and processing requirements. It is highly-customisable and integrate-able and can well fit into the existing application infrastructure for standardisation and streamline of business process.

The platform's Functional Model is a thinking model used to facilitate and coordinate the data submission solutions around people, process and technology. It is a is a secured web-based application that provide the unified platform to facilitate the submission of data from Reporting Entities (REs) and external data providers (or data sharing partners). Its configurable nature provides the flexibility to support various data submission needs of data collectors and can be implemented within short duration.

In general, it will help the data collectors to:

- Reduced REs' reporting overhead via enabling submission through a single reporting platform;
- Reduced REs' reporting overhead via streamlined reporting via a standardised and unified submission/ reporting framework;
- Enhanced data management processes with an integrated solution for end-to-end data management processes from data capturing, transformation and dissemination;
- Enhanced data management processes by trigger and alert capabilities for immediate actions by both RE and data collector; and
- Ease of maintenance and support as only one system need to be maintained.

Handshakes corporate intelligence solution

Sachvinder Singh
Handshakes Technology (M) Sdn Bhd

Abstract

Handshakes is an award winning financial technology ("FinTech") and regulation technology ("RegTech") company whom specializes in Corporate Intelligence and Artificial Intelligence solutions. We were Incorporated in 2011 and are a regional SME with presence in Singapore, Malaysia, Thailand, Vietnam Hong Kong and China.

There's an abundance of company information around, but most are either unstructured or unavailable in the public domain. At Handshakes, we help solve this challenge and deliver this data in a format that is easy for interpretation and accurate decision making.

At the heart of our collective intelligence capability is the technology that is the backbone of our solutions. With our proprietary Data Analytics and Deep Learning Artificial Intelligence, our technology not only enhances our data, but can also be seamlessly integrated with our client's processes and database.

Handshakes flagship product **"Handshakes"** is an interactive platform for delivering corporate intelligence about people and companies. Handshakes is powered by big data and proprietary information researched from official data sources such ACRA (Singapore), SSM(Malaysia), SAIC (China) company registry information, and stock exchange (Bursa, SGX, HKEX) disclosure documents.

The Handshakes platform is capable of disambiguating duplicate entities and fusing multiple sources of data, to allow multiple layers of discovery and analysis. The Handshakes platform can handle data in multiple languages, including character-based data like Chinese data. The Handshakes platform is also powered by "**SEER",** a natural language processing artificial intelligence engine that processes text and automatically extracts data about people and companies. The extracted data can be seamlessly delivered to the Handshakes platform to enrich the main database. (For example, SEER can automatically process a news article that mentions companies and people and add that data into the main database. SEER extraction is accurate and reliable enough for regulators to rely on, our Capital Markets dataset which regulators use daily processed by SEER.

Keywords

Data; Technology; HST; Corporate Intelligence

1. Introduction

Handshakes is an award winning financial technology ("FinTech") and regulation technology ("RegTech") company whom specializes in Corporate Intelligence and Artificial Intelligence solutions. We were Incorporated in 2011 and are a regional SME with presence in Singapore, Malaysia, Thailand, Vietnam Hong Kong and China.

There's an abundance of company information around, but most are either unstructured or unavailable in the public domain. At Handshakes, we help solve this challenge and deliver this data in a format that is easy for interpretation and accurate decision making.

At the heart of our collective intelligence capability is the technology that is the backbone of our solutions. With our proprietary Data Analytics and Deep Learning Artificial Intelligence, our technology not only enhances our data, but can also be seamlessly integrated with our client's processes and database.

We connect our clients to official registries that include Singapore ACRA to access and retrieve all the latest and historical data they require. This information include shareholders identity and directors information to incorporation data and entities related to the business- all presented in interactive data maps, allowing teams to make sense of the data faster.

With Handshakes, organisations today are conducting superior research about companies private and public to make better informed decisions. From research departments seeking highly accurate data information to audit teams desiring fast and speedy data analysis, our solutions are being used by companies in a multitude of industries to solve their everyday challenges. Financial Institutions use us to make better informed investment decisions, regulators utilise our concierge service for custom market research that ranges from capital markets to retail and journalism.

Handshakes solution is an acclaimed company developing prize winning solutions which includes;

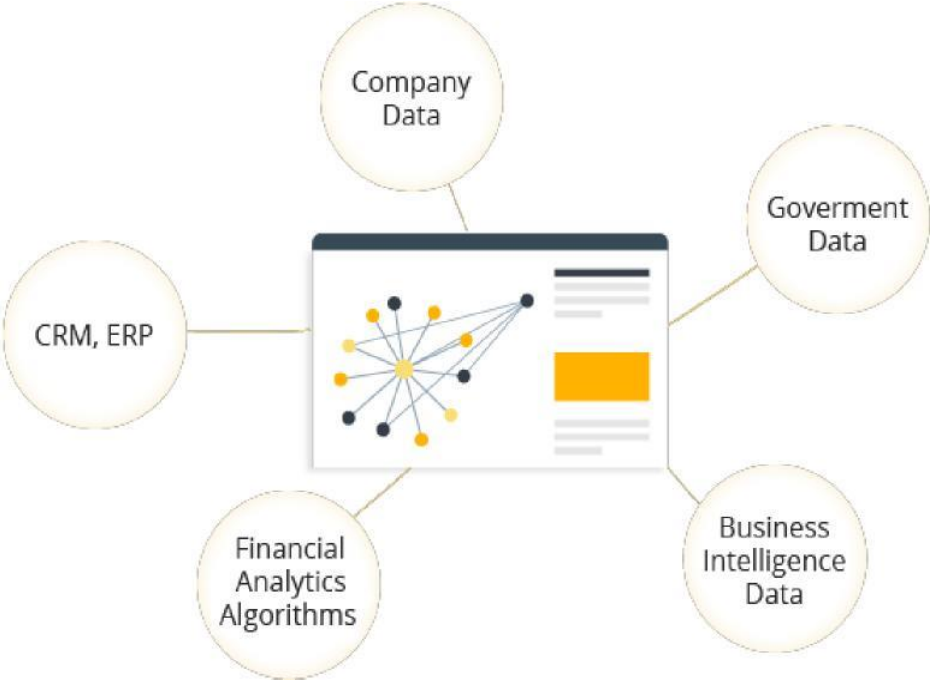
- National Infocomm Awards 2016
- Singapore InfoComm Technology Federation Awards 2017.
- "Most Innovative Business Idea" at the India-ASEAN Summit in Delhi.
- IDC's Innovators 2018 (Asia Pacific Data As A Service Provider)

2. Methodology

Handshakes Solutions

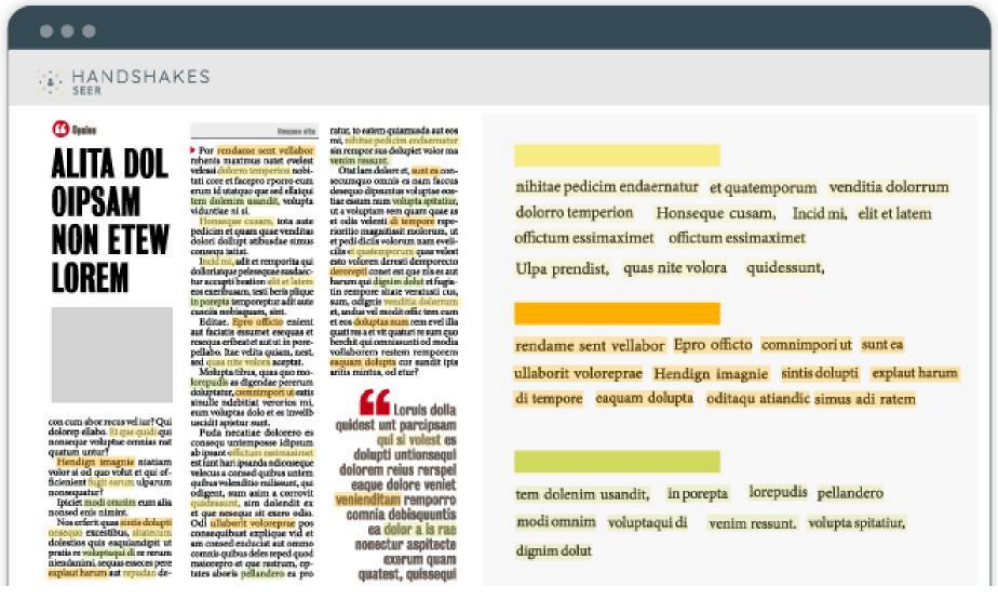
Handshakes App, API and FUSE

- Handshakes flagship product "**Handshakes**" is an interactive platform for delivering corporate intelligence about people and companies. Handshakes is powered by big data and proprietary information researched from official data sources such ACRA (Singapore) SSM (Malaysia), SAIC (China) company registry information, and stock exchange (Bursa, SGX, HKEX) disclosure documents. The Handshakes platform is capable of disambiguating duplicate entities and fusing multiple sources of data, to allow multiple layers of discovery and analysis. The Handshakes platform can handle data in multiple languages, including character-based data like Chinese data.
- Via Handshakes FUSE the clients can now also bring the power of the Handshakes technology platform and data straight into their organization. The will be able to customize the Handshakes platform bu fusing their data, business processes and analytics algorithms directly into a secured private premise copy of Handshakes. This will assist clients whom love our App but feel they can improve it further as well as those whom want to broaden their data horizons by combining multiple public and private data sets.



Handshakes SEER

- Context and linguistics analysis, entity and relationship extraction, tabular and paramedic data harvesting- all in one self – evolving, machine learning package. The Handshakes SEER can be used to monitor global websites and news; process volumes of customer feedback; extract knowledge from emails, categorise terabytes of documents and so much more. The SER technology can be applied o any large collection of text too unwieldy (or impractical) for any individual to process- saving many days effort and rendering previously impossible tasks, possible.
- The Handshakes platform is also powered by "SEER", a natural language processing artificial intelligence engine that processes text and automatically extracts data about people and companies. The extracted data can be seamlessly delivered to the Handshakes platform to enrich the main database. (For example, SEER can automatically process a news article that mentions companies and people and add that data into the main database. SEER extraction is accurate and reliable enough for regulators to rely on, our Capital Markets dataset which regulators use daily processed by SEER.



- Technology that Handshakes Possess includes;
  - ✓ Automated Entity Resolution:
  - ✓ Matching entities and names automatically, reliably.
  - ✓ Network Analytics Algorithms: E.g. Interconnection, Beneficial Ownership...etc
  - ✓ Named Entity Recognition
  - ✓ Semantic Classification

- ✓ Semantic Clustering

- Valuable Data that Handshakes Possess includes;
  - ✓ Capital Markets (Listed Companies): Singapore, Malaysia, Hong Kong.
  - ✓ Company Registry (Unlisted Companies): Singapore, Malaysia, China, Indonesia, Vietnam.
  - ✓ Regulatory Action across 80 sources worldwide.

3. Results

- The result? The ability to save valuable time poring over mountains of data, discover relevant content in seconds and guarantee consistent performance eliminating any human -error.
- The Handshakes technology is today widely used by Regulators (e.g. central banks, market regulators, tax authorities) ,enforcement Agencies (e.g. police, white collar crime, anti-corruption),Financial Institutions, Investment Banking Professionals (e.g. issue managers, placement agents, sponsors, corporate governance advisors) ,Lawyers (e.g. onboarding teams, collection teams, corporate litigation) ,Researchers (e.g. Universities, Trade Associations, Professional Associations) ,Private Investigators ,Journalists ,Auditors for the following purposes;
  - ✓ Uses Background Checks: Check on directors, shareholders, beneficial owners and adverse track records.
  - ✓ Investigations: Unified cross-border data, interconnecting entities, syndicate detection, identify indirect controllers, semantic searching.
  - ✓ Surveillance: Push alerts on directorship or shareholding changes, real-time news surveillance and alerts.
  - ✓ Procurement Checks: Detect conflicts of interest, collusion.
  - ✓ Automation of internal processes.
- The results of the Handshakes solutions have proven effective and were even showcased in high profile news such as the following;
  - ✓ Linked Firms Eyeing Same Contract
    (The Business Times ,12th September 2017)
    Data by Handshakes - a portal that generates interactive network maps of people and entities - shows four cases of certain tenderers for the same jobs being connected to one another by common directors - past or present - and/or shareholders plus other links such as a common company secretary and in one case, registered address. https://www.businesstimes.com.sg/government-economy/linked-firms-vying-for-same-public-contracts
  - ✓ A peek inside SixCap, a firm on MAS Investor Alert List

(The Straits Times, 5th September 2017)
Handshakes used to identify linked firms. https://www.straitstimes.com/business/a-peek-inside-sixcap-a-firm-on-mas-investor-alert-list

- ✓ Many HK penny investors caught in cross – holdings web
(The Business Times, 10th July 2017)
Research from corporate intelligence portal Handshakes, an artificial intelligence platform run by Singapore startup DC Frontiers, now indicates that key individuals from the Enigma network are linked to a wider web of Hong Kong small caps. https://www.businesstimes.com.sg/companies-markets/many-hk-penny-investors-caught-in-cross-holdings-web
- ✓ Man Behind Recent Investmenr Woes.
(The Business Times, April 25th 2016)
THOUSANDS of diamond and wine investors in Singapore and Malaysia who have been left high and dry by investment companies might have more in common than first thought. https://www.businesstimes.com.sg/companies-markets/man-behind-recent-investment-woes
- ✓ Handshakes App Used to Uncover Relations Behind Hai Di Lao's Recent IPO.
Handshakes was used to uncover the relationships between any shareholders in Hai Di Lao and the suppliers the company deals with. https://www.handshakes.com.sg/handshakes-App-Uncovering-The-Relations-Behind-Hai-Di-Lao-s-Recent-IPO.html

4. Discussion and Conclusion

- ➢ The business landscape is becoming increasingly interconnected and information surrounding it is growing at an exponential pace. While this complexity opens up ends possibilities, how do we converge and make sense of this vast amount of data? Every Industry has their own unique challenges and systems which are manpower intensive. This often leads to inaccurate data and analysis that could lead to flawed conclusions. From time to productivity loss to wrong decision making, we saw the need for a combination of quality official data coupled with adaptive technologies. Corporate intelligence has to evolve and at Handshakes we are the catalyst.

Exploratory study of key traits for the fourth industrial revolution among employees of financial institutions in Malaysia

Haniza Yon[1], Mazlina Muhammad[3], Devika Balan[4], Kok Mun Yee[2], Nur Ayu Johar[2], Amrit Lakra, Nurul Fatin Shakira Helmi[3]

[1]Global Psytech (Malaysia)
[2]Global Psytech Sdn. Bhd.
[3]Bank Negara Malaysia
[4]Bank Islam Malaysia Berhad

Abstract

The fourth industrial revolution (4IR) is changing the way people live and work at an unprecedented speed. For success during this revolution, employees require future-focused skills such as problem-solving, creativity, and critical thinking. In the financial services industry, one effect of the revolution is the rise of the financial technology (FinTech) sector. The accompanying technological restructuring, which includes a shift towards a peer-to-peer financial system, is expected to affect business models, profitability, productivity, and employment needs in the financial industry. To survive the FinTech wave, financial institutions will need to foster innovation and attract entrepreneurial talents by adopting new methods of talent acquisition and management. This paper reports the results of our measurement of behavioural competencies among 211 employees in the Malaysian financial services industry, with a focus on attributes that are important in a 4IR work environment in which FinTech is taking centre stage. We also report measures of the psychometric and statistical properties of the instrument we used, as well as of its construct validity.

Keywords

psychometrics; talent analytics; validity

1. Introduction

Financial Technology (FinTech) is the use of new technologies in the financial services industry. Innovations in FinTech, such as mobile banking, cryptocurrency, and robo-advisers, are leading to a restructuring in this industry. This restructuring, which includes a shift towards a peer-to-peer financial system, is changing business models and affecting profitability, productivity, and employment needs.

The Fourth Industrial Revolution (4IR) has opened up new opportunities and challenges for many industries. Sectors such as healthcare, travel, and agriculture have greatly advanced using 4IR technology, but in the financial services sector, this revolution remains at an earlier stage (World Economic Forum, March 2019). Perhaps the sophistication and complexity of the

technology already used in the financial services sector, which make it difficult for most people to understand, have impeded the adoption of even more advanced 4IR technology. Despite the industry's slowness to adapt, change inevitably continues; for instance, the number of mobile banking users has been increased year by year. Financial institutions will need to continue to foster innovation and to attract entrepreneurial talents in order to give users greater access to the financial system.

Researchers have found personality factors to be important predictors of workplace performance, turnover, and citizenship work behaviour (Barrick & Mount, 1991; Borman & Motowidlo, 1997; Boudreau et al., 2001; Campbell, 1990; Campbell & Knapp, 2001). We embarked on this research to study the patterns of behaviour among employees in the financial services industry in Malaysia, with a focus on attributes that are important in a 4IR work environment in which FinTech is taking centre stage. Specifically, this behavioural skills study, which includes understanding and validating noncognitive factors relevant to the 4IR work environment, represents one of the important contributions we are making to both the financial services industry and human resource literature.

Table 1: Summary of Psychometric Properties

Factor	Range Item Location	Range Item Outfit	Person Reliability	Gender (p<.01) No. showing DIF	Gender Main Effect (t(134))
Flexibility	-1.15 - 1.32	0.83 - 1.42	0.80	-	-2.09*
Customer-Service Orientation	-1.25 - 1.35	0.82-1.34	0.79	2	1.95
Creativity	-1.26 - 1.00	0.82-1.46	0.81	-	-0.38
Empathy	0.46 - 0.63	0.79-1.58	0.77	-	0.33
Prob.-Solv. & Resourcefulness	-1.57 - 0.92	0.81-1.70	0.79	1	-1.61
Initiative	-0.87-1.24	0.84-1.72	0.79	-	0.61
Effectiveness	1.02-1.37	0.86-1.31	0.77	-	0.32
Entrepreneurship	-1.01-1.29	0.71- 1.40	0.76	-	-0.83
Emotional Intelligence	-1.4-1.36	0.92-1.06	0.83	-	-0.19
Resilience	-0.58-0.60	0.86-1.10	0.84	-	-0.91
Execution	-0.95-0.84	0.85-1.22	0.79	-	1.21
Prob.-Solv. & Decision-Making	-0.88-0.46	0.96-1.10	0.75	1	0.17
Self-Confidence	-0.97-1.26	0.84-1.19	0.75	-	2.20*
Productivity	-0.79-1.11	0.93-1.31	0.79	-	1.60
Innovation	-1.16-1.06	0.88-1.09	0.78	-	-1.89

* p < .05

The five-factor personality model developed by workers such as Goldberg (1990) has emerged as the dominant framework in personality research. It describes personalities along five dimensions: openness to experience,

conscientiousness, extraversion, agreeableness, and neuroticism. However, there are researchers who have identified more comprehensive frameworks of narrow personality traits such as those of NEO-PI-R by Costa, McCrae & Dye (1991), and the International Personality Item Pool (IPIP) by Goldberg et al. (2006). Drasgow et al. (2012), on the other hand, identified 21 narrow personality facets that can be categorised under the hierarchical structure of the five-factor personality model. As in the aforementioned research, our approach identified a large number of factors (15), and these are shown as the row headers in Table 1.

Despite evidence that the five-factor personality model can predict workplace success, there has been little research addressing the relevance of personality measures in a 4IR work environment. The Future of Jobs report published by the World Economic Forum (January 2016) discussed the skills that 4IR would demand of future workers. The top ten skills identified for workers in 2020 were complex problem solving, critical thinking, creativity, people management, coordinating with others, emotional intelligence, judgment and decision making, service orientation, negotiation, and cognitive flexibility (World Economic Forum, January 2016). Creativity was projected to be one of the top three skills by 2020. Active listening was projected to drop out of the top ten and to be replaced by emotional intelligence.

Some researchers have explored the use of forced-choice items to combat faking on personality tests (Gordon, 1951; Ghiselli, 1954). Forced-choice personality tests normally present participants with pairs of descriptive statements, each statement in the pair representing a different desirable personality trait. Participants are to select the statement that describes them better. This method produces ipsative test scores – that is, it compares the strength of different constructs within each individual rather than comparing them across individuals (Hicks, 1970). Advances in modern psychometrics, however, have made it possible to obtain normative test scores using forced-choice measures (Stark, Chernyshenko & Drasgow, 2005; Stark & Drasgow, 2002).

This paper aims to investigate patterns of behaviour among employees in the Malaysian financial services industry, with a focus on attributes that are important in a 4IR work environment in which FinTech is taking centre stage. First, we describe our sample of test takers, our forced-choice instrument, and our methods of data analysis. Next, we present measures of the psychometric and statistical properties of our instrument, as well as of its construct validity. Finally, we give an overview of the implications of the results. Throughout, we take gender as the main independent variable to be studied.

2. Methodology

Respondents. We administered our instrument as an online survey. A link was sent to a sample of the target population (employees in the Malaysian

financial services industry) by bank officers through e-mail. The participants were given a deadline by which to complete the instrument. The sample, which was randomly selected, was composed of 211 employees (84 males and 127 females) in the Malaysian financial services industry. The respondents' experience in the industry ranged from a few months to more than 26 years. All participants were required to accept a data protection agreement before participating in the survey.

Questions. The instrument called Work 4.0 which we developed, included questions about demographic variables such as gender, ethnicity, place of work, and working experience. In addition, a total of 36 questions were included to measure behavioural competencies in creativity, innovation, entrepreneurship, productivity, problem-solving, self-confidence, empathy, emotional intelligence, and resilience. In the present context "faking" is likely to occur when using Likert-style items; this problem cannot be resolved using consistency scales. Therefore, we used a forced-choice (FC) approach to combat faking (see, e.g., Brown & Bartram, 2009; Bartram & Burke, 2013, Hontangas et al., 2015, Bartram & Tippins, 2017). FC items are best suited to high-stakes situations with high demand characteristics, such as in personnel selection, when applying for bank credit, or when being forced to reveal other high stakes information – including value measurement (see, e.g., Brown & Bartram, 2009; Hontangas et al., 2015).

IRT. We analysed our data with Winsteps (Linacre, 2019) software. We used the one-parameter logistic model to measure psychometric and statistical properties of our instrument, including its model fit and gender-related biases, and to assess its construct validity. The one-parameter logistic model is a kind of item response theory model. According to the Columbia University Mailman School of Public Health:

> The item response theory (IRT), also known as the latent response theory refers to a family of mathematical models that attempt to explain the relationship between latent traits (unobservable characteristic or attribute) and their manifestations (i.e. observed outcomes, responses or performance). They establish a link between the properties of items on an instrument, individuals responding to these items and the underlying trait being measured. IRT assumes that the latent construct (e.g. stress, knowledge, attitudes) and items of a measure are organized in an unobservable continuum. (Item Response Theory, *n.d.*)

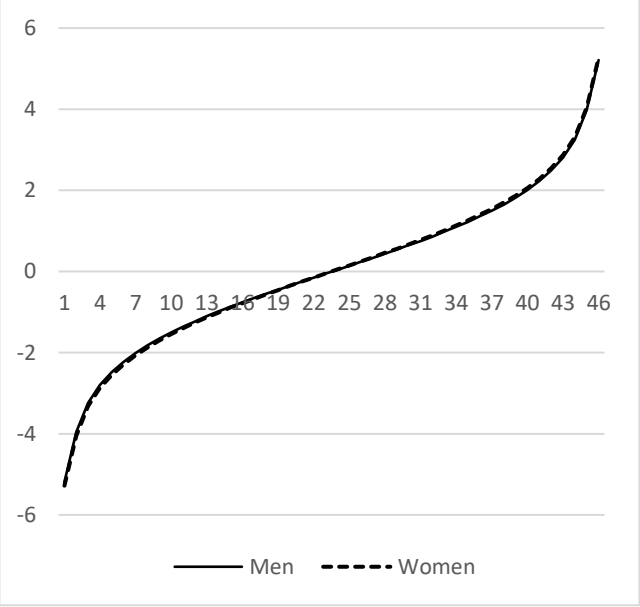
For additional information about IRT, we refer the reader to van der Linden and Hambleton (1995).

Gender Bias. IRT can be used to screen for bias for or against particular sub-groups of respondents. Bias can occur at the item level and at the test level; at the item level, it is called Differential Item Functioning (DIF). We tested for bias by gender at both levels.

3. Results

Item Fit and Reliability. The results of scaling analyses of 15 factors are summarised in Table 1. Acceptable item *MS-Outfit* statistics (i.e., *Outfit* $\leq$ 1.40, see Linacre, 2019) were obtained for all but five of the factors. For each of the remaining five, the misfit was caused by a single misfitting item, and removing this item did not affect the overall person measures (r > 0.96). Table 1 further shows that each of the factors has acceptable reliability (all values $\geq$ 0.75).

Figure 1: Raw-to-Logit Transformation for Women and Men: Customer-Service Orientation



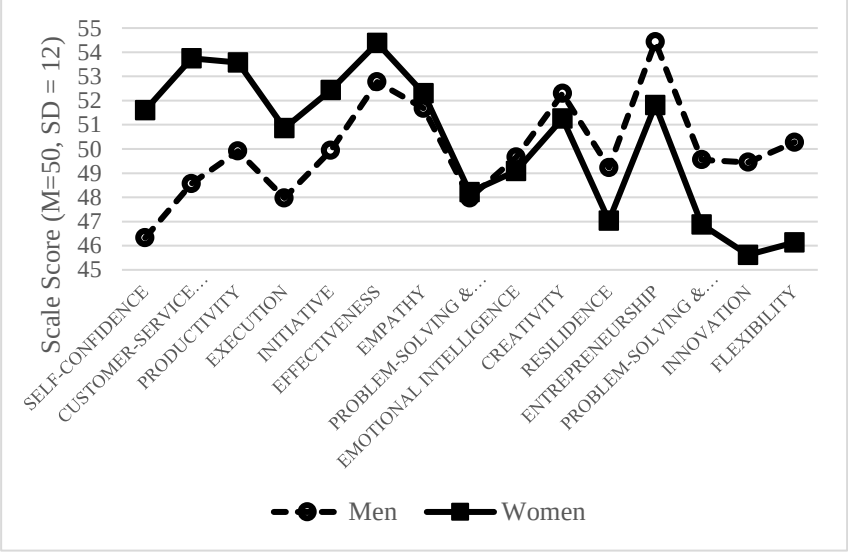
Gender Bias. Table 1 shows that three factors contained at least one item with statistically significant DIF by gender (i.e., the item parameters differ between men and women). However, given the large number of pairwise comparisons made at $p < 0.01$, this number is to be expected by chance alone (Binomial Distribution, p > .20).

Most importantly, the gender-related DIF had little effect at the test level, in the estimation of respondents' traits. When the raw-to-logit transformations for men and women were computed separately, the differences were negligible. For instance, "Customer-Service Orientation" had the largest number of items showing DIF (two). Yet, as is illustrated in Figure 1, the test-level distortion is negligible; the raw-to-logit curves for men and women essentially coincide. Very similar graphs were obtained for "Problem-Solving & Resourcefulness" and "Problem Solving & Decision-Making", the other two factors for which items showed statistically significant DIF. Accordingly, we conclude that the factor estimates were essentially unbiased by gender.

Gender Main Effects. Gender DIF can occur regardless of the presence or absence of any gender main effects on the estimated factor measures. In our data, the traits for which some items showed DIF were not those with the largest differences in average scores between men and women. Figure 2 below

shows the factor means for women (solid lines) and men (dotted lines), with the factors arranged along the X-axis. The factors are ordered according to the difference (M_{Women}–M_{Men}) between the female and male means. Rather surprisingly, women reported greater self-confidence and productivity but less flexibility and innovation. These results should be interpreted cautiously, as statistical tests for pairwise group differences (see Table 1) indicate that only two of the differences reached statistical significance at $p < .05$.

Figure 2: Factor Means by Gender



4. Discussion and Conclusions

All our 15 factors showed acceptable model fit and reliability, indicating that we succeeded in constructing a set of internally valid scales in accordance with the one-parameter IRT model. A very small number of misfitting items was detected, and further research is being conducted to understand the issues involved.

However, no substantive gender biases were found, and we conclude therefore that any gender differences in average observed trait scores reflect genuine trait-level differences, rather than (say) item- or test-related artifacts.

We observed some surprising gender differences: the women in our sample showed greater self-confidence, customer service orientation, productivity, execution, and initiative than did the men. Conversely, men showed greater flexibility, innovation, problem-solving and resourcefulness, and entrepreneurship. These findings are not what one might expect in a male-dominated business culture. Admittedly, these differences were not very strong, and most of them fell short of statistical significance. Even to the extent that the differences are real, they may be limited to Malaysia. Nonetheless, the observed gender differences are interesting and present an opportunity for further study.

Our instrument appears to yield reliable results when used on employees in the financial services industry. Further validation of the instrument is ongoing, and we aim to report the outcomes of this research in future publications.

References

1. Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26.
2. Bartram, D. & Burke, E. (2013). Industrial/Organisational testing case studies. In J.A. Wollack & J.J. Fremer (Eds.), *Handbook of Test Security* (pp. 313-332). New York: Routledge.
3. Bartram, D. & Tippins, N. (2017). The Potential of Online Selection. In H.W. Goldstein, E.D. Pulakos, J.Passmore & C. Semedo (Eds.), *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection & Employee Retention* (pp.271-292). New Jersey: Wiley-Blackwell.
4. Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99–109.
5. Boudreau, J. W., Boswell, W. R., Judge, T. A., & Bretz, R. D. (2001). Personality and cognitive ability as predictors of job search among employed managers. *Personnel Psychology, 54,* 25–50.
6. Brown, A., & Bartram, D. (2009). *Doing Less but Getting More: Improving Forced-Choice Measures with IRT*. Paper presented at the 24th Annual conference of the Society for Industrial and Organizational Psychology, New Orleans.
7. Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. M. Hough (Eds.), *Handbook of 52 industrial and organizational psychology* (2nd ed., Vol. 1, pp. 687–731). Palo Alto, CA: Consulting Psychologists Press.
8. Campbell, J. P., & Knapp, D. J. (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Erlbaum.
9. Costa, P. T., Jr., McCrae, R. R., & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences, 12,* 887–898.
10. Drasgow, F., Stark, S, Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification decisions* (Technical Report No. 1311). Fort Belvoir, VA: Army Research Institute for the Behavioral and Social Sciences.
11. Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *Journal of Applied Psychology, 35,* 407–412. doi:10.1037/h0058853

12. Ghiselli, E. E. (1954). The forced-choice technique in self-description. *Personnel Psychology, 7,* 201–208.
13. Goldberg, L. R. (1990). An alternative "description of personality:" The Big Five factor structure. *Journal of Personality & Social Psychology, 59,* 1216–1229.
14. Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public domain personality measures. *Journal of Research in Personality, 40,* 84–96.
15. Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74,* 167–184.
16. Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced choice tests. *Applied Psychological Measurement*. Advance online publication, doi:10.1177/0146621615585851
17. Item Response Theory (*n.d.*). Retrieved on April 16, 2019, from https://www.mailman.columbia.edu/research/population-health-methods/item-response-theory.
18. Linacre, J. M. (2018). *Winsteps® Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com.
19. Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29,* 184–203.
20. Stark, S., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26,* 208–227.
21. Van der Linden, W.J., & Hambleton, R.K. (1995). *Handbook of Modern Item Response Theory*. New York: Springer.
22. World Economic Forum (Jan, 2016). *The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution.* Retrieved from: http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf
23. World Economic Forum (Jan, 2016). *The 10 skills you need to thrive in the Fourth Industrial Revolution.* Retrieved from: https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/
24. World Economic Forum (March, 2019). *It's time for financial services to embrace the Fourth Industrial Revolution.* Here's why. Retrieved from: https://www.weforum.org/agenda/2019/03/its-time-for-financial-services-to-embrace-the-fourth-industrial-revolution-heres-why/

Using covert response biases in psychometric assessments to bolster job candidate interviews: an example with hospitality roles

James Houran[1], Bruce Tracey[2], and Rense Lange[3]

Abstract

Psychometric testing and *structured behavioural interviews* are two best-practice approaches to pre-employment screening and selection. However, an ongoing challenge has been to maximize their effectiveness by empirically and procedurally aligning the two tactics to work together as a cohesive process. New generation applications of Modern Test Theory offer one viable solution to this problem. Using the example of the proprietary 20|20 Skills™ assessment that was designed specifically for service-hospitality industries, this paper shows how psychometric assessments can be designed to yield Human Resources data that transcend mere raw-scores to provide "hidden" information about job candidates' likely areas of strengths or weaknesses related to Execution, People, and Cognitive Skills. This information follows from covert response biases (i.e., IRT residuals) that candidates exhibit to assessment items, which subsequently can be utilized to frame and guide structured behavioural interviews. Candidates are unaware of such statistical outcomes in assessment reports, thereby providing an extra level of test security that job-seekers cannot readily anticipate or "game." This improved approach is innovative in that it essentially tailors structured behavioural interviews to individual job candidates, while maintaining consistency and legal-defensibility in its general framework and process.

Keywords

profiling; psychometric testing; rasch scaling; employee selection; behavioural interviewing

1. Introduction

Although there is no fail-proof method for evaluating applicants or incumbents in recruitment or promotion contexts, Human Resources (HR) professionals have long recommended that organizations implement a triangulated system of checks-and-balances to gather and evaluate candidate information related to technical competencies, role fit, and compatibility with

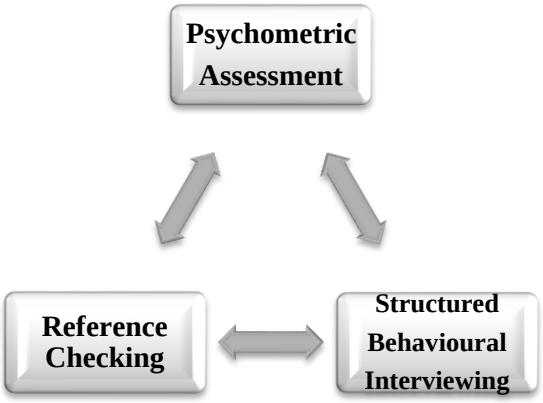
[1] AETHOS Consulting Group, Dallas, Texas and corresponding author: jameshouran@eathos.com)

[2] School of Hotel Administration, SC Johnson College of Business, Cornell University, Ithaca, New York

[3] Global Psytech, Selangor, Malaysia, and Lab. for Statistics and Computation, ISLA, Vila Nova de Gaia.

a company culture (see Fig 1 below). Arguably, the only demonstrably objective part of this triangulated process is the use of psychometric testing or standardized assessment. Despite their popularity, many such tools have been criticized on theoretical, empirical, or legal grounds. An obvious and pervasive confound is that outcomes on many traditional psychometric assessments can be intentionally and positively skewed by respondents via cheating or impression management. Moreover, many tools have questionable internal validity, legal-defensibility, and practical value given their construction and validation with Classical Test Theory as opposed to Modern Test Theory (Lange, 2017; Lange & Houran, 2015). Accordingly, new assessments must be developed that transcend raw-scores based on test content so that job-seekers cannot easily "game" the test.

Figure 1: Best-Practice Triangulated Process of Candidate Due Diligence in Screening and Selection.



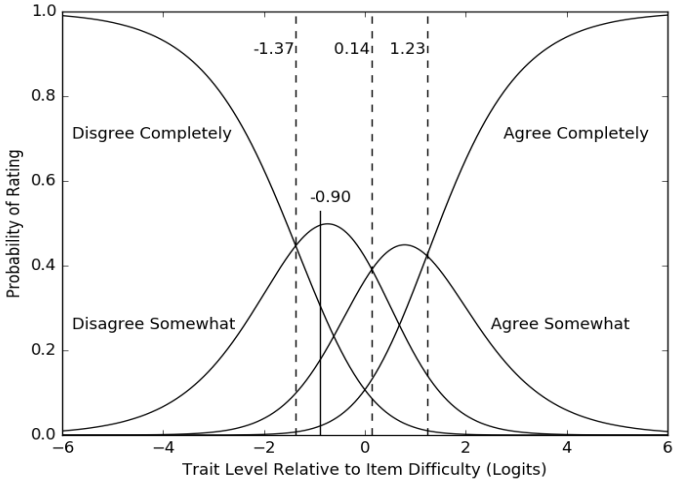
Decades of research have demonstrated that employment interviews alone have limited validity in predicting job performance (Hunter & Hunter, 1984). However, researchers have noted the strong tendency for interviewers to make decisions based on superficial observations. For example, one simulation found interviewers rated applicants more highly if they showed greater amounts of eye contact, head movement and smiling, as well as other non-verbal behaviour. Such physical clues accounted for eighty percent of the variance in candidate ratings (Tessler & Sushelsky, 1978). More recent analyses suggest that the effectiveness for both structured and unstructured interviews may be better than is traditionally assumed (Huffcut & Arthur, 1994). Also, most applicants believe the interview is an essential component of the selection process (Rynes & Gerhart, 1990), perhaps many organizations use at least one interview in their selection process (Howell & Dipboye, 1982). Research suggests that candidates view the employment interview as the most suitable measure of their relevant abilities (Schuler, 1993). Smither and colleagues (1993) also found that applicants perceived interviews as more job

related than other procedures. Similarly, Rynes and Connerley (1993) found that the interview was perceived to possess high job-relatedness. Schuler (1993) suggested that selection methods which are perceived as controllable by the candidate, obvious in purpose, providing task relevant information, and offering a means of feedback are considered the most socially-valid or acceptable. We propose that the validity and effectiveness of the combined approach in Figure 1 can be bolstered by empirically and procedurally aligning assessments and interviews to work collectively as a seamless and integrated process. Using insights gained from psychometrics and data-mining, we outline one solution that uses empirical insights gleaned from test-takers' responses. This solution has a long (> 15 years) track record of actual usage in the area of Human Resources (HR) testing. The approach relies heavily on the statistical machinery provided by Item Response Theory (IRT), and Rasch scaling in particular (see van der Linden & Hambleton, 1997).

2. An IRT Approach

The Rasch (1960/1980) latent-trait model uses the simplest IRT where items are described solely by their difficulty, at least for binary items. In the case of rating scales, the structure of the rating scale, as described by transition from one rating to another, has to be taken into account as well. The Rasch rating scale model is probabilistic and revolves around the log odds of P_{ijk}, i.e., the probability that item i will receive from person j the rating k. In particular,

Figure 2: Category structure of hypothetical item (see text).



$$\log\left(\frac{P_{ijk}}{Pij(k-1)}\right) = T_j - D_i - \{F_{ik}\} \quad (1)$$

In Equation 1,

- The left-hand side represents the ratio of P_{ijk} over $P_{ik(k-1)}$, i.e., the log odds of observing a rating in category k relative to probability of observing one in category k-1.
- T_j represents the trait level (or ability) of person j, and most applications focus on the T_j only.
- D_i is the "difficulty" of the item, or the magnitude of the trait level needed to elicit the response k. Higher values of D indicate that higher trait levels are needed to obtain a higher rating.
- In general, higher D yields lower ratings and higher T will yield higher ratings.
- As the trait level T_j increases we expect to see the categories 0, 1, 2, …., m to be used in this order – at least probabilistically. The F_{ik} denote the locations at which the ratings k and k-1 are equally likely, and such F_{ik} are typically referred to as "step-values."
- Although the following assumes that items share the same step-values, the subscript i indicates that the step-values are actually allowed to vary across items. We use the convention that $\sum_k F_{ik} = 0$ and it is mathematically convenient to define $F_{i0} = 0$.
- Note that all parameters vary along the same latent trait variable and they are expressed in the same unit of measurement, i.e., the *logit* as defined by the left-hand side of Equation 1.

3. Example

The various definitions are illustrated in Figure 2 which shows the Pijk of each rating k for a rating scale with four categories *("Disagree Completely," "Disagree Somewhat," "Agree Somewhat," and "Agree Completely")* across the underlying latent Rasch dimension. In the example, $D_i = 0$ and F_k *= {0, -1.37, 0.14, 1.23}*, and these step-values are shown relative to Di. It can be seen that for very low trait levels the rating *"Disagree Completely"* is virtually certain, but that "Disagree Somewhat" becomes equally likely at -1.37 logits and the latter is more likely above this point. Similarly, *"Disagree Somewhat"* and *"Agree Somewhat"* are equally likely at 0.14 logits, and *"Agree Somewhat"* and *"Agree Completely"* are equally likely at 1.23 logits. Note that the P_{ijk} sum to unity across the various categories. For instance, at *-0.9 logits* (solid vertical line) the probabilities of the 4 categories (in order) are about *0.31, 0.45, 0.21*, and *0.03*, respectively.

4. Background

Obtaining images such as Figure 2 requires solving for P_{ijk} in Eq. 1:

$$P_{ijk} = \frac{\exp \sum_{t=0}^{k} [T_j - D_i - F_t]}{\sum_{k=0}^{m} \exp \sum_{t=0}^{k} [T_j - D_i - F_t]} \qquad k = 0, 1, \dots, m \tag{2}$$

It follows from Eq. 2 that raw sums of observations (ratings coded as 0, 1, 2, …) are minimally sufficient statistics to estimate respondents' trait levels T_j as well as the item parameters D_i (Wright & Masters, 1982). Various approaches to estimating the model parameters are described in this work as well. To avoid introducing additional notation, the following makes no distinction between the estimated and the true parameter values. Given Eq. 2 it is possible to derive the expected value (E_{ij}) of a person's rating x_{ij}, and its' standard deviation SD_{ij} (see Wright and Masters, 1982). Thus, we can define an observation's residual as:

$$y_{ij} = x_{ij} - E_{ij} \tag{3}$$

and its' standardized form z_{ij} as:

$$z_{ij} = \frac{y_{ij}}{SD_{ij}} \tag{4}$$

5. Misfit

It may be assumed that the z_{ij} follow an approximately normal distribution with M=0 and SD=1. Thus, the summed squared values z_{ij}^2 then follow a χ^2 distribution and a person index of fit can be obtained by aggregating the z^2 over items answered by this person (see Wright & Masters, 1982). Similarly, aggregating across persons will provide an index of item fit. When items' z are aggregated across different subgroups of respondents, they also serve to identify Differential Item Functioning (DIF), also called item bias. For instance, if an item's z_{ij} has higher mean for men than for women then this item is biased against women. Thus, the z_{ij} can be interpreted as indices of idiosyncratic preferences and subjective biases.

6. Using Misfit

Traditionally, testing in HR, education, and psychology focuses exclusively on estimating the T_j, the respondent's overall trait level and a test taker's items' fit is typically ignored. Yet, an observation with large $|z_{ij}|$ should be deemed aberrant because it is implausible given the model parameters. It is the central thesis of this paper that such aberrations are worthy of study in their own right. Large z_{ij} (e.g., $|z_{ij}| > 2$) can be caused by a variety of factors: the question may be ambiguous (e.g., due to poor wording of the question), or the test-taker was distracted (e.g., ambient noise) or the person lacked motivation. However, when such factors can be excluded as the causes of misfit, the z_{ij} often reflect a respondent's idiosyncrasies and biases. As is illustrated below, misfit conveys valuable information that goes beyond a person's overall trait estimate (test scores) and can be exploited for diagnostic purposes.

7. A Case Study: The 20|20 Skills™

The following describes how response residuals can be used to suggest to

interviewer useful follow-up questions during behavioural interviews. We use the example of the 20|20 Skills assessment, which contains over 125 questions that address eleven HR related factors, including: Leadership, Ethical Awareness, Service Orientation, Creativity, Problem Solving, Self-Efficacy, Awareness of Diversity, Team-Building, Group Process, Company Loyalty, and Humour Appreciation (Houran & Lange, 2007; Lange & Houran, 2009). Each factor was assessed via ten questions using a 4-point rating scale. The assessment also contains a "Lie Scale" to gauge potential impression management. The design and development of the 20|20 Skills™ assessment is based on collaborative effort between the School of Hotel Administration at Cornell University and two private companies (Integrated Knowledge Systems, Inc. and AETHOS Consulting Group, Inc.). Across the eleven factors, the average reliability was 0.84 (*Median* = 0.82, *SD* = 0.04), indicating high reliability. Inspection of items' group-specific residuals indicated that item misfit (if any) cannot be attributed to age-, gender-, job-level-, or ethnicity-related biases.

Figure 3: 20|20 Skills Item-level Feedback map.

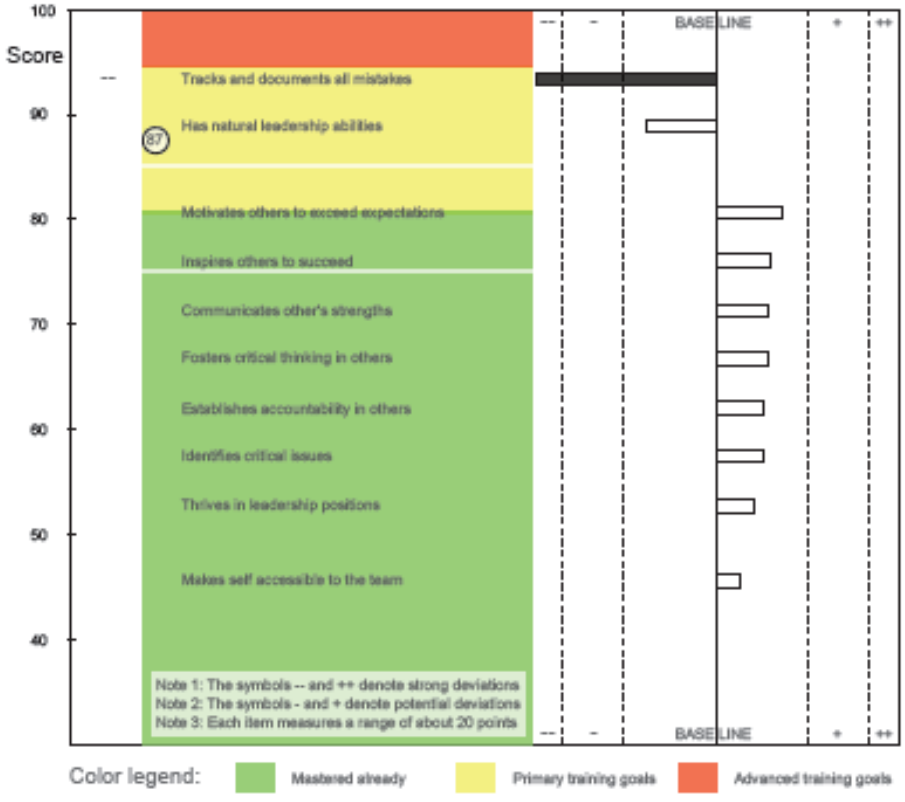


Figure 3 shows an example where a hypothetical test-taker obtained an IRT scaled score of 87 - which is a linear transform of T_j, (in *Logits*) - on the Leadership subscale. Condensed versions of the 10 Leadership items are shown on the left sorted in order of their difficulty D_i ("easy" items occur near

the bottom, "hard" items occur near the top), together with the scale score (inside circle). The green area indicates the areas that are already mastered by this respondent, the red area marks the topics yet to be mastered (none here), while the yellow area marks the primary training goals. Most importantly, the right-hand side of Figure 3 shows the model deviations of this respondent's test answers in terms of their z_{ij} (see, Eq. 6 above). Note that the answers to the easiest questions (bottom part of figure) do not reach any of the vertical dotted lines, where the first line away from 0 (solid black line) represents $z_{ij} = 1.65$ and the second $z_{ij} = 2.0$. Thus, while the answer residuals are not exactly 0, these deviations can be seen as random. However, this is not the case for *"Tracks and documents all mistakes"* which is highly negative ($z < 2$). In other words, while most of the responses of this test taker are only to be expected given his/her T_j, this item stands out significantly $(p < .05)$. From a practical perspective this means that this issue may need further attention. The identification of items with particular low or high z_{ij} is exploited to formulate or guide the most pertinent and useful questions to pose the test-taker (i.e., job candidate or incumbent) during subsequent job-interviews. For instance, in the situation depicted in Figure 3 it is suggested that the interviewer might ask use the follow-up question *"Do you think it is important that complete records should be kept of employees' mistakes?"* Note that for each question this approach presumes the availability of a library of follow-up questions corresponding to low vs. high residuals.

8. Discussion

The 20|20Skills™ system outlined above has been in commercial use for over 15 years. Over 100,000 applicants or incumbents in the service-hospitality industry have been tested globally across a wide range of roles (consumer non-facing and facing) up-and-down the organizational chart – and it has been translated into several languages other than English. Conceptually, this paper demonstrated a specific and proven method of combining (i) tailor-made individual feedback obtained via *psychometrics* with (ii) *structured* (or *situational*)-*based interviews* by human interviewers to create an integrated process of pre-employment screening and selection. The 20|20Skills™ served as proof-of-concept for the development of a customized "line-level" assessment system that combines artificial intelligence and machine-learning with the response residual approach. The new system is optimized for a particular company culture to quickly and accurately pinpoint test-takers with attitudes and behaviours characteristic of markedly-high levels of *service orientation* and *contextual performance* (Houran, Tracey, & Lange, 2017). In combination with the use of "big-data" in HR (daily attendance information, service latencies, customer evaluations, … etc.), we anticipate that residuals also can play an important role in performance evaluations. In addition, we expect that response residuals proves useful equally in engagement surveys,

customer satisfaction surveys, and employee 360-degree performance reviews. Clearly, this opens a new area for research and practical applications.

References

1. Houran, J., & Lange, R. (2007). State-of-the-art measurement in Human Resource assessment. *International Journal of Tourism and Hospitality Systems*, *1*, 78-92.
2. Houran, J., Tracey, J. B., & Lange, R. (2017). The hospitality 'X Factor' in non-management employees, psychometrically speaking**. Paper presented at the Cornell Hospitality Research Summit (CHRS), Oct 5-7, Ithaca, New York.**
3. Howell, W. C., & Dipboye, R. L. (1982). *Essentials of industrial and organizational psychology*. Homewood, IL: Dorsey Press.
4. Huffcutt, A.I., & Arthur, W. Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79,* 184-190.
5. Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72-98.
6. Lange, R. (2017). Rasch scaling and cumulative theory-building in consciousness research. *Psychology of Consciousness: Theory, Research and Practice*, *4*, 135-160.
7. Lange, R., & Houran, J. (2009). Perceived importance of employees' traits in the service industry. *Psychological Reports*, *104*, 567-578.
8. Lange, R., & Houran, J. (2015). "Quality of measurement" – the implicit legal cornerstone of HR assessments. *Employee Relations Law Journal*, *40*, 46-60.
9. Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: U. of Chicago Press.
10. Rynes, S. L., & Connerley, M. L. (1993). Applicant reactions to alternative selection procedures. *Journal of Business and Psychology*, *7*, 261-277.
11. Rynes S. L., & Gerhart, B. (1990). Interviewer assessment of applicant 'fit:' An exploratory investigation. *Personnel Psychology*, *43*, 13-22.
12. Schuler, H. (1993). Social validity of selection situations: A concept and some empirical results. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: individual and organizational perspectives* (pp. 11-26). Hillsdale, NJ: Erlbaum.
13. Tessler, R., & Sushelsky, L. (1978). Effects of eye contact and social status on the perception of a job applicant in an employment interviewing situation. *Journal of Vocational Behavior*, *13*, 338-347.
14. van der Linden, W. J., & Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer.

15. Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Developing an integrated survey for admin-based population estimates and labour market statistics

Pete Jones
Office for National Statistics, Fareham, UK

Abstract

Surveys are an essential component in the production of official statistics in the United Kingdom (UK). While the Office for National Statistics (ONS) is transforming population, migration and social statistics with increased use of administrative data, an Integrated Survey Framework (ISF) is needed to support the transition towards alternative sources of data in official statistics. One of the major ambitions for ONS transformation is to deliver an Administrative Data Census for England and Wales after the next Census in 2021. This will offer more frequent statistics for small area geographies, and the potential for new outputs that have not been collected in previous censuses. Having published research outputs highlighting the coverage patterns in administrative data population estimates, ONS is developing a Population Coverage Survey (PCS) to measure and adjust for coverage errors in administrative records. With an anticipated annual sample size of 500,000 households per annum, the scale of the proposed survey corresponds with sample size expectations for the future transformation of ONS social surveys. Given the scale of the proposed survey transformation at ONS, one of the central aims of the Integrated Survey Framework is to merge the PCS and LMS into a single integrated household survey. While there are significant overlaps in the data collection requirements underpinning these two surveys, establishing a sample design that can produce robust population estimates and labour market statistics remains a challenge. The current proposal is based on the concept of a 'master wave', where households are sampled using a two-stage stratified design. The master wave will be central to the future Integrated Survey Framework, provided a sampling frame for other ONS social surveys. This paper will describe the principles behind the ONS Integrated Survey Framework, research undertaken so far on the sample design for the LMS/PCS survey, and test plans to support its implementation between now and 2022.

Keywords

Population; administrative data; labour market; survey integration

1. Introduction

The Office for National Statistics (ONS) is working towards the transformation of the population, migration and social statistics system in England and Wales. The Data Collection Transformation Programme (DCTP)

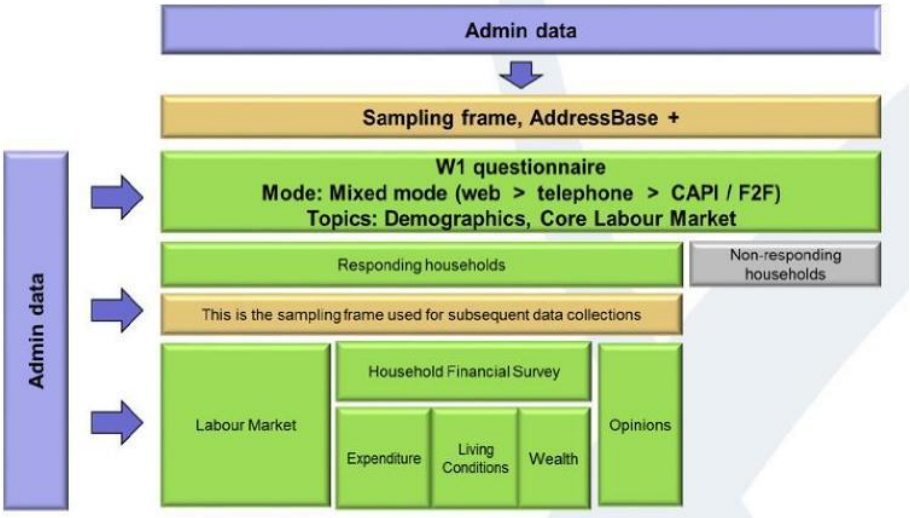
seeks to rebalance ONS's data collection activity significantly toward wider, more integrated use of administrative and other non-survey data sources, thereby reducing our reliance on large population and business surveys. While this does not eliminate a need for surveys, it does mean ONS's traditional approach to surveys will now differ. The ONS Integrated Survey Framework (ISF) has been set up to deliver modernisation of social surveys with three underlying principles to support transformation; administrative data first, digital by default, statistical redesign and rationalisation.

Central to the ISF is a major redesign of the ONS Labour Force Survey (LFS). The LFS is the largest social survey conducted in the UK, collecting longitudinal information across five waves from approximately 160,000 households annually. Topics collected from the LFS have continued to expand over recent decades. Presently there are nearly 600 questions across the different waves and sub-modules, covering topics that extend beyond traditional labour market content. As censuses are undertaken every ten years in England and Wales, the LFS provides a crucial source of information for updating statistics for key census topics in the period between censuses. This has been supported with strong user demand to increase samples sizes in local areas, leading to the supplementary Annual Population Survey (APS) boost to support the LFS collection. Using combined data from the LFS and APS, annual statistics are produced for a range of topics including labour market, health, ethnicity, households and families.

Consistent with the general trend towards declining response to social surveys (de Leeuw and de Heer, 2002[1], and De Leeuw, Hox and Luiten, 2018[2]), there has been a notable reduction in LFS response rates in recent decades. The LFS collection currently uses a combination of face to face and telephone interviewing, and between the ten-year period 2008 to 2018 response rates have dropped from 58.2% to 40.3%. Similar to experience amongst other National Statistics Institutes, the roll out of electronic questionnaires designed for self-completion is now key to the future development of social survey and census collections. The US Census Bureau document the challenges of moving questionnaires online[3], however successful transition to online will reduce collection costs and support strategies for reversing the recent trend towards lower response rates.

Beyond the LFS, the ONS also conducts a range of other social surveys covering different topics and questions, as well as using different sample designs and collection modes. These are also the target of transformation with examples including; the Household Financial Surveys (HFS), which comprise three surveys covering expenditure, living conditions, and wealth and assets, and the Opinions and Lifestyle Survey (OPN). Figure 1 below shows the concepts behind the proposed ISF to transform these collections.

Figure 1: Concepts underpinning the ONS Integrated Survey Framework (ISF)



The use of administrative data is an important feature of the proposed framework in two ways. Firstly, the increasing availability of alternative data sources collected by government has the potential to replace survey questions and reduce additional burden on respondents. Secondly, information from administrative records can be used to enhance the address frame used for social survey sampling. ONS are still in the process of assessing the quality of the Ordnance Survey AddressBase product[4], however the integration of household and person characteristics from administrative data will enable more flexibility in the future design of social surveys.

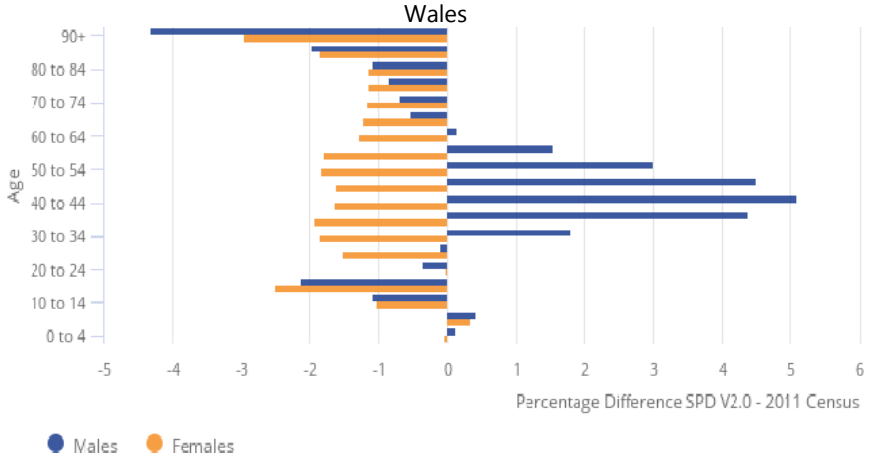
Central to the ISF is the concept of redesigning and upscaling the wave 1 questionnaire that is used in the current LFS model. To maximise response, the wide set of topics currently included in wave 1 of the LFS has been reduced to cover only core labour market content. This redesigned survey, which is mid-way through a five-year testing plan is known as the ONS Labour Market Survey (LMS). Using a mixed-mode approach, selected households will be invited to take part in an online survey for the first two weeks of collection, before using face to face and telephone interview follow up. The overall sample size for the LMS wave 1 – referred to as the 'master wave', will be increased on the basis that ONS will be collecting this data to support wider collection requirements under the ISF.

The responding households in wave 1 that indicate willingness to take part in further surveys will be used as a sampling frame for other social surveys. These sub-modules will include longitudinal collection of labour market data following wave 1 and also capture any residual social survey requirements. Survey modules will be streamlined accordingly to reduce respondent burden and make use of administrative data where possible.

Another major aim of the ISF is to incorporate requirements for new survey collections to support Population, Migration and Social Statistics

transformation at ONS. Since 2011, ONS has been working towards the UK government ambition to deliver future censuses using administrative data. An increasing number of countries are now aiming to transition from traditional census collections in favour of reusing existing data collected by government. For countries that have a population register, the existing pre-conditions are in place to produce reliable population estimates. For the UK and other countries that do not have a population register, there are considerable challenges in compiling administrative data to have accurate coverage of the resident population. The ONS have developed methods to link multiple administrative sources, including health, tax and benefits, and education to construct a Statistical Population Dataset (SPD)[5]. Figure 2 below shows the coverage patterns of the SPD population estimates relative to the 2011 Census.

Figure 2: Statistical Population Dataset (SPD) estimates compared to 2011 Census, England & Wales



The comparison in figure 2 shows that the SPD is characterised by under-coverage (typically a consequence of non-registration on administrative data), but also over-coverage (for example, individuals that are registered on administrative data but have since left the country). To measure and adjust for these coverage biases, an ongoing Population Coverage Survey (PCS) is proposed for collection on a continuous basis. The PCS is based on similar concepts as the Census Coverage Survey (CCS)[6] which has been used as a post-enumeration survey to adjust for non-response bias in the 2001 and 2011 Censuses. An important difference regarding the CCS is that it has been designed as an independent survey that is collected during the weeks immediately following a Census collection. The aim of the PCS is to integrate with wave 1 of the new Labour Market Survey, which will necessitate continuous collection and redesign of the sampling approach.

2. Methodology

ONS have started testing a 'master wave' LMS/PCS survey that has the dual purpose of assessing coverage of administrative data and collecting labour market data. The proposal to merge labour market questions with population coverage is desirable based on similarities underpinning the collections, including sample size requirements for social surveys and population estimation, the use of mixed mode collection and the inclusion of questions that enumerate all household members.

While merging the two surveys will greatly reduce the operational costs of running two large surveys, certain aspects of the integrated design need to be developed and tested. To measure overall feasibility, ONS undertook an operational test for three versions of an online survey in Autumn 2017. Tranche 1 of the test was designed to measure take up rate to an online version of the new LMS, using different incentives. No additional follow up modes were pursued with non-responding households that were invited to take part online. Tranche 2 of the test was designed as a standalone PCS covering questions typically used in previous CCS questionnaires. Non-responding households to the online survey were followed up with face to face interviewing and the option to take part in a telephone or postal survey. Tranche 3 of the test was designed as an integrated household survey, merging questions from both the PCS and the LMS. Insights from previous cognitive testing was used to harmonise questions between the two surveys, and a similar mixed mode approach was adopted as that used in tranche 2. The results of the 2017 test are summarised in section 3.

Between November 2018 and April 2019, data has been collected to support a statistical test comparing LMS and current LFS outputs. This data has been collected using a similar version of the integrated survey described in tranche 3 of the 2017 test, and results will be available later in 2019. In May 2019, a first attempt at using administrative data to support the integrated sample design is being tested in the field, with the aim of increasing the number of contacts with migrant households. The results of this test will be used to determine whether a migrant boost should be incorporated into the LMS/PCS survey to support transformation of migration statistics.

A collaboration between ONS and University of Southampton has commenced to develop an integrated sample design. The current LFS is based on equal probability sampling, whereas it's assumed that a coverage survey to support administrative data population estimates will require a stratified design to adjust for varying levels of coverage error across geographic areas and age-sex groups. Using derived variables from the SPD we have developed a first iteration of a stratification index to predict over-coverage and under-coverage propensities for small areas. To identify the optimum number of strata required to measure under-coverage and over-coverage we have used the Delanius-Hodges method of minimum variance stratification[7]. We are in

the process of calculating the design effects that a stratified sample will have on labour market statistics, with a view to determine what level of clustering and overall sample sizes will best support PCS and LMS outputs.

3. Results

Figure 3 shows that for tranche 1 of the 2017 test, the online LMS survey achieved an online response rate of 22.5%. This is broadly in line with the findings of other UK government surveys that have switched to online collection. An LMS incentivisation strategy, which was also tested in tranche 1 and has since been used for subsequent tests achieved a response rate of 27.7%. Online response rates to tranches 2 and 3 achieved higher non-incentivised online take-up, possibly the result of interviewers encouraging households to respond online as an alternative to the face to face interview requested in follow up. Additional follow up modes including face to face, telephone and postal option achieved an overall response rate of 67.6% for the mixed mode PCS survey. This was higher than the PCS/LMS integrated version (56.8%). This may partly be due to the shorter amount of time it took respondents to complete the PCS only version. However, it is also a consequence of the PCS sample being highly clustered enabling interviewers to make more repeat visits than the integrated PCS/LMS field work.

Figure 3: Results from the 2017 test tranches

	Tranche 1: LMS online survey*	Tranche 2: PCS survey	Tranche 3: Integrated PCS/ LMS survey
Overall mixed mode response rate	-	67.6%	56.8%
Online	22.5%	25.6%	25.4%
Face to Face	-	39.3%	27.9%
Telephone	-	1.0%	0.6%
Postal	-	1.8%	2.9%
Average household completion time	-	9m 46s	16m 43s
Online	18m 12s	11m 54s	21m 46s
Face to Face	-	7m 50s	13m 42s
Household partial complete rate (online)	7.0%	12.0%	9.6%
Individual partial complete rate (online)	8.0%	11.6%	9.2%

*for comparability across tranches we present results from the non-incentivised LMS survey tested in tranche 1

The results of the 2017 test indicated that online self-completion could support the strategy for increasing response rates to social surveys, and that an integrated LMS/PCS was worth pursuing for more detailed statistical testing.

A key milestone for ONS social survey and population statistics transformation is to deliver a large scale LMS/PCS test in 2020. The aim of this test is to produce population estimates and labour market statistics for comparison with official estimates at national level. Early simulations have identified a potential approach based on a two-stage stratified clustered design. Under this design, small areas (ONS output areas)[8] are selected from strata at random as the primary sampling unit. Within the output areas selected, a fixed number of addresses will be randomly selected to take part in the survey. The stratification is based on assigning each output area into 1 of 6 strata for over-coverage and under-coverage using the minimum variance stratification method. The design variable for over-coverage is based on measuring numbers of people moving into and out of addresses in the area. For under-coverage we assessed a number of predictors, including migration within the area, ethnicity, and other evidence of administrative data activity not captured on the SPD. Using principle components analysis, we have developed a composite design variable to stratify for under-coverage derived from these variables. An initial assessment of model fit gives an R2 value of 0.44 for the over-coverage measure and 0.48 for under-coverage. These are broadly comparable with the model used to predict Census non-response as part of the 2011 hard-to-count index (0.47)[9]. Sample sizes for the 2020 test will be calculated once the impact of a stratified design is fully understood for the precision of labour market statistics.

4. Discussion and Conclusion

Initial operational tests for an integrated LMS / PCS survey have demonstrated potential for combining survey questions on labour market and population coverage. Further research is needed before concluding that the proposed design will produce robust statistical outputs. From the PCS perspective, an understanding of how response rates to a voluntary survey will impact the quality coverage assessment is needed. While the CCS, which is used to support the traditional Census estimation framework is also a voluntary survey, it achieves high response rates (90%+) on the basis that respondents associate the collection with the mandatory Census that is conducted a few weeks prior to the CCS. The proposed integrated LMS/PCS survey will undoubtedly achieve lower response rates, posing an increased risk of bias in the resulting estimates. Understanding the relationship between survey non-response and registration on administrative data is of important consideration when evaluating the impact of lower response rates, and further work will be undertaken in this area. In addition, a method for adjusting over-coverage within the dual-system-estimation framework is also needed in preparation for the 2020 test.

From an LMS perspective, ONS are currently in the process of assessing the impact of a stratified design on labour market statistics. Compared to the

equal probability sampling used for the current LFS, the proposed design to have unequal sampling fractions across PCS strata will result in unequal design weights for labour market estimates. In addition, the proposal to select a fixed number of cases within the areas selected in each strata will further reduce the effective size of the LMS sample. We are currently in the process of running simulations to estimate the design effects that the proposed sample design will have on LMS outputs. Depending on the results we will consider ways to improve the sample design if it is unable to deliver the current level of precision obtained from the LFS collection.

A considerable challenge will be in developing the operational capacity at ONS to carry out the necessary field work to ensure response to the survey is maximised. Running a continuous survey of this scale is unprecedented up until now, and the current aim is to be at full scale by 2022.

References

1. de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, A. D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse* (pp. 41-54). New York, NY: John Wiley & Sons
2. de Leeuw E., Hox J. & Luiten A. (2018), International Nonresponse Trends across Countries and Years: An analysis of 36 years of Labour Force Survey data. Survey Insights: Methods from the Field. Retrieved from https://surveyinsights.org/?p=10452
3. Census Bureau: New Technologies in Census Data Collection. Part 2: Developing an Electronic Questionnaire, 2016
4. AddressBase is a database of address points in Great Britain, maintained by Ordnance Survey
5. Methodology of Statistical Population Dataset V2.0, Office for National Statistics, October 2015
6. The 2011 Census and Coverage Adjustment Process, Office for National Statistics, July 2012
7. Tore Dalenius and Joseph L. Hodges, Jr, Minimum Variance Stratification, Journal of the American Statistical Association, Vol. 54, No. 285 (March 1959), pp. 88-101
8. ONS output areas are the smallest level of geography for which estimates of population size are produced in England and Wales (on average 125 households per OA)
9. Predicting patterns of household non-response in the 2011census http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/statistical-methodology/predicting-patterns-of-household-non-response-in-the-2011-census.pdf

Recent progress on implementing a Bayesian approach to population estimation from an administrative list subject to under and over-coverage

Patrick Graham, Anna Lin
Statistics New Zealand, Christchurch, New Zealand

Abstract

Several statistical agencies are exploring replacing or enhancing traditional census-based population estimation systems with administrative data. Administrative data is prone to both under and over-coverage. Directly estimating genuine list over-coverage due to erroneously enumerated individuals no longer in the target population is challenging, because it is often difficult to obtain definitive evidence of absence. We have been investigating a Bayesian method for estimating both under and over-coverage of an administrative list, which is based on a model for the joint distribution of inclusion in the target population and the list. The model is fitted to the union of a sample survey of the target population and the list. Estimation of list over-coverage from the sample-list union is possible, given good information on sample inclusion probabilities. In this paper we review the basic ideas of our estimation methodology and report on recent progress with implementation and evaluation of the model.

Keywords

population estimation; administrative data; bayesian inference; missing data

1. Introduction

Statistical agencies in several countries are investigating methods for replacing traditional census-based population estimation system with approaches based on administrative data (see, for example, Bycroft, (2015)). Administrative lists may fail to include some people who are in fact in the target population and also include people who are no longer in the target population, due, for example, to undetected out-migration. Relative to a traditional census, the latter problem (over-coverage) may be a more significant issue for population estimation based on administrative data. By population estimation we mean, not just the total size of the population, but also the distribution of population across categories of key demographic variables such as age, sex, ethnic group and area. We assume that it is possible to conduct a highly quality survey of the target population and that this sample can be linked to the list without error. We assume no other fieldwork. In particular, the methodology outlined does not require any sampling from the list. Thus, our approach makes use of the important insight of Zhang (2015) that estimation of list over-coverage is possible without sampling directly from

the list. However, our problem differs from that discussed by Zhang (2015), because we assume a single list, supplemented by a survey, whereas Zhang (2015) assumed a data structure comprising two (or more) lists and a sample survey of the target population (which could be replaced by a third list known only to suffer from undercoverage). Our focus is on small domain population estimation and production of a corrected unit record file and we take a Bayesian approach to inference. In contrast, Zhang (2015) concentrated on frequentist estimation of total population size. A detailed account of our methodology can be found in Graham and Lin (2019). Here we provide a brief account of the main ideas and discuss some details of implementation, particularly with respect to the sample survey of the target population. As in Graham and Lin (2019) we ignore issues of measurement error or misclassification of list variables and linkage error.

Table 1: Cross tabulation of target population estimation and an administrative list

		List		
		1	**0**	
Target	1	n_{11}	n_{10}	N_T
	0	n_{01}	0	
		N_L		

Table 2: Underlying cell-probabilities for population-list union at some setting x of covariates

		List	
		1	**0**
Target	1	$\phi_{11}(x)$	$\phi_{10}(x)$
	0	$\phi_{01}(x)$	0

2. Basic set-up. To establish basic concepts, suppose a target population (e.g. usually resident population of New Zealand) could be cross-tabulated with an administrative list that is thought to overlap the target population. The resulting table would have the structure shown in Table 1. Note that Table 1 does *not* represent the data structure for a dual systems (DSE) population estimation problem. It is a conceptual representation of the relationship of the target population (which is not directly observed) and an administrative list that overlaps the target population.

The only directly observable quantity in Table 1 is the total number of people on the list, N_L. An unknown number $n_{01,}$ of individuals on the list are not in the target population. These n_{01} people constitute "over-coverage" of the list with respect to the target population. If we had an indicator for inclusion or otherwise in the target population it would be straightforward to exclude people not in the target population from population estimation. However, we assume no such indicator and therefore identifying the n_{01} people included

on the list but not in the target population is a missing data problem. The missing data absent from the list are the indicators for inclusion in the target population. If we could determine the over-coverage, $n_{01,}$ then since the list total, N_L, is directly observed, we could immediately obtain the number of people both in the target population and on the list as $n_{11}= N_L - n_{01}$. On the other hand, an unknown n_{10} individuals are in the target population but not on the list. This group represents the "under-coverage" of the list with respect to the target population. If we could estimate n_{11}, then given an estimate, $\hat{n}_{10}$ of n_{10} we could obtain an estimate of the target population total N_T as

$$\hat{N}_T = \hat{n}_{11} + \hat{n}_{10} = \hat{N}_L - \hat{n}_{01} + \hat{n}_{10}.$$

Ideally, we would like to estimate not just the total population size N_T but the number of people in the target population by characteristics such as age, sex, ethnic group and area. Therefore, we assume a structure such as Table 1 for each combination of these variables. We let **X** denote the covariates of interest and **X** = **x** a particular combination of these variables.

Allowing for dependence on the covariates, Table 2, describes a probability model underpinning the cross-tabulation of the target population and the list. The probabilities for the three occupied cells in Table 2 sum to one. Under this model, an individual in the target population-list union, with covariates x is allocated to one of the three possible cells with the probabilities given in Table 2. Thus, at the unit level, we posit a multinomial model,with one trial. Given the cell probabilities from Table 2 we can define the under-coverage probability, Pr(not on list |in Target, **X** = **x**) as $\phi^{under}(\boldsymbol{x})$ = $\phi_{10}(\mathbf{x})$/($\phi_{11}(\mathbf{x})$ + $\phi_{10}(\mathbf{x})$) and the over coverage probability for the list, Pr(not in Target|on list, **X** = **x**) as $\phi^{over}(\boldsymbol{x})$ = ϕ_{01} (**x**)/(ϕ_{11} (**x**) + $\phi_{01}(\mathbf{x})$). Since ϕ_{11} (**x**) + $\phi_{10}(\mathbf{x})$ + $\phi_{01}(\mathbf{x})$ = 1 we need specify only two of the cell probabilities to fully specify the multinomial model implied by Table 2. A convenient approach is to model $\phi^{under}(\boldsymbol{x})$ and $\phi_{01}(\mathbf{x})$. The remaining cell probabilities can then be obtained as ϕ_{11} (**x**) = (1 – $\phi_{01}(\mathbf{x})$)(1 – $\phi^{under}(\boldsymbol{x})$), $\phi_{10}(\mathbf{x})$ = (1 – $\phi_{01}(\mathbf{x})$) $\phi^{under}(\boldsymbol{x})$.

Table 3: Cell-probabilities for the sample-list union at setting x of the covariates

		List	
		1	**0**
Sample	1	$\lambda(x)\phi_{11}(x)$	$\lambda(x)\phi_{10}(x)$
	0	$(1-\lambda(x))\phi_{11}(x)+\phi_{01}(x)$	$(1-\lambda(x))\phi_{10}(x)$

Notice that the number of people in the (0,0) cell in Table 1, corresponding to "not in the target population and not on the list" is assumed to be 0. In fact, most of the world's population falls in the cell! However, we are not interested in estimating the population of the world but of some specific target population such as the usually resident population of New Zealand, and we are seeking to use an administrative list for this purpose. For this problem, only people in the target population or on the list or in both are relevant. That is our conceptual starting point for estimation is the union of the target population and the list (cf Zhang (2015)). We let N_U denote the size of the target-list union.

If a sample has been drawn from the target population with sample inclusion probabilities, $\lambda(\mathbf{x})$, independently of list inclusion, and the sample is linked to the list without error, cross-tabulation of sample and list produces a 2 x 2 table (at each setting of $\mathbf{X}$) underpinned by the probabilities shown in Table 3. For simplicity, we regard the $\lambda(\mathbf{x})$ as known. In practice $\lambda(\mathbf{x})$ may need to be estimated. From Table 3 it can be seen the sampling process transfers some people from the (1, 1) cell in the target-list union to the (0, 1) cell in the sample-list union, and some people from the (1, 0) cell in the target-list union to the (0,0) cell in the sample-list joint distribution. This cell is, in reality not observable. This needs to be accommodated in the analysis. An important point is that Table 3 does not represent a traditional capture-recapture, or dual-systems population estimation problem. Whereas the latter involves two or more samplings from a target population we have a single sampling from the population which is linked to a list that overlaps the target population. The observed (0, 1) cell comprises a mix of people from the target population that were not included in the sample and people genuinely not in the target population. Traditional DSE methods cannot accommodate the latter group.

3. Inference. We base inference on the posterior predictive distribution of a corrected list from which individuals not in the target population have been removed and the target population members missed by the list have been added. If we can generate corrected lists from this distribution, then for each draw we could obtain population counts for all cells of interest by simple

tabulation. The tabulations obtained by repeating this for each simulated corrected list, represent a sample from the joint posterior distribution of the cell counts. Summaries of this distribution such as the median, other quantiles, and approximate credible intervals can be obtained straightforwardly.

Introducing the notation Y to denote the cell-location for an individual in the target-list union, $\tilde{Y}$ to denote the cell location in the sample-list union,

$$[\mathbf{X}_i|\boldsymbol{\xi}] \overset{\text{indep}}{\sim} G(\boldsymbol{\theta}),\ i = 1, \ldots, N_U$$

$$[Y_i|\mathbf{X}_i, \boldsymbol{\xi}] \overset{\text{indep}}{\sim} \text{Multinomial}(1, \boldsymbol{\phi}(\mathbf{X}_i)),\ i = 1, \ldots, N_U \quad (1)$$

$$[\tilde{Y}_i|Y_i, \mathbf{X}_i, \boldsymbol{\xi}, \boldsymbol{\lambda}] \overset{\text{indep}}{\sim} H_Y(\boldsymbol{\lambda}, \mathbf{X}),\ i = 1, \ldots, N_U \quad (2)$$

letting $\xi = (\phi, \theta)$ where ϕ denotes the vector of parameters for the models for $\phi^{under}(\boldsymbol{x})$, and $\phi_{01}(\mathbf{x})$, letting $\phi(x) =$ = (ϕ_{11} (x_i), ϕ_{10} (x_i), ϕ_{01} (x_i)) denote the vector of cell probabilities at covariate setting x, and assuming the covariate values in the target-list union are drawn from some distribution G(θ) we have the model

where if Y is the Bernoulli distribution with possible values (1, 1) and (0, 1) with Pr($\tilde{Y}$ = (1, 1)| Y = (1, 1), **X**, **λ**) = λ(**X**);if Y = (1, 0), $H_{(1,0)}$(**λ**, **X**) is the Bernoulli distribution with possible values (1, 0) and (0, 0) with Pr($\tilde{Y}$ = (1, 0)| Y = (1, 0), **X**, **λ**) = λ(**X**); if Y = (0, 1), $H_{(0,1)}$(**λ**, **X**) is the degenerate distribution with Pr($\tilde{Y}$ = (0, 1)| Y = (0, 1), **X**, **λ**) = 1. Sampling of the target population has no impact on the group that is on the list but not in the target population. To complete the model we must specify a prior for the model parameters. We assume a priori independence so $p(N_U, \phi, \theta) = p(N_U)p(\phi)p(\theta)$. Further prior specification details will be application specific. The observed data is $\boldsymbol{D}^{obs} = (X_i, \tilde{Y}_i; i : \tilde{Y}_i \neq (0,0)0$. This is the sample-list union. The extra information required to obtain the complete target-list union can be characterised as $\boldsymbol{D}^{mis}$ = ($\boldsymbol{Y}^{mis}$, $\boldsymbol{X}^{mis}$) where $\boldsymbol{Y}^{mis}$ are the unobserved target-list cell locations for individuals not in the sample but on the list (i.e. in the ($\tilde{Y}$ = (0, 1) cell), and $\boldsymbol{X}^{mis}$ represents the covariate values for people missed by both the list and the survey (i.e in the $\tilde{Y}$ = (0, 0) cell). Note that since people observed in the $\tilde{Y}$ = (0, 1) group are a mix of those on the list but not in the target population (Y =(0,1)) and people on the list and in the target population (Y = (1, 1)) but not selected into the population sample, the true target-list cell location for individuals in this group are not directly observed. Given $\boldsymbol{D}^{mis}$ we could obtain the target population by first forming $\boldsymbol{D}^{full}$ = ($\boldsymbol{D}^{mis}$, $\boldsymbol{D}^{obs}$) and then dropping records with Y = (0, 1). The primary inferential task is therefore to obtain $p(D^{mis}|\, D^{obs})$:

$$p(N_U, \boldsymbol{D}^{mis}|\, \boldsymbol{D}^{obs}) = \int p(\boldsymbol{D}^{mis}|\, \boldsymbol{D}^{obs},\, N_U, \boldsymbol{\xi})\, p(N_U, \boldsymbol{\xi}|\boldsymbol{D}^{obs}) d\boldsymbol{\xi}$$

$$= \int p(\boldsymbol{D}^{mis},\, N_U,\, \boldsymbol{\xi}|\boldsymbol{D}^{obs}) d\boldsymbol{\xi}$$

A Gibbs sampling approach can be applied to simulate the joint distribution of unknowns $p(\boldsymbol{D}^{mis}, N_U, \boldsymbol{\xi}|\boldsymbol{D}^{obs})$. The generated draws of $\boldsymbol{D}^{mis}$ can then be used in conjunction with $\boldsymbol{D}^{obs}$ to produce a Monte Carlo representation of the posterior distribution for the population counts. The Gibbs sampler alternates between sampling from the following full conditional distributions: (i) $p(\theta|\emptyset, N_U, \boldsymbol{D}^{mis}, \boldsymbol{D}^{obs})$; (ii) $p(\emptyset|\theta, N_U, \boldsymbol{D}^{mis}, \boldsymbol{D}^{obs})$; (iii) $p(\boldsymbol{D}^{mis}, N_U|\boldsymbol{D}^{obs}, \emptyset, \theta)$. Steps (i) and (ii) amount to reasonably standard Bayesian computations since they are conditional on the full data. The components of $\boldsymbol{D}^{mis}$ can be simulated sequentially using the following decomposition

$$p(\boldsymbol{D}^{mis}, N_U|\boldsymbol{D}^{obs}, \emptyset, \theta) = p(N_U|\boldsymbol{D}^{obs}, \emptyset, \theta)p(\boldsymbol{X}^{mis}|N_U, \boldsymbol{D}^{obs}, \emptyset, \theta)p(\boldsymbol{Y}^{mis}|N_U, \boldsymbol{X}^{mis}, \boldsymbol{D}^{obs}, \emptyset, \theta).$$

An alternative approach to inference for the target population that is simpler computationally uses the idea that at each setting of the covariates the target population count, N_T (x) can be approximated as

$$N_T(x) = N_L(x)\frac{1-\phi^{over}(\boldsymbol{x})}{1-\phi^{under}(\boldsymbol{x})}$$

where $\phi^{over}(\boldsymbol{x}) = \phi_{01}(x)/(\phi_{11}(x) + \phi_{01}(x))$, which is the list over-coverage, and $N_L(x)$ is the list count at **X** = **x**. Justification for this approximation is given in Graham and Lin (2019). A straightforward strategy for implementing this approach involves weighting each list record by the ratio $w_i = \frac{(1-\phi^{over}(\boldsymbol{x_i}))}{(1-\phi^{under}(\boldsymbol{x_i}))}$. Estimated total population and sub-population counts can then be obtained by summing the weights for individuals in the populations of interest. The weights are a function of coverage model parameters, and posterior inference therefore follows directly from the posterior distribution for the coverage model parameters. In practice, we compute a set of weights for each draw from the posterior for the coverage model parameters. Posterior distributions for population counts can be obtained simply by computing the relevant weighted counts for each set of weights. The administrative list supplemented with a set of replicate weights also provides a convenient unit-record representation of the target population. The variation across the replicate sets of weights represents uncertainty due to estimating and adjusting for under and over-coverage. In several simulated examples we have found close agreement in estimates obtained from the weighting approach and from directly estimating finite population counts. Consequently, in what follows we concentrate on this weighting approach.

Since the weights that adjust for under and over-coverage depend only on the coverage model parameters, our inference task is simplified because we need only obtain the posterior distribution for these model parameters. As a further simplification we use the conditional likelihood for ∅, which can be

computed directly from the observed data, in place of the full likelihood which is difficult to compute because it involves integrating over the missing data. The conditional likelihood for ∅ is

$$L^{cond}(\emptyset) = p(\tilde{\boldsymbol{Y}}^{obs}|(\tilde{\boldsymbol{Y}} = (0,0), \boldsymbol{X}^{obs}, \emptyset)$$
$$= \prod_{i:\tilde{Y}_i=(1.1)} \frac{\lambda(\boldsymbol{x}_i)\emptyset_{11}(\boldsymbol{x}_i)}{1-(1-\lambda(\boldsymbol{x}_i))\emptyset_{10}(\boldsymbol{x}_i)} \prod_{i:\tilde{Y}_i=(1.0)} \frac{\lambda(\boldsymbol{x}_i)\emptyset_{10}(\boldsymbol{x}_i)}{1-(1-\lambda(\boldsymbol{x}_i))\emptyset_{10}(\boldsymbol{x}_i)} \prod_{i:\tilde{Y}_i=(0,1)} \frac{\emptyset_{01}(\boldsymbol{x}_i)+(1-\lambda(\boldsymbol{x}_i)\emptyset_{11}(\boldsymbol{x}_i)}{1-(1-\lambda(\boldsymbol{x}_i))\emptyset_{10}(\boldsymbol{x}_i)} \quad (3)$$

and an approximation to the posterior for the coverage model parameters can be obtained as $p\left(\emptyset\middle|\boldsymbol{D}^{obs}\right) \propto p(\emptyset)L^{cond}(\emptyset)$.

4. Sample design and the inclusion probability. A standard design for household surveys in official statistics involves a two-stage area-based sampling design, where, at the first stage, small geographic areas known as primary sampling units or PSUs are sampled, and at the second sampling stage households are selected within PSUs. If there is no within household non-response, the number of responding households in a PSU, divided by the total number of households in the PSU is the PSU level inclusion probability (λ). Thus, when supported by a well-maintained household list or register, a conventional area based multi-stage sample design can yield PSU-specific inclusion probabilities. In order for the PSU specific inclusion probability to apply to all individuals within a PSU, two assumptions must hold: (i) there must be no within household non-response; (ii) household non-response must not vary with dimensions of household composition that relate to covariates included in the analysis. For the time-being we make these assumptions but note that further refinement of the inclusion probability may be possible if household level covariates that are reasonable proxies for household composition are available.

In order to support the multi-stage design we model the coverage probabilities, $\phi^{under}(\boldsymbol{x})$ and $\phi_{01}(x)$ at the PSU level and specify hierarchical models to pool information over the PSUs. The within PSU likelihood has the form, given by (3), with $\lambda(\mathbf{x})$ set to the PSU-specific inclusion probability. The overall model likelihood is obtained by multiplying the PSU specific likelihoods. For realistic applications, posterior inference for ∅ using the conditional likelihood approach requires MCMC methods. We have implemented the model using the Bayesian modelling software STAN (Carpenter et al, 2017).

5. Application. We constructed a simulating target population by sampling 800,000 records from the 2013 Census usually resident population while keeping a hierarchy of household, PSU, stratum, territorial authority, and region. From this we drew a subset of 31,881 records to represent under-coverage. These records were excluded from the simulated administrative list.

A sample of 59,223 records were selected from the target population to represent the over-coverage group. Thus, the simulated list comprised 800, 000−31, 881 = 768, 119 records from the initial target population selection of 800, 000, plus an additional 59, 223 records representing over-coverage. The total size of the simulated list was therefore 828, 342. The under and over-coverage proportions were 4% and 7%, respectively. The records selected into the under and over-coverage groups were selected using coverage probabilities chosen to reflect plausible patterns of variation in coverage by age, sex and area. Finally, we simulated a coverage survey by taking a two-stage sample of 5% from the target population. We applied two selection methods in the second sampling stage: a standard Stats NZ household survey design approach of sampling 12 households from each sampled PSU, and an alternative approach of selecting all households. For both scenarios we assumed a household level response probability of 0.9 for PSUs, and no within household non-response. The number of sampled PSUs was 1377 for the former and 231 for the latter. We adopted weakly informative priors for all model parameters, similar to the prior specifications given in Bryant et al (2017, pp 10-12). Estimates of the list over-coverage probabilities, ϕ^{over} by age and sex under the two designs are shown in Figure 1. The estimates obtained for the ϕ^{over} parameter under the "full-PSU" approach are clearly more precise than under the "12-household per PSU" approach, even though the overall sample size is the same in both cases. The reason for this is that increasing the within PSU inclusion probability strengthens inference for ϕ_{01} because, within PSUs, the observed (0, 1) cell then contains a higher proportion of true over-coverage. Our hierarchical modelling takes advantage of this design by modelling at the PSU level. The results presented in Figure 1 suggest our proposed methodology, based on conditional likelihood, is promising.

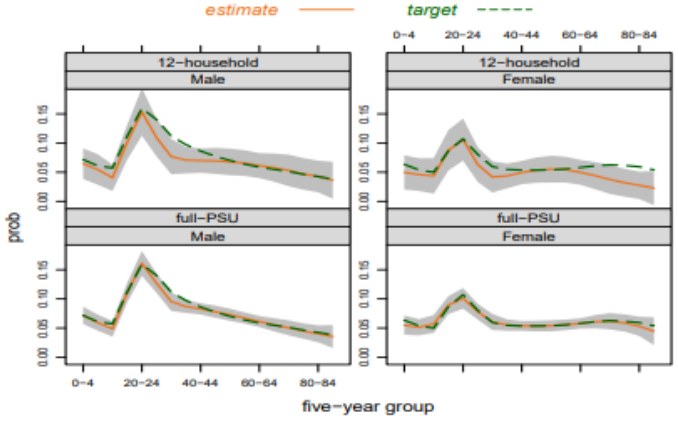


Figure 1: Plot of estimated over-coverage with 95% credible intervals compared to true values

6. Conclusions. We have outlined a Bayesian approach to estimating the size and distribution of a population using an administrative list in conjunction with a coverage survey sample drawn from the target population and linked to the list. In a simulated

data example our methodology showed encouraging results, particularly when the "full-PSU" sampling method was used. Currently our methodology assumes no within household non-response to the survey, and that, within PSUs household response does not vary by household composition or other household characteristics. Analysis of patterns of response to existing household survey data may shed light on the validity of these assumptions, and perhaps, help build prior models to adjust λ, for within PSU response variations. An alternative stategy is to extend our methodology to include a second administrative list, in which case the survey inclusion probability can be estimated within the model (Zhang, 2015). However, introducing a second administrative list raises the very real possibility of linkage error between the two lists. The methodology outlined here also needs to be extended to cope with measurement error or misclassification of list variables, as well other practical challenges. We are currently investigating these issues.

References

1. Bryant, J., Dunstan, K., Graham, P., Matheson-Dunning, N., Shrosbee, E., Spiers, R. (2016) Measuring Uncertainty in the 2013 Base Estimated Resident Population. Statistics New Zealand Working Paper No 16-04. Statistics New Zealand, Wellington NZ.
2. Bycroft, C. (2015) Census Transformation in New Zealand: Using administrative data without a population register. Statistical Journal of the IAOS, 31(3), 401-411.
3. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J. Li,P., and Riddell, A. (2017) Stan: A probabilistic programming language. Journal of Statistical Software 76(1). DOI 10.18637/jss.v076.i01.
4. Graham, P., Lin, A. (2019) Bayesian and approximate Bayesian methods for small domain population estimation from an administrative list subject to underand over-coverage. Unpublished manuscript available on request from the authors.
5. Zhang, L.C. (2015) On modelling register coverage errors. Journal of Official Statistics, 31(3), 381-396.

Population size estimation from incomplete multisource lists: A Bayesian perspective on latent class modelling

Davide Di Cecco[1], Marco Di Zio[1], Brunero Liseo[2]
[1] ISTAT, via Cesare Balbo, 16, 00184 Rome
[2] MEMOTEF, Sapienza Rome University, viale del castro laurenziano 9, 00161 Rome

Abstract
We propose a capture–recapture model for estimating the size of a population of interest based on a set of administrative sources and/or surveys in the presence of out-of-scope units (false captures). Our Bayesian approach makes use of a certain class of log - linear models with a latent structure. We also address the presence of sources providing partial information implementing a Gibbs Sampler algorithm which generates from the posterior distribution of the population size in presence of missing data. The proposed method is applied to simulated data sets.

Keywords
Bayesian Analysis, Capture–Recapture, Latent Class

1. Introduction

The use of administrative data for the production of official statistics is providing many new opportunities and methodological challenges. In estimating the size of the usual resident population by municipality, in almost all national statistics institutes the use of traditional censuses is gradually being replaced with the use of administrative sources, which provide "signs of life" for the population of interest. While undercoverage was the main issue in the former approach, overcoverage is the main concern with administrative data. By overcoverage we mean the erroneous inclusion in the lists of units which do not belong to our population, i.e., out-of-scope units. Of course, overcoverage can be encountered in surveys and census too, but almost always it consists of duplicated records generated by linkage errors, which are now commonly addressed even in capture–recapture contexts. In administrative data, on the other hand, linkage errors constitute just one of the factors, in a number of possible reasons for erroneous captures. In general, administrative data are gathered by other organizations for non-statistical purposes. Hence, units and variable definitions may not align perfectly. For example, the available information pertaining the registered events, their temporal description, their legal definition may vary in each source, and their harmonization can be difficult. As a consequence, each list may contain different subpopulations of out-of-scope units, and the assignment of the units to our target population may not be error free. Obviously, any piece of

available information should be included in the process of identification of the erroneous cases in the lists. Ideally, recognizing and deleting spurious cases should constitute a first phase of our analysis, after which some capture-recapture technique might be used on the "cleaned" data. However, in many cases, the available information does not suffice to single out every false capture, and there will remain a certain portion of uncertainty for which we have no capability of discerning the cause of error. In practice, the main approach in official statistics is the following: all available administrative sources are integrated into a unique population statistical registers. The register is coupled with an ad-hoc coverage survey (in the same way as censuses were coupled with an additional post enumeration survey) to exploit a Dual Systems Estimator (DSE). Then, the overcoverage rate is estimated on the basis of the comparison between the (supposedly) error-free survey and the administrative data via some supervised model, and then used to "correct" the DSE in some way. An original approach, called Trimmed DSE, and proposed in Zhang et al (2017), consists in an iterative procedure which removes units and estimate a DSE until a stopping criterion is satisfied. The authors prove that, if the survey has no overcoverage, the procedure has some optimal properties of convergence. The Dual System approach, including the aforementioned, has the remarkable property of being partic¬ularly robust (see, e.g., Chao et al 2001), and it does not rely on any complex model specification. Our approach, on the converse, relies on a Multiple Record System, where one considers the various adminis¬trative sources separately, in order to exploit the information redundancy. There exist various proposals in literature which use complex model to deal with false captures in multiple lists, particularly in animal abun¬dance problems, see, e.g., da Silva (2009), Wright et al (2009), and Link et al (2010). However, in all those works, the false captures are essentially duplicate linkage errors. To our knowledge, the only contributions dealing with false captures with no restrictive hypothesis on the source of error in multiple record systems are Overstall et al (2014) and Fegatelli et al (2017). The former proposes a Bayesian log-linear model, the latter extends that work in order to include latent variables. However, in both cases, only a single source list is assumed to suffer from false captures. When considering administrative sources separately, a series of methodological issues arises:

- It is necessary to take into account possible dependencies among the various sources.
- While DSE is known to be robust with respect to violation of basic hypotheses (e.g., the homogene¬ity of capture probabilities), this is not true in general in Multiple Record Systems.
- In our framework, administrative sources often target specific categories of citizens (e.g., people in a certain age range), leaving subset of the population with null probability of being captured.

Our proposal relies on the following assumption: all possible erroneous captures are defined as random classification errors under a binary classification model. That is, we hypothesize two subpopulations: one comprising the out-of-scope units, and the other the in-scope units. Then a two-component latent class model would adequately describe our data. To model possible dependencies among captures of a same individuals in different sources, we relax the classic conditional independence assumption of latent class models and assume a general log-linear model for the joint distribution. To address the problem of subpopulations that are uncatchable for some sources, we treat the uncatchable units as missing information and develop an inferential approach to deal with missing data. This model has been proposed in Di Cecco et al (2018). Here we present a Bayesian approach to estimate the size of the population, addressing the challenges listed above.

2. Methodology

Assume k lists or capture occasions are available, and let Yi be the random variable indicating whether a unit is included in the i–th list, $i = 1, ..., k$ (i.e., has been captured in the i–th occasion):

$$Y_i = \begin{cases} 1 \; if \; a \; unit \; is \; captured \; in \; the \; i-th \; list; \\ 0 \qquad\qquad\qquad\qquad otherwise. \end{cases}$$

Let $Y = (Y_1,\dots, Y_k)$ denote the capture profile of a unit, and let $\{P(Y = y) = p_y\}_{y\in\{0,1\}^k}$ be the associated probability distribution. Let $U(i)$ be the set of units that are catchable by list i, and let U be $U_i U(i)$. Let U_1 be our target population, with $U1 \subset U$. The cardinality of U is N, the one of U_1 N_1. Let X be the latent variable identifying the units belonging to our target population:

$$X = \begin{cases} 1 \qquad\qquad if \; a \; unit \; belongs \; to \; U1 \; ; \\ 0 \qquad\qquad otherwise. \end{cases}$$

Let n_y be the number of units having capture profile $\mathbf{y}$, of which $n_{x,y}$ belong to the latent class x so that $n_{0,y} + n_{1,y} = n_y$. The total number of observed unit is n_{obs}, while the units having capture history $\mathbf{y} = \mathbf{0} = (0, \dots, 0)$ are unobserved, so that $\sum_{y\neq 0} n_y = n_{obs}$, and $N = n_{obs} + n_0$. Note that $n_{1,0}$ is the number of units in U_1 that are not captured, while $n_{0,0}$ is the number of uncaptured units which are in U but not in U_1. We are interested in estimating $N_1 = \sum_y n_{1,y}$. The latent class model under the conditional independence assumption (CIA) can be equivalently expressed as the mixture model

$$p_{\mathbf{y}} = \sum_{x=0,1} p_x \prod_{i=1}^{k} p_{y_i|x}, \tag{1}$$

where $p_{y_i|x}$ indicates the conditional probability $P(Yi = y_i|X = x)$, or as in the log-linear model notation

(2) $$[XY_1][XY_2]\cdots[XY_k],$$

which reports only the higher order interactions (generators) of the model. Any additional interaction term in (2) represents a relaxation of the CIA.

2.1 Prior distributions: The usual priors for log–linear models are based on Multivariate Gaussian distributions. Here we propose a different prior based on Dirichlet distributions. We find this approach easier in terms of elicitation of prior knowledge, and also from a computational point of view, since it allows us to develop a Gibbs sampler for obtaining a sample from the posterior distribution of $N1$, so avoiding the use of a Metropolis–Hastings algorithm. To illustrate our proposal we start with decomposable models. In this case the prior distribution is simply the product of Dirichlet densities. In Dawid et al (1993) it has been demonstrated that, if G is the dependence graph of the decomposable model, $\{\mathcal{L}_1,\ldots,\mathcal{L}_g\}$ are the maximal cliques of G, and $\{\mathcal{L}_1,\ldots,\mathcal{L}_g\}$ are defined as

$$\mathcal{S}_i = \mathcal{C}_i \cap \bigcup_{j=1}^{i-1} \mathcal{C}_j \qquad i = 2,\ldots,g,$$

the joint distribution can be written as the product of conditional distributions:

(3) $$p_G = \prod_{i=1}^{g} p_{\mathcal{C}_i} \left(\prod_{j=2}^{g} p_{\mathcal{S}_j} \right)^{-1} = p_{\mathcal{C}_1} \prod_{i=2}^{g} p_{\mathcal{C}_i|\mathcal{S}_i},$$

where p over a (sub)graph is the (marginal) distribution over the variables included in the (sub)graph. Let θ be the vector of parameters $\theta = (P_{\mathcal{C}_1}, P_{\mathcal{C}_2|\mathcal{S}_2}, \ldots, P_{\mathcal{C}_g|\mathcal{S}_g})$. We define a prior distribution on θ as follows: for each $P_{\mathcal{C}_i|\mathcal{S}_i}$ and for each value of $\mathcal{S}_i$ we set a Dirichlet distribution defined for each possible combination of values $\mathbf{y}_{\mathcal{C}} \in \{0,1\}^{|\mathcal{C}|}$ of the variables in $\mathcal{C}$. The Dirichlet densities are independent by construction, and this class of priors is conjugate to (3). In the case of a general log–linear model, we made use of the "Bayesian iterative proportional fitting" described in Schafer (1997) in order to sample from a "Constrained Dirichlet". That is, we generate samples from a Dirichlet distribution which satisfies the constraints given by the log–linear model. This prior has been rarely utilized in literature, and, as far as we know, has never been utilized in capture–recapture analysis. Regarding N, in

accordance with the literature on Bayesian capture–recapture, sensible options include:

i) Jeffreys' prior, i.e. $\pi(N) \propto 1/N$;

ii) a hierarchical Poisson prior: $N \sim Poi(\lambda)$, $\lambda \sim Gamma(a, \beta)$;

iii) Rissanen's prior (Rissanen 1983), $\pi(N) \propto 2^{-\log^*(N)}$, where $\log^*(N)$ is the sum of the positive terms in the sequence $\{\log_2(N), \log_2(\log_2(N)), ...\}$.

We further assume that N and θ are a priori independent.

2.2 Missing data: We propose a strategy useful to properly include sources which do not operate over certain subpopulations ("incomplete lists"). In fact, if we treat the uncatchable units as sampling zeros, the final population size estimate would be biased. The idea is to treat the incomplete lists as Missing at Random (MAR) information, i.e. assuming that, if they could operate on the whole population, they would retain the same joint distribution as in the observed subpopulations. In addition, we assume that we can distinguish whether a unit has not been captured in a list by chance or because it is out of the scope of that list, i.e., we can divide the population in strata where different set of lists operates. Then, certain profiles of the captured units are considered as partially observed, and we develop a data augmentation algorithm that imputes the complete capture histories using the rest of the data given the model. We distinguish completely observed capture profiles, $\mathbf{y}$, from the partially observed capture profiles $\mathbf{y}_{mis}$. In addition, for each stratum, we have a structural zero $\mathbf{z}$ consisting in a different combination of zeros and missing values. For example, in a 4-lists scenario with 2 strata, one where all lists operate and one where the first list does not operate, we have the structural zero $n_{0,0,0,0}$ in the first strata, and $n_{*,0,0,0}$ in the second, where the asterisk denotes the missing information. Then, our Gibbs algorithm at iteration $t + 1$ has the following steps:

(1) we sample the components of $\theta^{(t+1)}$ from their posterior conditional Dirichlet distributions (constrained or not);

(2) for each observed $\mathbf{y}$ and $\mathbf{y}_{mis}$, we randomly divide all the observed values n_y and $n_{y\,mis}$ into the corresponding consistent complete sequences $n_{x,y}$ according to their conditional probabilities;

(3) if we adopt $\pi(N) \propto 1/N$, it has been demonstrated in Manrique-Vallier et al (2014) that we can sample all structural zero cells counts n_z from a Negative Multinomial distribution. Otherwise, if we choose an informative prior for N, we can use a Metropolis-Hasting step to generate a value for $N(t+1)$ and then conditionally sample the structural zero cells such that $\sum_z n_z = N - n_{obs}$;

(4) for each generated n_z, we sample all complete sequences $n_{x,y}$ consistent with $\mathbf{z}$.

3. Simulations

We report the results of a simulation for empirically assessing the proposed algorithm. We considered 5 lists, *A*, *B*, *C*, *D*, *E*, and defined a scenario with three strata: one with all sources, one where 4 out of five sources operate, and another one where just three sources operate. We set N = 10000 and a proportion of out–of–scope units (both captured and non-captured) equal to 40%, so that the desired total $N1$ is 6000 in expectation. The model parameters have been set in such a way that the proportion of unobserved units (both in–scope and out–of–scope) is about 30%.

Model selection is a critical issue for capture–recapture modeling as population size estimate can be sensitive to changes in the parameterization. To have a hint on the robustness of the procedure under mis-specification of the model, we generated a sample from model *[XABC][XD][XE]*, and estimated $N1$ under two different models: the CIA model *[XA][XB][XC][XD][XE]*, and the (non – decomposable) model including all 15 second order interactions but no higher order parameters. Results regarding the second model can be viewed in Figure 1, where one sees that the true value of $N1$ is comprised in the 95% credibility interval, despite 5 parameters are missing (those relative to *[ABC], [XAB], [XAC], [XBC]* and *[XABC]*).

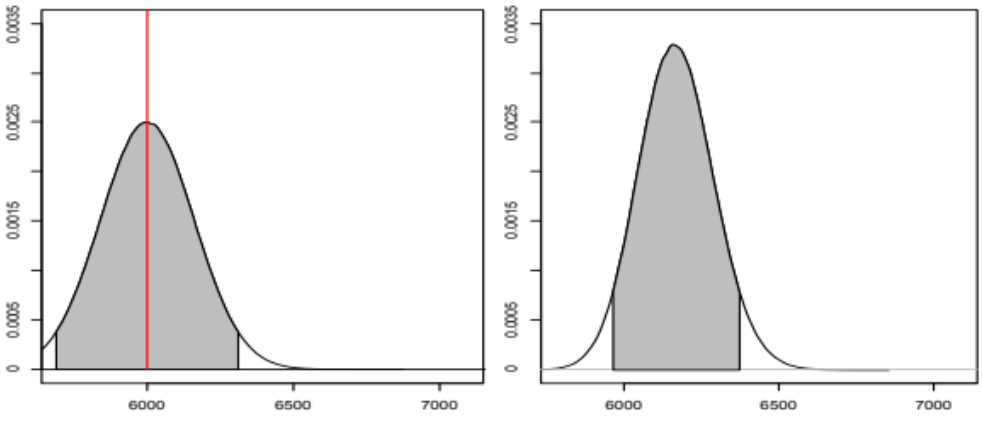


FIGURE 1. Posterior distributions of *N1* under the generating model *[XABC][XD][XE]* (left) and under the all–second–order–interactions model (right). The orange line indicates the true value of *N1,* the gray area the 95% HPD.

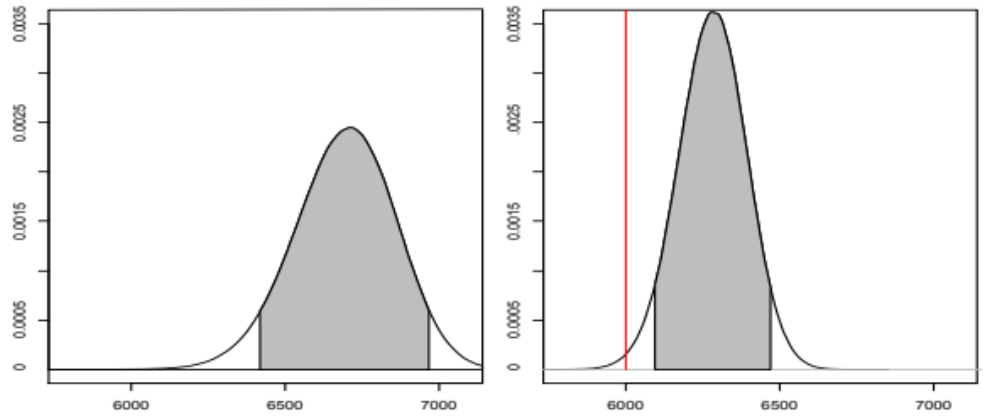


FIGURE 2. Posterior distributions of *N1* under the CIA model, flat priors(left) and informative priors (right). The orange line indicates the true value of *N1,* the gray area the 95% HPD.

On the converse, the left panel of Figure 2 shows that the estimated posterior distribution of *N1* under the CIA model is far from the real value. To evaluate the influence of the prior distributions to compensate for the model misspecification, we set an informative prior in the following way: we mimicked an informative context coming from an audit sample by taking a 5% sample of the generated complete population *[XABCDE],* and fixed the parameters of the Dirichlet prior equal to the observed counts in that sample. As one can see in the right panel of Figure 2, even though informative priors influence the posterior in the right direction, their contribution seems insufficient to even include the true value of N *1* in the credibility interval.

References

1. Chao, A., Tsay, P., Lin, S., Shau, W., and Chao, D. (2001). The applications of capture-recapture models to epidemiological data. Statistics in medicine, 20(20):3123–3157.
2. da Silva, C. Q. (2009). Bayesian analysis to correct false-negative errors in capture-recapture photo-ID abundance estimates. Brazilian Journal of Probability and Statistics, 23(1):36–48.
3. Dawid, A. P. and Lauritzen, S. L. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. The Annals of Statistics, 21(3):1272–1317.
4. Di Cecco, D., Di Zio, M., Filipponi, D., and Rocchetti, I. (2018). Population size estimation using multiple incomplete lists with overcoverage. J. Off. Stat., 34(2):557–572.
5. Fegatelli, D. A., Farcomeni, A., and Tardella, L. (2017). Bayesian population size estimation with censored counts. In Capture-Recapture Methods for the Social and Medical Sciences, pages 371–385. Chapman and Hall/CRC.

6. Link, W. A., Yoshizaki, J., Bailey, L. L., and Pollock, K. H. (2010). Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. Biometrics,66(1):178–185.
7. Manrique-Vallier, D. and Reiter, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. Journal of Computational and Graphical Statistics, 23(4):1061–1079.
8. Overstall, A. M., King, R., Bird, S. M., Hutchinson, S. J., and Hay, G.(2014). Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. Statistics in medicine, 33(9):1564–1579.
9. Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. The Annals of statistics, 11(2):416–431.
10. Schafer, J. L. (1997).Analysis of incomplete multivariate data. CRC press.
11. Wright, J. A., Barker, R. J., Schofield, M. R., Frantz, A. C., Byrom, A. E., and Gleeson, D. M. (2009). Incorporating genotype uncertainty into mark–recapture-type models for estimating abundance using DNA samples. Biometrics, 65(3):833–840.
12. Zhang, L.C., and Dunne, J. (2017) Trimmed dual system estimation. In Capture-Recapture Methods for the Social and Medical Sciences, pages 237–257. Chapman and Hall/CRC

Forecasting banknote demand at the Reserve Bank of Australia

Richard Finlay[1]
Reserve Bank of Australia

Abstract

I detail the Reserve Bank of Australia's method for forecasting banknote demand and deciding on the level of contingency banknote stocks to hold. In approaching this, the RBA has two main objectives: most importantly, to always have sufficient banknotes to meet public demand; and second, subject to the first requirement, to minimise cost. Regarding banknote forecasts, we use statistical autoregressive models rather than models based on economic and financial variables. There are two main reasons for this: autoregressive models have historically performed reasonably well; and using explanatory variables such GDP, the number of ATMs, etc. to forecast banknote demand necessitates forecasting these variables, which in practice has been difficult to do accurately. Autoregressive models will miss turning points or economic shocks, however, and so we hold sufficient buffer stocks to allow for this and ensure that we do not run out of banknotes. I also briefly discuss methods that we have employed to estimate the proportion of banknotes in circulation used for transactional and store-of-value purposes.

Keywords

bank note demand; forecasting; Central bank; contingency banknote stocks

1. Introduction

Banknotes, being complex physical objects, cannot be created instantly. In fact, the sophistication of modern banknote designs and the number of different components that they contain means that it can take many months between notifying a printworks that more banknotes are required, and the finished product being supplied to the central bank. Demand for banknotes, on the other hand, can change very rapidly, as was the case in Australia and many other countries around the onset of the global financial crisis. This miss-match between a relatively slow production process and potentially fast changes in demand, coupled with central banks' unwillingness to countenance running out of banknotes, means two things: central banks try to forecast future banknote demand as accurately as possible; and, knowing that they will nonetheless get things wrong, they employ backup strategies such as holding substantial contingency stocks. This paper discusses the Reserve Bank of

[1] The author is from Note Issue Department and would like to thank Ben Smagarinsky for his input.

Australia's approach to these two issues, and also briefly summarises recent work on estimating the contribution to overall banknote demand from various sections of the economy and society.

2. Forecasting banknote demand

The RBA currently uses an ARMA time-series model to forecast seasonally adjusted banknote demand. The general model takes the form:

$$y_t = \alpha + \delta t + \sum_{i=1}^{p} \beta_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

where y_t is the logarithm of seasonally adjusted banknote demand for a given denomination, t is the time period, and ϵ_t is an error term. This model is simple and robust and has performed relatively well over time. Its main drawback, however, is the lack of any economic factors driving the model, meaning that structural changes to banknote demand will only be picked-up with a lag. The RBA has also investigated using models with more economic and financial market content, such as error correction (ECM) or vector error correction (VECM) models. These models are based on cointegrating relationships between banknote demand and various macroeconomic variables. The general form of the ECM investigated is

$$\Delta y_t = \beta_0 + \delta_1 \Delta y_{t-1} + \gamma \Delta X_t + \lambda(y_{t-1} - \alpha X_{t-1}) + \epsilon_t$$

$$X_t = [CashRate_t;\ ATM_t;\ EFTPOS_t;\ TWI_t;\ GDP_t;\ Population_t]$$

where y_t is seasonally adjusted banknote demand for a specific denomination in period t, X_t is a vector of variables including the cash rate, number of ATMs, number of EFTPOS terminals, the trade weighted index (TWI) of the Australian dollar exchange rate, nominal GDP, and population, and α is a vector of long-run parameters.[2] A global financial crisis dummy variable is also included for the higher denominations. All variables except the cash rate are in logarithms. The speed of adjustment term is λ. The main drawback of the ECM model is that it relies on various macroeconomic series, and these need to be forecasted in order to produce banknote demand forecasts. To do this we use: vintages of the RBA's real GDP and CPI forecasts (to construct a nominal GDP forecast); historical overnight forward rates as implied by financial markets (for the cash rate); long run average growth rates to project forward data on ATMs, EFTPOS terminals and population, as no forecasts are available; and we assume that the exchange rate (TWI) remains constant.

[2] Nominal GDP and population capture the effect of rising household income, and more people, on banknote demand. The cash rate accounts for the opportunity cost of holding cash. I include the number of ATMs and EFTPOS machines to control for the ease of withdrawing currency. The exchange rate captures foreign demand for Australian banknotes.

A VECM is similar the ECM above, except that instead of estimating a separate equation for each denomination, one estimates all denominations concurrently, which allows shocks to one denomination to effect demand for another. Mathematically, the equation we estimate has the same basic form as the ECM above, except that scalars and vectors become vectors or matrixes (with y_t now a vector of the log demand of each denomination at time t).

To compare the ARMA and ECM models we use rolling out-of-sample forecast performance. Graphs 1-5 below show the mean absolute percentage error (MAPE), relative to observed outcomes, for forecast horizons from one month to three years into the future, where our sample starts in 2003 and runs to between 2008 and November of 2015 (mean squared percentage errors show a similar profile).

One can see that in terms of forecasting performance, the ARMA model is as good as or better than the ECM for all denominations except the $50. The superior forecast performance of the ECM for the $50 denomination (but not the other denominations) appears to be due to a better fitting ECM model, and in particular demand for $50 banknotes being more closely aligned with the macro factors considered than is the case for the other denominations (for example more factors are significant in the ECM than for the other denominations, and the adjusted R2 is over twice as high as for the next-best model). This likely reflects that the $50 is the main ATM banknote, and so is the denomination that will respond most to changes in consumer spending. This is also reflected in the $50 displaying the most seasonality of all denominations. In contrast, the lower denominations, which are used as change, appear to adjust velocity rather than quantity in response to changes in spending, while demand for the $100 is in general hard to model accurately.

3. Determining contingency stocks

While we make every effort to forecast future banknote demand as accurately as possible, our forecasts will inevitably contain errors, sometimes large ones. Historically, the median absolute forecast error one year ahead, expressed as a percentage of total outstanding banknotes, has ranged from slightly less than 1 per cent for the $10 denomination to 3 per cent for the $50 denomination. The largest error, which occurred around the onset of the global financial crisis for the $50 denomination, was 11 per cent. Given that the Reserve Bank has a strong aversion to running out of banknotes in the event of a spike in demand, we hold substantial contingency stocks. These are calibrated to cover, for each denomination, a one-year outage to the printworks during which time no new banknotes are delivered, and, on top of this, an increase in demand proportional to that seen during the global financial crisis. This delivers a contingency stock ranging from about 15 per cent of circulation for the $100 denomination to 35 per cent of circulation for the $50 denomination. Holding contingency stocks is not costless, although

we judge the costs to be acceptable and in any case less than the costs to society that would be incurred if the RBA was not able to supply public demand. For a central bank, the costs associated with holding banknote inventory primarily consist of (i) the opportunity cost associated with interest foregone on the purchase price of the banknotes (which will vary with the interest rate cycle, but is currently low for most countries given the setting of global interest rates); and (ii) when a banknote series is withdrawn, the full purchase price of any unused old-series banknotes, which become obsolete. The note-issuing function of the central bank cannot easily effect the opportunity cost without reducing its buffer holdings, but the second cost component can be reduced by running-down old-series banknote stocks as far as is prudent ahead of a new series being released, and/or allowing the old and new series to co-circulate so that the life of the old-series is not cut short.

4. Sources of banknote demand

To form a long-term forecast of banknote demand and plan strategically, it helps to understand what factors have driven banknote demand in the recent past. There are a number of ways one could approach this, including estimating models of demand using macroeconomic and financial factors as explanatory variables similar to those discussed above, and then constructing various scenarios. In this section we take a different approach and use a number to techniques to estimate how much of recent demand has been driven by transactional uses of cash, compared with non-transactional uses. For further details see Finlay, Staib and Wakefield (2018) and Wakefield and Finlay (2018).

4.1 Transactional banknote demand

The most visible source of banknote demand is for banknotes that are used to facilitate day-to-day transactions in Australia, which we call 'transactional demand'. Transactional demand is also the easiest to estimate since transactional banknotes continuously flow through the cash distribution system. As a result, we are able to employ a number of different methods to estimate the size of this source of demand. We first describe each method and then present a summary of our combined results at the end of this section.

4.1.1 The counting method

Our first approach is to estimate the stock of cash held in various physical locations that are part of the transactional stock, including banknotes in wallets, ATMs and bank branches, cash depots, tills and self-service checkouts and gaming machines, and banknotes held by tourists. These figures are aggregated to form an economy-wide estimate. This calculation by necessity relies on a number of assumptions and will miss any cash held in locations not directly considered. Despite these limitations, the approach is

useful as it provides a broad sense-check on other estimates arrived at through more abstract means and also offers a tangible basis from which to think about the transactional stock of cash. We use two approaches to estimate the stock of cash held in each location:

- estimating the number of a given location (e.g. the number of tills) and multiplying this by an estimated average amount held per location; and
- converting flow data to a stock by making assumptions about the velocity of cash through a particular location.

This method suggests that the transactional stock of cash has risen from around $9 billion at the end of 2002 to around $13 billion as at June 2018. This corresponds to an annualised growth rate of around 2 per cent, which is well below the 6 per cent growth rate in total outstanding banknotes over the same period. As a result, the transactional stock's share of the total is estimated to have fallen from 30 per cent to around 20 per cent by value according to this method (Graph 6).

4.1.2 The banknote life and banknote processing methods

We now assume that the non-transactional stock of cash consists only of hoarded $50 and $100 banknotes. While this may not be exactly true, it is probably not far off the mark: for example, almost all large claims for damaged banknotes that are submitted to the Reserve Bank are for the $50 and $100 denominations. We then try to find some data affected by this hoarding. In each of the methods below, this involves data where the $50 and $100 banknotes behave very differently to the other denominations. This difference can then be used to estimate transactional demand for the $50 and $100. Adding this to the value of outstanding $5, $10 and $20 banknotes gives an estimate of overall transactional demand.

The banknote life method

Banknotes reach the end of their lives (become 'unfit') for two main reasons: excessive inkwear, which will tend to increase in a relatively linear fashion with banknote use; and mechanical defects such as tears, which can be thought of as random events that can occur at any stage, but whose cumulative probability of having occurred also increases with use. Given that all denominations of banknotes are initially of similar physical quality, the speed at which certain denominations become unfit is closely related to the frequency with which they are handled. Since banknotes are most commonly handled when used as a means of payment, banknotes used in transactions should have a shorter lifespan than banknotes not used in transactions.

If we assume that all banknotes used in transactions wear out at a similar rate, then the 'excess life' of high-denomination banknotes relative to low-

denomination banknotes can be attributed to hoarding. Based on this insight, we estimate that over the past three decades the share of $100 banknotes used for transactions has fallen from around 20 per cent to just 3 per cent; the share of $50 banknotes used for transactions has fallen from around 35 per cent to 25 per cent; and the transactional share by value of all banknotes has fallen from around 45 per cent to around 20 per cent (Graph 6).

The banknote processing method

One can apply the same idea to data on the frequency with which different banknote denominations are processed by cash depots. In particular, cash depots process and fitness-sort banknotes lodged by commercial banks and large retailers, but do not process any banknotes that are hoarded or otherwise are not part of the transactional stock of cash. Thus, broadly speaking, only the transactional stock of banknotes passes through cash depots, and the rate at which banknotes pass through depots is an indication of transactional cash use. Given this, if we assume that the processing frequency of transactional $50 and $100 banknotes is equal to the processing frequency of the $20 banknote, then the difference between the observed processing frequency of $50 and $100 banknotes and that of the $20 is the result of hoarding. In fact the true processing frequency of transactional $50 and $100 banknotes is likely to be higher than the $20 denomination as almost all $50 and $100 banknotes received by retailers will be banked, whereas some $20 banknotes will be given as change. This suggests that this method will deliver an upwardly biased transactional share estimate. Applying the same technique used in the banknote life calculations suggests that the transactional stock has fallen from around 55 per cent by value of total outstanding banknotes in the late 1990s to around 40 per cent now (Graph 6).

4.1.3 The velocity method

Another way to estimate the stock of cash used for transactions is to first estimate the flow of cash payments made by consumers, and then convert this flow into a stock. The flow of cash payments and the stock of banknotes used to make them are related, but one banknote can be used in multiple transactions; banknote velocity ties the two concepts together, as described in the equation below.

$$\text{Flow of cash payments} = \text{Velocity of transactional stock} \times \text{Value of transactional stock.}$$

We estimate the flow of cash payments through time by scaling the value of card payments with the cash-to-card payment ratio as recorded periodically in the Reserve Bank's Consumer Payment Survey (CPS). To estimate the velocity of transactional cash, we map out the cash cycle: banknotes start at a

cash depot, are transported to an ATM or bank branch, pass to a consumer's wallet or purse, get spent at a business, and then get returned to a bank and/or cash depot. For some legs of this journey we have accurate data – for example, we know the flow into and out of cash depots, and so can calculate the average time a banknote spends in a depot – whereas for other aspects we need to use judgement. Our estimates suggest that the velocity of transactional cash has declined over the past decade, and that, on average, a transactional banknote takes a little over one month to complete a full cycle. To estimate the transactional stock of cash we divide our estimates of cash payments by our estimates of velocity. With cash payments estimated to be broadly stable and velocity estimated to be falling, we estimate the transactional stock to be gradually increasing over recent years and in the range of $15–25 billion currently. These results suggest that transactional cash accounts for around 20–30 per cent by value of total banknotes (Graph 6).

4.1.4 The seasonality method

The final way we estimate the transactional share of banknotes is via the seasonality present in banknote demand. The logic works as follows: demand for cash displays a predictable seasonal pattern, with a peak around Christmas and a trough in the winter months. This seasonality resembles that of consumer spending, which suggests that it is driven by seasonality in transactional cash demand. On the other hand, non-transactional cash demand (for example, hoarding for store-of-value or numismatic purposes) is unlikely to contain significant seasonality. As a result, if most cash is transactional, then the seasonality of cash demand should closely match the seasonality of cash spending; conversely, if non-transactional demand is more important, then there will be less seasonality in cash demand than in spending. As such, and similar to the banknote life and banknote processing methods, the degree of seasonality present in cash demand, when compared with the seasonality of cash spending, is an indication of the share of cash used for transactional purposes. To account for the stock/flow mismatch between outstanding banknotes and cash lodgements, we adjust the seasonality of the lodgement data with three estimates of the seasonality present in the velocity of transactional cash, and then average over the three estimates. Our results suggest the transactional stock of cash has been largely unchanged over the past decade. Converting to a share of the value of banknotes outstanding suggests that transactional demand has declined from around 40 per cent of banknotes by value in 2009 to 25 per cent currently (Graph 6).

4.1.5 Overview

Overall, the methods that we employ suggest that somewhere between 20 and 40 per cent by value of outstanding banknotes are used to facilitate transactions within Australia (Graph 6). Notably, all methods show

that this share is in decline. Although each estimation method is imperfect, we take comfort from the fact that a number of different methods yield a broadly similar trend.

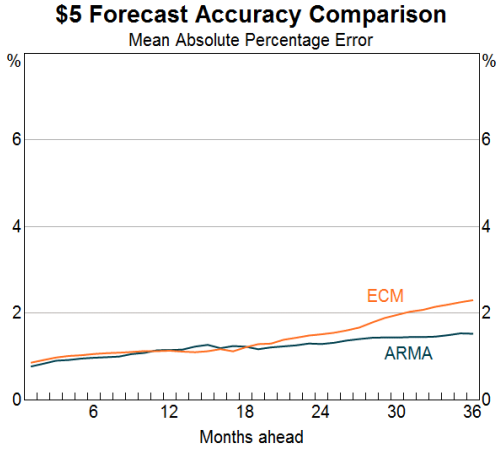
4.2 Non-transactional banknote demand: the residual

While we presented the banknote life, banknote processing and seasonality estimates above as indirect estimates of transactional cash demand, they can equally be seen as indirect estimates of non-transactional demand. These methods suggested that 20–40 per cent of outstanding banknotes by value were used to facilitate transactions, implying that 60–80 per cent by value are used for non-transactional purposes.

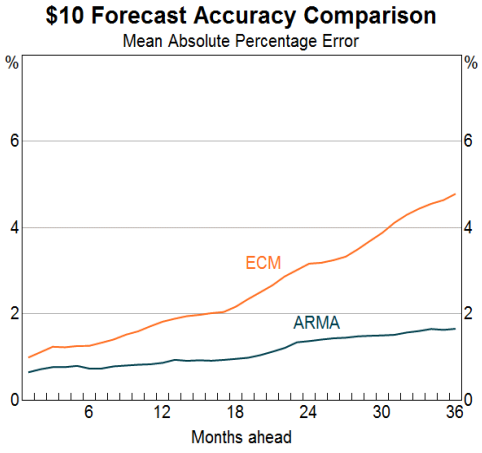
References

1. Finlay R, A Staib and M Wakefield (2018), 'Where's the money? An investigation into the whereabouts and uses of Australian banknotes', RBA Research Discussion Paper 2018 12.
2. Wakefield M and R Finlay (2018), 'Understanding Demand for Australia's Banknotes', RBA Bulletin, December.

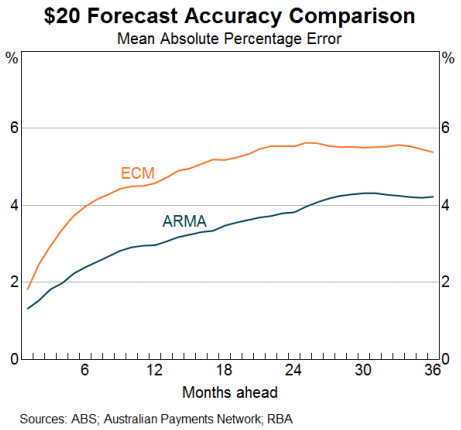
Graph 1



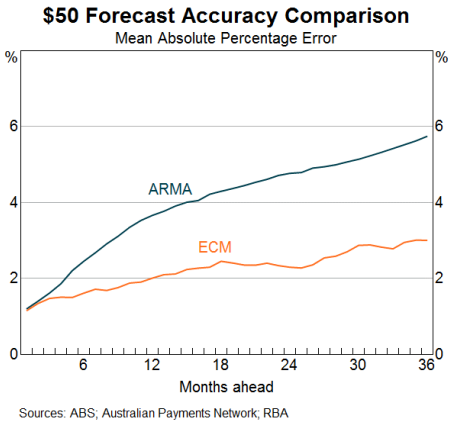
Graph 2



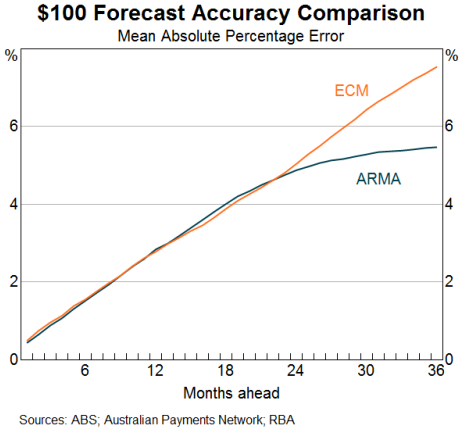
Graph 3



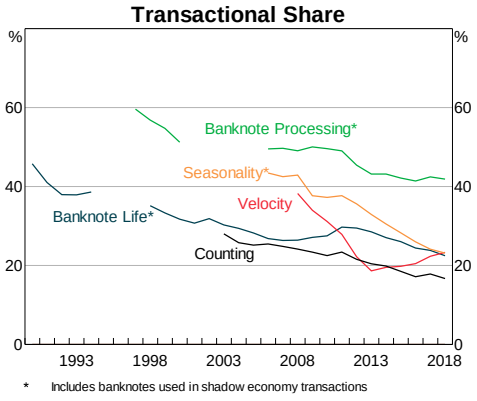
Graph 4



Graph 5



Graph 6



Currency demand forecasting: The Philippine experience

Iluminada T. Sicat
Bangko Sentral ng Pilipinas, Currency Management Sector
Manila, Philippines

Abstract

One of the functions of Central Banks (CBs) is to issue currency to ensure that that there will be ample note and coin to support, on time and in the denomination required, the transaction payment needs of the economy, including sufficient inventory to cover spikes in demand due to cyclical factors, unexpected shocks or delay in currency production. In this respect, many CBs have developed econometric models to estimate their currency demand using factors such as economic, financial, and demographic and other explanatory variables as determinants. A forecast that results to over-estimation of currency demand could exert pressure on CB resources (particularly budgetary allocation for currency printing/minting, vault capacity, among others). However, an underestimation of currency forecast has far more serious impact since this may pose reputational risks on the CB by compromising the country's clean note and coin policy in the short-run to serve the economy's currency requirement. Hence, the optimum econometric currency model is one that minimizes forecast error (i.e., difference between actual and forecast currency demand). In the case of the Philippines, its currency demand equation models have evolved over time with the view to capture structural changes in the economy, thus improve the forecasting model's robustness, reliability and goodness of fit.

Keywords

currency order, currency in circulation, forecast performance, buffer stock, mean absolute percent error

1. Introduction

The Bangko Sentral (BSP) is the agency in-charge of maintaining monetary and financial stability, and a safe and efficient payment and settlement system. In line with these mandates, one of its functions involves currency issue. Maintaining the value and confidence on the Philippine currency is a key part of its responsibility. In keeping with this responsibility, the BSP has to ensure that there is sufficient supply of good quality notes and coins to support the requirement of the economy and facilitate financial transactions. Failure to supply sufficient volume of required currency may hamper payment and settlement transactions, thereby adversely affect economic activity, or worsen

the quality of bank notes in circulation, which may, in turn, lead to inability to distinguish genuine notes from counterfeits.

The BSP also has the exclusive authority to withdraw from circulation notes and coins which are already unfit and mutilated and replace them with good quality ones. In line with the authority to issue currency, and withdraw and replace the unfit ones from circulation, the BSP plans the currency order based on the forecasts of currency demand. Specifically, it determines how much banknotes and coins should be ordered and in what denomination, to serve the requirement of the economy, while keeping enough inventory from its vault as a safety precaution to meet unexpected spikes in demand.

The BSP plans the currency order two years in advance to cover a ten-year forecast horizon. As a matter of procedure, the estimates for first two year period of the forecast horizon are binding, which means that once the forecasts have been approved by the BSP Monetary Board[1] then these become the basis for the production plans of SPC, plant capacity assessment, budget allocation for procurement of banknote printing work materials and coin blanks needed, as well as budget and planning for shipment of currency, including logistics, and mobilization of security personnel.

Because of the many important considerations attached to currency order, it is imperative that estimates of the BSP on annual currency demand must be near accurate.

2. Role of Cash in the Philippines

Cash remains an essential medium of exchange in the Philippines. Data indicate that it continues to grow and has yet to show signs of decline in either use or preference. Results of the 2014 Financial Inclusion Survey indicated that 98 percent of retail transactions in the Philippines continued to be facilitated through cash.[2] Currency in circulation, which comprises 97 percent banknotes and 3 percent coins, has been steadily growing in tandem with economic growth. In 2018, currency in circulation which grew by 17.6 percent outpaced the growth of nominal gross domestic product (GDP) at 10.2 percent during the same period. Moreover, share of currency in circulation (CiC) to total liquidity (M3) has been stable moving within a narrow band of 12 - 16 percent over the past 18 years. New payment systems such as usage of electronic transaction for payment and settlement can lessen demand and use of cash. However, such remain very low in the Philippines. Based on a 2015 report by Better than Cash Alliance, only 1 percent of the 2.5 billion monthly retail payments were done electronically.[3] This is attributed, in part, to the high ratio

[1] The BSP Monetary Board is the policy making body at the BSP.

[2] www.bsp.gov.ph/payments/nrps_overview.asp

[3] Gilberto M. Llanto, Maureen Ane D. Rosellon, and Ma. Kristina P. Ortiz. E-finance in the Philippines: Status and Prospects for Digital Financial Inclusion (Discussion paper series no. 2018-22).

of unbanked population. According to the 2014 survey of Philippine Consumer Finance, 86% of Filipino households remain to be unbanked.[4] Access to electronic payments is mostly associated with maintenance of a deposit account with the bank. Concerns on dependable IT infrastructure, interoperability concerns across digital financial service providers and IT-related security issues associated with digital financial criminality such as cyber thievery, fraud, identity theft and cyber-attack, are among the reasons cited for the low usage of electronic and online payment transactions. While the BSP has implemented many initiatives in the regulatory front and continue to introduce many more initiatives as part of its financial inclusion program to address the barrier and unveil greater use of electronic payments in retail transactions, many believe that cash will remain a dominant medium of exchange in the retail payment space over the medium- to long-term. The BSP hopes to bring up share of non-cash transactions to 20 percent by 2020. Given this premise, it is thus vital that the BSP should have a robust forecasting model of currency demand.

3. Currency Demand Framework

The process in currency forecasting in the Philippines consists of 2 levels, one involves forecasting at the national (aggregate) level, and the other, is for regional allocation. Also, part of the forecasting exercise is to estimate the denominational mix. Forecasted currency by denomination is based primarily on historical trend but is also adjusted to take stock of the changes or anticipated changes in denominational preference over time, caused by emergence of alternative payment infrastructure such as use of electronic fund transfer at point of sale, use of plastic cards for payments, and increase in the number of automated tellering machines (ATMs). It has been observed that demand for higher denomination banknotes rose as the number of ATMs installed increased. Similarly, the number of electronic fund transfer at point of sale such as use of debit card, has a negative effect on the demand for low denominations. A comparison of the denominational mix of the currency in circulation in 2000 and 2018 in the Philippines showed an increase in the share of higher denominations, particularly for 1000P, 200P, and 100P notes. By contrast, the share of 20P declined. This shift in denominational mix was in step with expansion in the number of off-site ATMs. Banks demand higher denominated notes to save them from more frequent reloading of cash in their machines.

Aggregate Demand. In forecasting the aggregate currency demand for the Philippines, the determinants are based on economic and non-economic variables. Specifically, currency demand is influenced by 3 factors, namely 1) the transaction demand to support economic needs and changes in price

[4] Bangko Sentral ng Pilipinas. Consumer Finance Survey (2014).

levels, 2) the level of currency stock necessary for safety cushion, and 3) the desired fitness level of the banknotes and coins in circulation before they shall be replaced. BSP's own statistical tests showed that economic activity, as captured by GDP, and the price level have long-run and strong co-integration relationship with currency demand. There exists linear relationship between demand for banknotes and these two variables. Specifically, increases in real economic growth lead consumers to increase their usage of all denominations, whereas an increase in price levels lead to an increase in the demand for higher denomination bank notes. The use of GDP targets and price level expectations provide the forward-looking information into the forecasts.

4. Currency Demand Forecasting Models

Forecasting model to estimate currency demand is assessed periodically for robustness of the model. In this respect, the BSP used different forecasting models over time to enhance the reliability of the forecasts under each model. In the late 1990s, currency in circulation for the forecast year is derived from the projected M3 based on the historical share of CiC to M3. Annual M3 projections, in turn, were based on a target real GDP growth and expected inflation rate. Currency demand is then calculated based on change in CiC.

***Model 1**: (1999 – 2004)* $f(\frac{CiC}{M3})$

Starting 2005, CiC for the forecast year is projected by multiplying previous year's CiC by a growth factor using macro assumptions of real GDP growth target and expected inflation rate.

***Model 2**: (2005 – 2-11)* $(CiC\ growth)_t = inflation + 1.17 * (real\ GDP\ growth)_t$

By 2012, CiC for the forecast year is based on an OLS (ordinary least square) model[5] using consumer price index (CPI) and GDP as explanatory variables. A dummy variable and an error-correction term are added to the baseline model to capture the effects of global financial crisis on demand for banknotes, as well as the difference between the forecast and the actual currency in circulation. It was noted that financial market uncertainty and financial volatility caused demand particularly for higher denomination banknotes to rise.

***Model 3:** (2012 – 2017)* $LOG(CiC)t = Ii0 + Ii1\ LOG(CPI)t + Ii2\ LOG(GDP)t + Ii3DFIN08t + t$

[5] A linear regression model that aims to minimize the sum of the squared error.

In 2018, the currency forecast model was again enhanced to add 2 more dummy variables to adjust the impact of seasonality in cash demand observed in Q2 and Q4, and adjustment terms to account for serial correlation.

Model 4: *(2017 - present) Model 3 + dummy variable for Q2 + dummy variable for Q4 + AR & MA terms*

Where:

DFIN08 – dummy variable for financial crisis
t – error correction term
dummy variable Q2
dummy variable Q4
AR & MA terms – to account for serial correlation

5. MAPE: Evaluating Reliability of Forecasting Models

The following presents an assessment of the performance of the models used relative to actual. The best performing model is that which produces the lowest forecast error on average, based on the mean absolute percent error or MAPE. MAPE refers to the variance or difference between the forecasts and actual. Lower MAPE indicates the model's improving goodness of fit. As can be gleaned from the table below, MAPE has been declining from a high of 33% to 6.9%, which can indicate improving forecast performance. Similarly, the accuracy of the forecast is even more enhanced with the introduction of dummy variables and autoregressive terms in Model 4, yielding a MAPE of only around 2% based on in-sample and out-samples estimates.

Model	Model 1	Model 2	Model 3	Model 4
MAPE	33.3 %	12.9 %	6.9 %	2.0 %

6. Other Factors Affecting Currency Demand

In addition to economic variables, the currency demand framework is also underpinned by the need to maintain a level of inventory for pre-cautionary needs. In the Philippines, the BSP Monetary Board, in the past, has approved to maintain 2 types of safety cushions called the "buffer" stock and "contingency reserves", equivalent each to 3 months average withdrawal for the past 3 years. The buffer stock is meant to serve as cushion or cover primarily against uncertainty regarding spikes in demand for cash arising from unexpected business cycle, or uncertainty arising from timing in supply delivery. Meanwhile, the contingency reserves are meant to provide supply in

case of extraordinary circumstances such as those arising from natural disaster or system-wide financial bank run. Buffer stocks are maintained for all denominations of notes and coins, whereas contingency reserves are only for high denomination banknotes. Recently, the BSP Monetary Board has integrated the 2 types of prudential reserves into one, and called it simply as "Buffer stock", for easier monitoring and to be consistent with the practice of other central banks. Currency demand forecasts also consider the lifespan of the banknotes/coins and the need to replace them whenever the quality of the banknotes and coins are no longer fit for circulation. Notes and coins in circulation must be of a certain quality, in order to promote easier detection of the authenticity of the notes, hence enhance the integrity of the currency. Because the Philippines consists of many islands surrounded by water, climate is humid, and wet markets exist, the Philippine banknotes, particularly, lower denominated ones have a shorter lifespan, hence faster replacement of notes in circulation. Over the years, the BSP has improved the Philippine peso substrates particularly for lower denomination banknotes in order to lengthen their lives. The aggregate currency demand in the Philippines is derived by summing the projections coming from these 3 variables.

$$\text{Currency demand}_t = f\ (\text{growth, mandated buffer, replacement rate})_t$$

To calculate how much banknotes and coins will be ordered, the end-of-period inventory or the stock on hand is deducted from the projected annual currency demand.

$$\text{Currency Order} = \text{Demand}_t - \text{Inventory}_{t-1}$$

Outcome of the currency forecast exercise indicate that the Philippines will need an average of about 4-4.5 billion banknotes annually for the next 5 years. Meanwhile, forecasts on denominational mix are based on historical trend. Error-correction term is incorporated to adjust for substitution of denominations due to unavailability of requested denomination. The error-term also captures change in denominational mix arising from financial innovations or alternative payment infrastructure.

Regional Currency Distribution. Having forecasted the aggregate demand for the entire country, the next step is to allocate the regional distribution according to each 23 currency distribution centers of the BSP. ARIMA[6] (auto-regressive integrated moving average) model using time series data is used

[6] Auto-regressive integrated moving average (ARIMA) models are a form of statistical time series which seek to capture the dynamic stochastic aspect of time-series data through a lagged variable structure while abstracting from more predictable features in the data (such as trends and seasonality). Univariate (single vector) ARIMA is a forecasting technique that projects the future values of a series based entirely on its own inertia. Its main application is in the area of short-term forecasting requiring at least 40 historical data points.

for regional demand forecasting. Specifically, historical monthly series on the net withdrawal is used owing to their high frequency data, and because they proxy the demand for currency associated with economic growth. Moreover, our exercises show that time-series banknote forecasts for the region simply seek to extrapolate from past behavior and have shown to have better forecast performance than their more theoretical counterparts.

7. Important Take Away In Forecasting Currency

1) Cash is expected to remain an important instrument to settle transactions in the Philippines. While new payment opportunities such as electronic payments may reduce the demand and use of cash in the Philippines, especially for lower denominations, their overall impact may still be negligible because of certain barriers and limitations.

2) Given these, forecasting the demand for cash using statistical models provides better basis of estimates of how much currency and in what denomination to order.

3) Be mindful of alternative payment infrastructure as they may alter preference for certain denomination.

4) Finally, it is important to constantly evaluate and test the models for robustness, reliability and goodness of fit.

Complex seasonal autoregressive model compared to machine learning methods for cash volume forecasting

K. Prokopenko, B. Bruijnis, M. Symotiuk
Giesecke+Devrient Currency Technology GmbH, Prinzregentenstrasse 159, 81677 Munich, Germany

Abstract

As expert in cash volume forecasting Giesecke+Devrient has invested in software that uses algorithms for cash volume forecasting based on data available from bank branches, cash devices and retail stores. This paper describes the validation of the data and the algorithms applied to this data. Based on statistical validated data from cash handling locations in Western Europe the seasonal behaviour of cash usage has been described. Based on the trends found three forecasting models have been validated: Complex Seasonal ARMA (CSARMA), LSTM recurrent neural network and XG Boost decision trees. The latter two are standard machine learning (ML) algorithms while CSARMA is a new approach based on complex seasonal autoregressive model of global trend, annual and weekly shapes. The forecast validation shows that for this data set the CSARMA algorithm outperforms the ML based algorithms.

Keywords

Cash volume forecasting; Seasonality; Time series; Machine learning; Auto regression; Neural networks; Decision trees

1. Introduction

Despite rumours about the cashless society in several countries like Georgia the absolute cash usage is still growing [5]. In other countries cash usage is in decline with Sweden as one of the most appealing examples. However, despite this movement to other means of payment, a fully cashless society is not there yet [6].

The combination of decreased cash usage and the desire to allow cash to remain as payment method demands Central banks and other actors in the cash supply chain to minimise the cost of handling cash while keeping cash availability without cash-outs. One cash cost improvement that doesn't require the infrastructure to be changed is the improvement of the cash usage forecast. In this paper the focus is on customer incoming and outgoing payments.

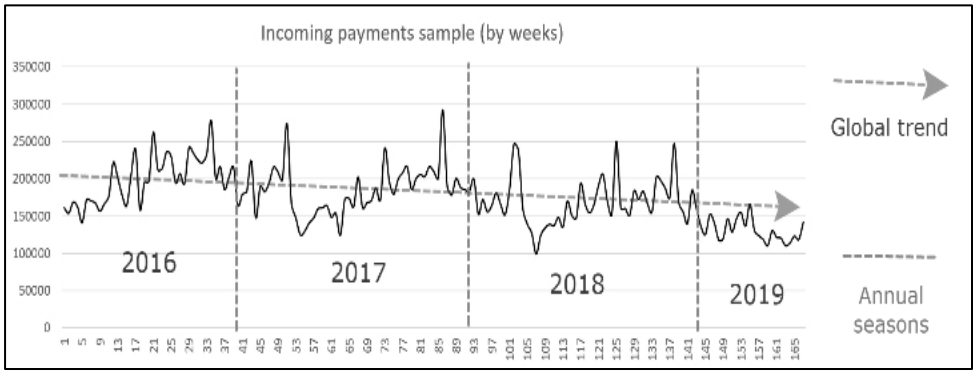
Using data of several companies in Western Europe Giesecke+Devrient has developed algorithms that predicts cash requirements in the market considering seasonality. This paper describes the decision making towards the algorithm to use. By knowing the demand from the end customer Central

Banks can use this input for the long-term planning for cash printing and destruction.

2. Data research

Analysis was provided using more than 3 years of daily total cash payments from 670 branches of a wide cash payments network in Western Europe. From the data processing perspective, cash payments (deposits and withdrawals) can be interpreted as daily time series which may contain trend and seasonality parts. It was highly important to analyze and understand the structure of the data and to find any regularities which should be used for the further modeling. The size of the daily-aggregated data sample is about 1200 points for incoming payments and 1200 points of outgoing payments for each branch.

Global trend and annual seasonality. Linear regression approach [2] was used for linear annual trend estimating. Autocorrelation function (ACF) was chosen as a tool for data structure analysis during research on weekly-aggregated data samples. There are strict spikes on 53th points which proves annual 52 weeks seasonality for both incoming and outgoing payments. The annual seasonality is more intensive for outgoing payments instead of less intensive annual and strict 4-5 weeks seasonality for incoming payments (Fig.1).



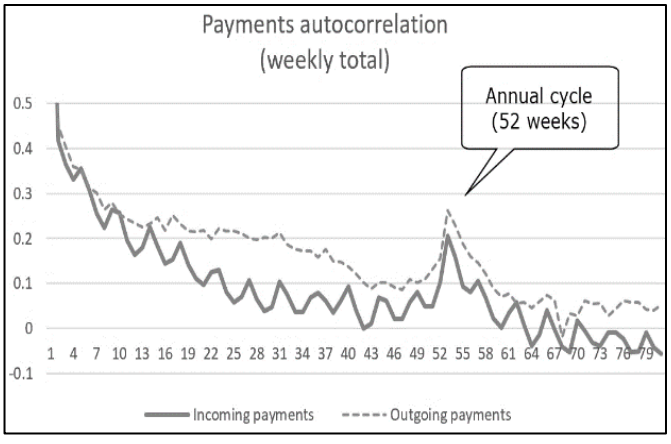


Fig. 1. An example of weekly-aggregated payments time series and averaged autocorrelation function of weekly-aggregated payments of 670 branches. 165 weeks (3+ years) of historical points and first 80 points of autocorrelation function are displayed.

Weekly seasonality. In addition to annual seasonality, strict weekly seasonality is present in the analyzed data both for incoming and outgoing payments data sets. Autocorrelation functions were calculated for each branch of the whole set of 670 branches and then its values were averaged for each point of autocorrelation function. Maximum values of ACF reflects weekly shape with values of 0.6152 for incoming and 0.752 for outgoing payments (Fig. 2).

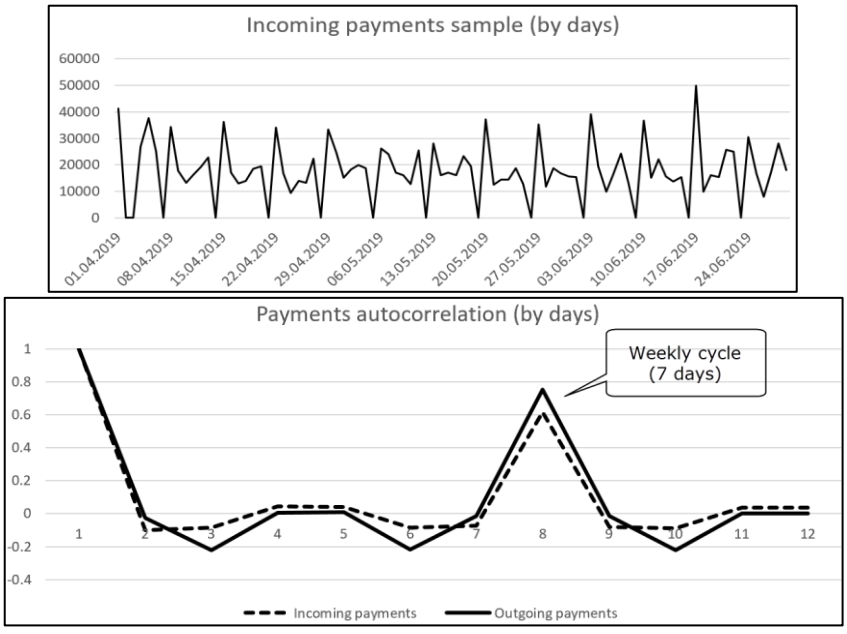
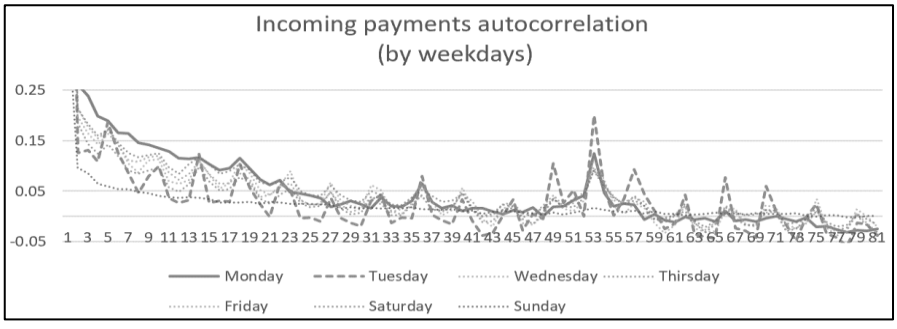


Fig. 2. An example of daily incoming payments time series and averaged autocorrelation function of daily-aggregated payments of 670 branches. Three months (90 days) of historical points and first 12 points of autocorrelation function are displayed.

Weekdays annual seasonality research. Incoming and outgoing payments data sets were investigated to search for annual seasonality of each weekday. Daily data samples were interleaved by weekdays and then autocorrelation functions were calculated for each weekday. Strict 52 weeks seasonality was detected, especially for Mondays and Tuesdays (Fig. 3.).



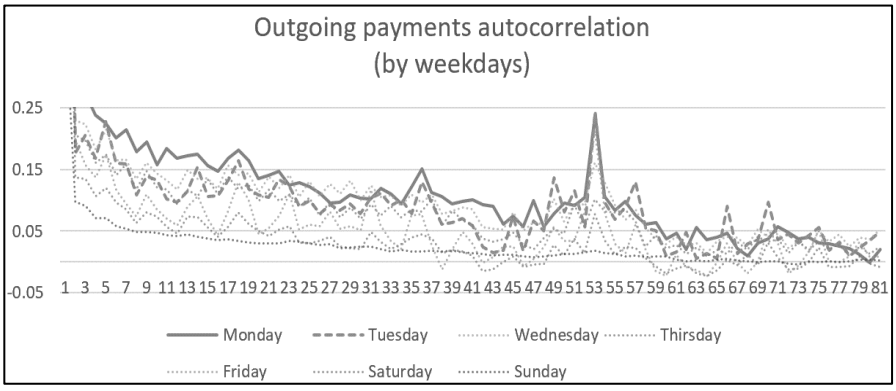


Fig. 3. Autocorrelation function of weekly incoming / outgoing payments interleaved by weekdays. First 80 points were calculated.
Strict 52 weeks seasonality is present especially for Mondays and Tuesdays.

Weekly shape normalization research. To avoid the impact of annual seasonality, the normalization procedure was applied on daily payments data sets. Thus, all daily payment values were divided by their total weekly value. Autocorrelation function of normalized daily payments was calculated (Fig. 4)

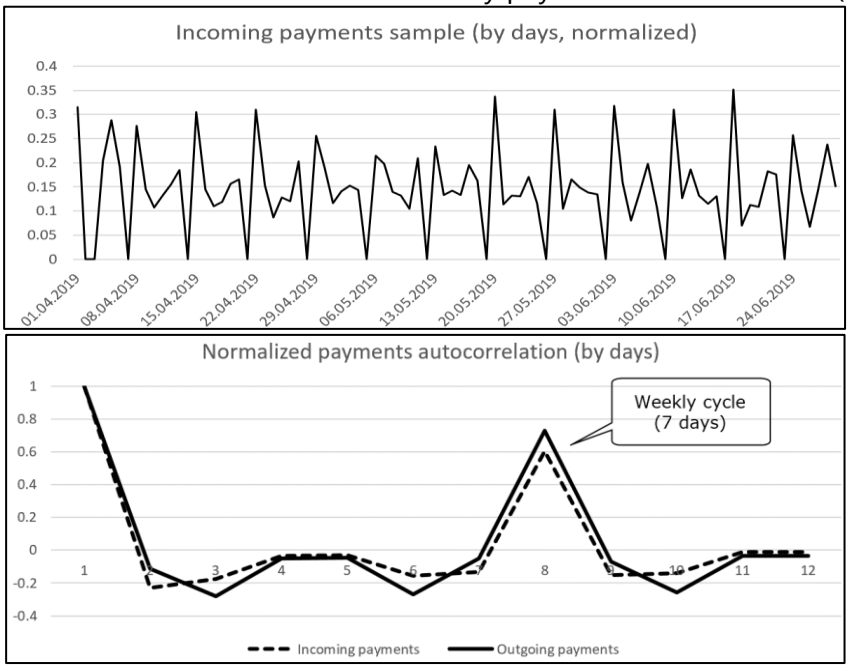


Fig. 4. An example of daily normalized incoming payments time series and averaged autocorrelation function of normalized daily-aggregated payments of 670 branches. Three months (90 days) of historical points (values were divided by their total weekly value) and first 12 points of autocorrelation function are displayed.

The autocorrelation function of normalized daily payments interleaved by weekdays was also calculated (Fig.5). It should be noted that the shape of normalized payments ACF is more meaning and informative than shape of original payment ACF.

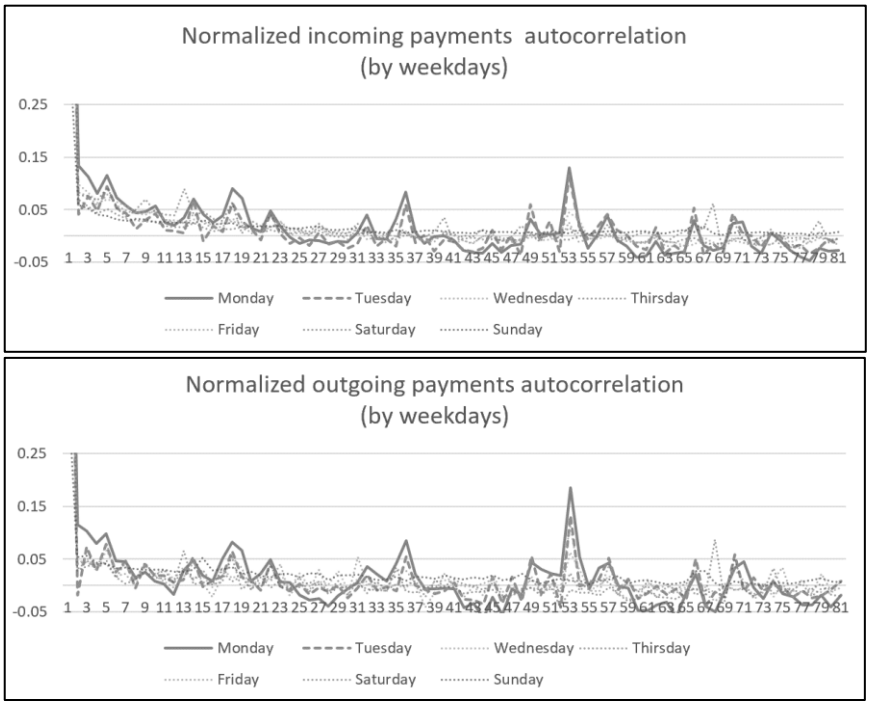


Fig. 5. Autocorrelation function of normalized weekly incoming / outgoing payments interleaved by weekdays. First 80 points were calculated. Strict 52 weeks seasonality is present especially for Mondays and Tuesdays

Stationarity analysis. Fisher criteria [1] with probability level = 0.01 was applied to evaluate the payments data stationarity. The results show the unstable behavior of most of branches (Fig.6).

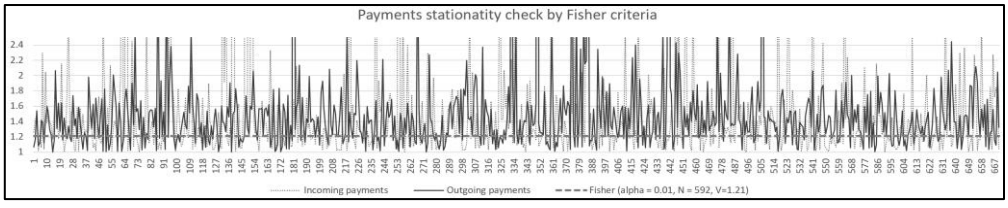


Fig.6. Stationarity check of 670 branches using Fisher criteria (F-test). Each branch has 1184 points were tested using Fisher criteria for alpha = 0.01, N = 592, V = 1.21. Number of branches with stationary incoming payments: 264 (39.40%), stationary outgoing payments: 195 (29.10%).

The normalization procedure was applied to deliver more stationary data sets. It should be noted that normalized payments have quite more stable behavior than original ones (Fig. 7).

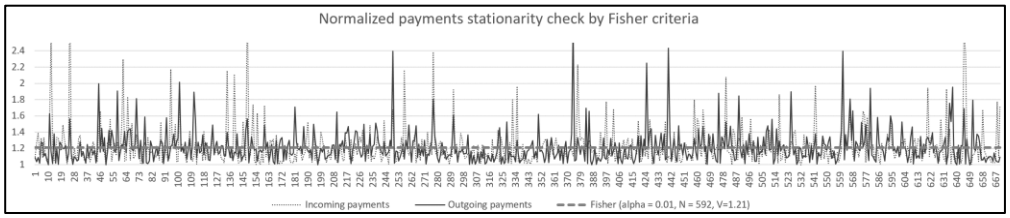


Fig. 7. Stationarity check of 670 branches using Fisher criteria (F-test). Values from each branch were divided by their weekly total value and tested using Fisher criteria for alpha = 0.01, N = 592, V = 1.21. Number of branches with stationary incoming payments: 428 (63.88%), stationary outgoing payments: 408 (60.90%).

3. Complex Seasonal ARMA model of cash payments

During research provided the following assumptions were stated, proved and then accepted as root points of the forecasting approach:

1. Weekly aggregated cash payments as time series has strict general trend and annual seasonality. For weekly aggregated series it usually has 52 full weeks seasonality.
2. Each branch has strict weekly coefficient shape. For example, for specific branch Monday's incoming payments can usually have 40% of total weekly incoming payments.
3. Coefficients of week days can also have seasonality behavior.
4. Payments normalization is a way to make data sets more stationary.

According to statements above cash payments model $CS\ ARMA\begin{pmatrix} p, a_1, \dots, a_p, P, A_1, \dots, A_p, \\ q, b_1^w, \dots, b_q^w, Q, B_1^w, \dots, B_Q^w, c^w, L \end{pmatrix}, w = 1, \dots, 7$ can be expressed as a recursive multi-layer seasonal autoregressive model of daily cash payments X_t:

$$X_t = T_{t\ div\ 7+1} * C_{t\ div\ 7+1}^{t\ mod\ 7+1} + \varepsilon_t \quad (1)$$

Where X_t is daily cash payment value at current date with index t, $T_{t\ div\ 7+1}$ is total payment of week where current date is placed, $C_{t\ div\ 7+1}^{t\ mod\ 7+1}$ is coefficient of week day of current date, ε_t is white noise.

Trend T in (1) is a seasonal autoregressive process explained by:

$$T_i = \sum_{j=1}^{p} T_{i-j} a_j + \sum_{j=1}^{P} T_{i-j-L} A_j + \varepsilon_i, i > L + Max(p, P) \quad (2)$$

where p and a_j are the order and coefficients of trend autoregressive model, P, A_j and L are the order, coefficients and lag of seasonal part of trend autoregressive model, ε_i is white noise. Weekdays coefficients $C^w, w = 1, \dots, 7$ are also set of seasonal autoregressive processes explained by:

$$C_i^w = c^w + \sum_{j=1}^{q} \left(C_{i-j}^w - c^w\right) b_j^w + \sum_{j=1}^{Q} \left(C_{i-j-L}^w - c^w\right) B_j^w + \varepsilon_i, i > L + Max(q, Q) \quad (3)$$

where $w = 1, \dots, 7$ is a weekday number, c^w is a constant, q and b_j^w are the order and coefficients of *w*-weekday's coefficient autoregressive model, Q, B_j^w and L are the order, coefficients and lag of seasonal part of *w*-weekday's coefficient autoregressive model, ε_i is white noise.

According to the theory of autoregressive processes [2], parameters $a_1, \dots, a_p, A_1, \dots, A_p, b_1^w, \dots, b_q^w, B_1^w, \dots, B_Q^w, c^w$ can be estimated by solving the system of linear equations composed from weekly totals and daily cash payment turnovers. Forecasting expression for daily payment value X_{N+t} is the following:

$$X_{N+t} = T_{(N+t)div\ 7+1} * C_{(N+t)div\ 7+1}^{(N+t)mod\ 7+1} \quad (4)$$

where $T_{(N+t)div\ 7+1}$and $C_{(N+t)div\ 7+1}^{(N+t)mod\ 7+1}$ were calculated according to (2), (3)

4. Experimental results validation

Validation of the CSARMA model was provided on payments data sets obtained from 100 branches. Each branches data sample has 3 years (1184 days) of both incoming and outgoing payments history. Train sample size = 1004 days, test sample size = 180 days. Error measures are explained in (Tab.1.).

Tab.1. Error measures used for forecasting accuracy calculation.
N is number of predicted points, x_i - actual values, $\bar{x}$ - mean of actual values, y_i - predicted values, $\bar{y}$ - mean of predicted values.

Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert y_i - x_i\rvert}{x_i} * 100\%$
Normalized Root Mean Square Deviation (NRMSD)	$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N}}, NRMSD = \frac{RMSD}{\bar{x}} * 100\%$
Determination coefficient (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - x_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
Relative Standard Error (RSD)	$SD = \sqrt{\frac{1}{N-2}\left[\sum_{i=1}^{N}(y_i - \bar{y})^2 - \frac{[\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right]}, RS = \frac{SD}{\bar{x}} * 100\%$

Competitive forecasting algorithms. ML LSTM recurrent neural network [3] and XG BOOST [4] decision trees approaches with different parameters were chosen for forecasting accuracy comparison. Examples of forecasts are shown on (Fig. 8, 9, 10)

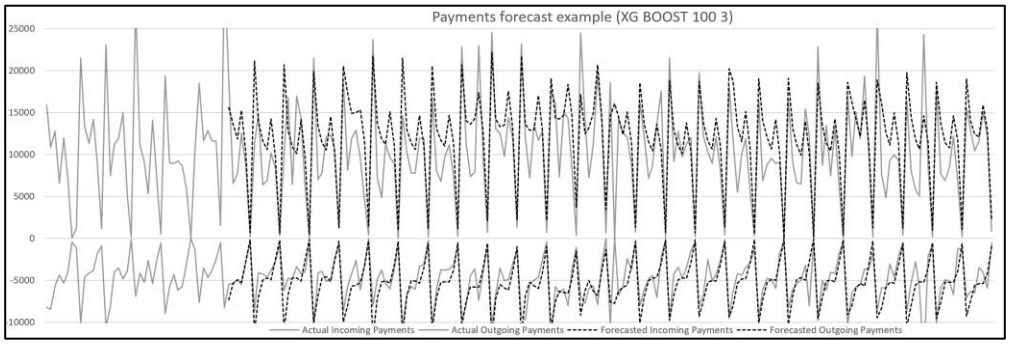


Fig. 8. An example of 180 days of payments forecast by XG BOOST (100, 3). Incoming payments accuracy: MAPE = 144.55%, RMSD = 39.22%, R2 = 0.56, RSD = 29.87%, Outgoing payments accuracy: MAPE = 74.80%, RMSD = 31.62%, R2 = 0.69, RSD = 25.85%

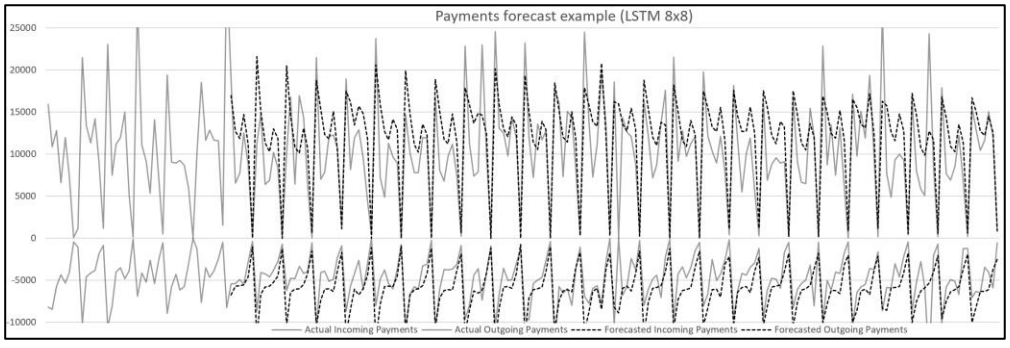


Fig. 9. An example of 180 days of payments forecast by LSTM (8x8). Incoming payments accuracy: MAPE = 81.21%, RMSD = 40.49%, R2 = 0.53, RSD = 31.70%, Outgoing payments accuracy: MAPE = 129.38%, RMSD = 39.77%, R2 = 0.52, RSD = 22.25%

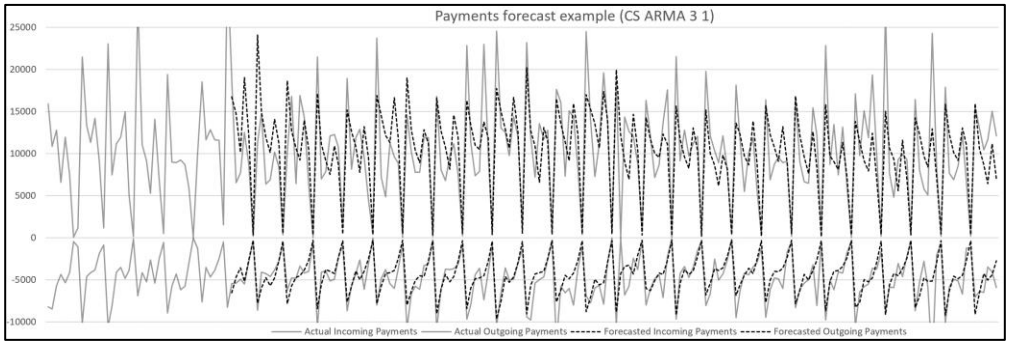


Fig. 10. An example of 180 days of payments forecast by CS ARMA (3, 1). Incoming payments accuracy: MAPE = 74.78%, RMSD = 37.91%, R2 = 0.58, RSD = 31.76%, Outgoing payments accuracy: MAPE = 38.44%, RMSD = 28.32%, R2 = 0.75, RSD = 25.87%

Forecast accuracy calculation results are combined into (Tab.2.)

Tab.2. Payments forecast accuracy estimations of XG BOOST, LSTM and CS ARMA predictors. For each class of predictors and accuracy measure two best values are highlighted with bold font. Cells with absolute best values are highlighted with gray shading.

Method	Description	Incoming payments MAPE (%)	Incoming payments NRMSD (%)	Incoming payments R2	Incoming payments RSD (%)	Outgoing payments MAPE (%)	Outgoing payments NRMSD (%)	Outgoing payments R2	Outgoing payments RSD (%)
XG BOOST (n_estimators, max_depth)	Python xgBoost library, gb.XGBRegressor(learning_rate=0.08, gamma=0, subsample=0.75, colsample_bytree=1), training sample size = 1004, testing sample size = 180, date features: "month, month day, year week, month week, week day, working day"								
(50, 2)		170.58	39.68	0.60	**27.66**	105.68	40.94	0.60	**23.24**
(100, 2)		103.20	**38.43**	**0.63**	**28.37**	**57.13**	**39.35**	**0.64**	**25.53**
(100, 3)		64.47	**38.17**	**0.63**	29.13	**66.73**	**39.08**	**0.65**	25.63
(100, 4)		**60.98**	38.89	**0.61**	30.25	114.63	40.13	0.63	26.32
(100, 5)		62.69	39.75	0.60	31.46	113.51	41.01	0.62	27.16
(250, 3)		**62.51**	39.35	**0.61**	30.81	78.54	40.95	0.62	27.21
(500, 3)		62.64	40.32	0.59	32.23	143.17	42.37	0.59	28.57
(500, 5)		85.52	42.66	0.54	35.65	70.24	43.99	0.57	31.15
LSTM (hidden layers)	Python Keras library, LSTM. Optimizer = "Adam", Loss function = "MSE", Batch size = 20, learning rate = 0.005, training sample size = 1004, testing sample size = 180, date features: "month, month day, year week, month week, week day, working day", training condition: while number of epochs <= 100 and EVS Error <= 0.81								
(4x4)		66.94	37.72	0.65	29.58	185.78	36.86	0.69	26.35
(8)		45.49	**37.57**	**0.65**	**29.11**	62.20	35.84	**0.72**	**26.14**
(8x8)		**44.78**	**37.70**	**0.65**	**29.26**	105.04	**35.70**	**0.71**	26.19
(16)		45.64	37.94	**0.64**	29.35	**46.69**	35.93	**0.71**	26.18
(16x16)		**44.96**	37.82	**0.64**	29.67	226.64	36.05	**0.71**	**26.14**
(32)		46.10	37.87	**0.64**	29.33	49.53	**35.8**	**0.71**	**26.08**
(64)		49.66	38.39	0.63	29.86	**42.56**	36.65	0.70	26.92
(128)		46.53	37.93	**0.64**	29.89	62.95	36.96	0.69	26.58
CS ARMA (p=P, q=Q)	CS ARMA (p, P, q, Q, L=52), training sample size = 1004, testing sample size = 180								
(1, 1)		36.94	**34.44**	**0.71**	**28.91**	47.22	31.86	**0.78**	28.23
(2,1)		**36.92**	34.50	**0.71**	28.99	47.27	31.91	**0.78**	28.29
(2, 2)		38.17	**34.49**	**0.71**	29.11	47.93	**31.44**	**0.78**	**27.73**
(3, 3)		38.68	34.88	**0.70**	29.64	**46.64**	31.95	**0.78**	28.27
(3, 1)		**36.86**	34.52	**0.71**	**29.05**	**47.14**	31.92	**0.78**	28.32
(1, 3)		38.65	34.64	**0.71**	29.30	**46.13**	**31.68**	**0.78**	**27.94**

5. Conclusion

The data sample for this research shows that cash usage shows annually, monthly and weekly distinctive patterns. Therefore, forecasting can be executed with a horizon of months or years. The newly developed Complex Seasonal ARMA has been compared against two existing models: XG BOOST decision trees and LSTM recurrent neural network. For the three algorithms different input parameters have been used and this has been measured

against several reliability output parameters. Especially the R^2 shows that the new CSARMA model is more stable and accurate when applied for this data. It is foreseen that with more historical data the ML algorithms will also detect seasonality better. Since CSARMA was designed based on research and strictly tuned on the specific data structure it is most likely that CSARMA will be less reliable than ML if for some reason seasonality changes. Therefore Giesecke+Devrient performs complete data research and produces forecasting algorithms as part of the software suite that are optimal for a given data structure. Central and Commercial banks can benefit from the high-quality forecasts for improving their expected cash flows based on end customer behaviour. Giesecke+Devrient will further extend the forecasting capabilities with inputs about the fitness levels of cash using data from counting machines inside cash centres.

References

1. Fisher RA. Statistical Methods and Scientific Inference. Ed 2 (rev) Edinburgh, UK: Oliver and Boyd; 1959
2. G E P Box, G M Jenkins and G C Reinsel 2000 Time Series Analysis - Forecasting and Control Prentice Hall New Jersey 1994
3. Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276.
4. Chen, Tianqi; Guestrin, Carlos (2016). "XGBoost: A Scalable Tree Boosting System". In Krishnapuram, Balaji; Shah, Mohak; Smola, Alexander J.; Aggarwal, Charu C.; Shen, Dou; Rastogi, Rajeev (eds.). Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM. pp. 785–794.
5. https://www.nbg.gov.ge/uploads/publications/annualreport/2018/eng_annual_2017.pdf (June 25, 2019)
6. https://www.npr.org/2019/02/11/691334123/swedens-cashless-experiment-is-it-too-much-too-fast?t=1559903488540 (June 7, 2019)

The emergence of legal and organisational arrangements to minimise global tax burden, and its impact on monitoring domestic economic activities[1]

Nadim Ahmad and Peter van de Ven
Organisation for Economic Co-operation and Development (OECD)

Abstract

Multinational enterprises (MNEs) move profits around the globe with, amongst others, the goal of minimizing their global tax burden. The potential to set up legal and organisational constructs for this purpose has grown tremendously with the increasing role of intangible assets in generating value added. Special Purpose Entities, in essence brass plate companies with hardly any physical presence, are being set up to charge fees for the use of intangible assets whose legal ownership has been transferred to these units. Digitalisation has further aggravated these problems to a significant degree. But it should be added that not only intangible assets, also other movable assets, such as transport equipment, lend themselves for setting up constructs, e.g. operational leasing agencies, to benefit from favourable tax conditions in some countries. Moreover, rock bands, football players and other wealthy persons create similar constructs to avoid or to minimize tax payments. The paper will discuss the impact these phenomena have on the accounting of economic activities of a country, as summarised in for example GDP-numbers and balance of payments. It will also propose some alternative ways of dealing with these issues.

Keywords

Multinational enterprises; national accounts; profit shifting; Special Purpose Entities; taxes

1. Introduction

1. Increased globalisation — through increased international trade, capital flows, and the growth of multinational enterprises — is one of the most important developments affecting the world economy during the last 25 years. Globalisation grew rapidly over this period, especially during the 15 years leading up to the 2007–2009 economic and financial crisis. Total world trade in goods and services, for example, increased from 41 percent of world GDP in 1993 to 61 percent in 2008, before dropping during the recession and then afterwards rebounding. The growth of foreign direct investment was perhaps even more dramatic. Global competition and the deepening of global supply chains have transformed many industries and led to profound economic

[1] For an important part, this paper is an excerpt from Moulton and Van de Ven (2018).

changes. Globalisation has been associated with "innovations" in business practices as corporations increasingly manage their production and sales activities at a global level.

2. The focus of this short paper is on economic measurement, and specifically on the impact of globalisation on the national accounts. How do economic activities that are often not constrained by national boundaries interplay with national accounts statistics, which are defined in terms of residence. In particular, one can observe that the effects of globalisation make national data harder to interpret, and, for certain types of analysis, they may even be considered as a distorting influence on the data. Especially the role of multinational enterprises (MNEs) and the growing emergence of global production arrangements make the compilation and interpretation of national accounts numbers less straightforward.

3. MNEs typically move profits around the globe with, amongst others, the goal of minimizing their global tax burden. The potential to set up legal and organisational constructs for this purpose has grown tremendously with the increasing role of intangible assets in generating value added. Special Purpose Entities, in essence brass plate companies with hardly any physical presence, are being set up to charge fees for the use of intangible assets whose legal ownership has been transferred to these units. Digitalisation has further aggravated these problems to a significant degree. But it should be added that not only intangible assets, also other movable assets, such as transport equipment, lend themselves for setting up constructs, e.g. operational leasing agencies, to benefit from favourable tax conditions in some countries. Moreover, rock bands, football players and other wealthy persons create similar constructs to avoid or to minimize tax payments.

4. The above type of arrangements have a direct impact on the allocation of value added across countries, and therefore on the estimation of Gross Domestic Product (GDP) of the countries involved. The paper discusses, in Section 2, in more detail the impact these phenomena have on the accounting of economic activities of a country. In Section 3, some proposals are being put forward to arrive at an alternative way of dealing with these issues, with less distorting consequences for national statistics. Section 4 concludes and provides some points on the way forward. In this paper, the issues surrounding the change in the interpretation of gross trade flows, resulting from global production arrangements which lead to a multiplication of cross-border flows in intermediate goods and services, will not be discussed. For a more detailed discussion of these issues and on-going work to improve the evidence base, reference is made to Ahmad and Ribarsky (2014).

2. The impact of globalisation on the measurement agenda of national accounts

5. On 12 July 2016, the statistical and economic policy community was shocked. The Irish Central Statistics Office published its latest national accounts data for 2015, revealing that real GDP growth was up 26.3% from 2014 (and up 32.4% in current prices). Commentators raised questions about the reliability of the numbers and about the conceptual basis for the measurement of GDP. Some quotes: "Ireland's Economists Left Speechless by 26% Growth Figure" (Bloomberg); "Why GDP growth of 26% a year is mad" (Economist); "It's complete bullshit, it's Alice in Wonderland economics" (Colm McCarthy, University College Dublin). The main reason for the particularly high Irish GDP growth rates lies in the fact that in recent years, attracted in large part by low corporation tax rates, a number of large multinational corporations have relocated their economic activities, and more specifically their underlying intellectual property, to Ireland. As a result, sales (production) generated from the use of intellectual property now contribute to Irish GDP rather than to other countries' GDP. Given the size of these companies, the boost to GDP growth has been correspondingly large.

6. There is ample evidence that the national allocation of value added and profits by MNEs is, for a significant part, driven by minimisation of global tax burden, through mechanisms such as transfer pricing, channelling funds through Special Purpose Entities (SPEs), "optimisation" of the recording of the economic ownership and use of intellectual property products, and the allocation of costs related to corporate services more generally. Lipsey (2010) shows that the ratio of profits to compensation of employees of affiliates that are majority-owned by US MNEs ranges from 0.579 for affiliates in Europe to 11.709 for affiliates in the Other Western Hemisphere.[2] Although it is clear that there is an economic rationale behind all of this, it hampers the analysis from an economic substance point of view, certainly when it comes to analysing national parts of MNEs. Bruner et al. (2018) find a 1.5 percent and 3.5 percent increase in measured U.S. GDP and operating surplus, respectively, if profits of UN MNEs would be allocated proportionally to compensation of employees and domestic sales. De Haan and Haynes (2018) show how, by using the "double Irish with a Dutch sandwich" construction, almost EUR 15 billion of revenues of Google/Alphabet disappear in the Bermuda triangle, and are left unaccounted for in World GDP. Using this construction, revenues are channelled through SPEs in Ireland and the Netherlands, and ultimately end up in a SPE registered in Bermuda. The presence of these SPE-type of units is

[2] Barbados, Bermuda, British Islands and Carribean (British Antilles, British Virgin Islands, Cayman Islands and Montserrat), Western Hemisphere n.e.c. (Anguilla, Antigua and Barbuda, Aruba, Bahamas, Cuba, Dominica, French Islands (Caribbean), Grenada, Haiti, Jamaica, Netherlands Antilles, St. Kitts and Nevis, St. Lucia, St. Vincent and the Grenadines, Trinidad and Tobago, British Islands (Atlantic)).

particularly large in some countries, in Europe for example in Ireland, Luxembourg and the Netherlands. In the latter country, the balance sheet total of SPEs amounted to 600% of GDP at the end of 2016. Also, the related in- and outflows of property income of SPEs in the Netherlands are very substantial, amounting to 20-25% of GDP in the years 2010-2016, while flows of imports and exports of services would add another 3-5%.

7. The background for the impact of the above issues on measurement of national accounts is related to the residency criterion, one of the core definitions or constructs of the international standards for compiling national accounts, the 2008 System of National Accounts (2008 SNA). It delineates the units that are part of the national economy, and, at least indirectly, defines all macro-economic aggregates that can be derived from the system. In § 4.10 of the 2008 SNA, the concept of residence is elaborated as follows: *"The residence of each institutional unit is the economic territory with which it has the strongest connection, in other words, its centre of predominant economic interest."* § 4.14 subsequently defines an institutional unit as having a centre of predominant economic interest *"when there exists, within the economic territory, some location, dwelling, place of production, or other premises on which or from which the unit engages and intends to continue engaging, either indefinitely or over a finite but long period of time, in economic activities and transactions on a significant scale."* For the period of time, one year is taken as a, somewhat arbitrary, operational definition.

8. For corporations and non-profit institutions, the above residency principle means that enterprises have a centre of economic interest in the country in which they are legally constituted and registered. MNEs obviously have centres of economic interest in quite a few countries. Even in the case in which a legal entity is not created, a unit without separate legal status that engages in substantial economic activities is considered a resident institutional unit. Furthermore, in § 4.55 – 4.67, the 2008 SNA addresses the residency of special purpose entities (SPEs), which are defined as having no employees and no non-financial assets; having little physical presence beyond a "brass plate"; always related to another corporation; and often resident in a country other than the country of residence of the related corporation. Although such legal units would normally not qualify as separate institutional units because they may not perform any activities of economic substance and would be consolidated with the related corporation, they are treated by convention as separate units, if they are resident on the economic territory of another country.

9. For MNEs, the above residency principles mean that the activities of each group of units belonging to an MNE that are located on the economic territory of a certain country are to be recorded as part of national economy of that country. This even holds in the case that the relevant unit, or group of units, has physical presence but no separate legal status (e.g. branches), only

performs ancillary activities for the corporation at large, as well as in the case of an SPE with legal status but hardly any physical presence.

10. Another complication in the recording of cross-border transactions of MNEs, and consequently also in the allocation of economies activities to national economies, concerns the application of the principle of economic ownership. In national accounts, transactions between units are based on the principle of change in economic ownership. In § 3.26 of the 2008 SNA, economic ownership is defined as follows: *"The economic owner of entities such as goods and services, natural resources, financial assets and liabilities is the institutional unit entitled to claim the benefits associated with the use of the entity in question in the course of an economic activity by virtue of accepting the associated risks."* The change in economic ownership depends, of course, on the delineation of institutional units in the SNA. An institutional unit, the unit for recording and classifying units in national accounts, is defined as a unit that is capable of owning assets, incurring liabilities, and engaging in economic activities and in transactions with other entities. It is also able to make economic decisions for which it is itself held to be directly responsible and legally accountable. The institutional unit generally consists of a legal unit or a limited group of legal units. Enterprise groups, in which a parent corporation controls several subsidiaries, are not to be considered as a single institutional unit (see § 4.51 – 4.52 of the 2008 SNA). A change in economic ownership would therefore typically coincide with a financial transaction between two institutional units and would therefore coincide with a change in legal ownership. But there are clear exceptions to this rule.

11. The principle of economic ownership is not necessarily straightforward within MNEs. All affiliates of an enterprise group are to some degree controlled by their parent, whereby the case of multinational enterprise groups has the added complication of having non-autonomous affiliates which are considered as institutional units by convention, simply because they are resident in an economic territory that is different from the parent's. Transactions between units of an MNE, or the absence of such transactions as recorded in business accounts, may therefore be at odds with the principle of economic ownership. On the other hand, in practice there may be no alternative to following business accounting, unless one performs a detailed and resource-demanding analysis of individual transactions of the relevant enterprise groups.

12. To add yet another layer of complexity, as noted before, modern economies are more and more knowledge based, in that the competitive edge of an enterprise and a country is often driven by the ownership on intellectual property products (IPPs), the use of which is neither physically nor locally constrained. Determining the economic ownership of IPPs, and therefore the allocation of the output and the use of these assets, is already challenging in a more traditional environment of MNEs owning a group of affiliated entities

producing goods and services, including corporate or ancillary services. But it gets even more complicated in a world where MNEs set up complex structures to allocate the receipts from IPPs and the payments for using them in the preferred way. For example, SPEs are being established in certain countries to reallocate the collection and distribution of royalties, license fees, or profits more generally, with the purpose of avoiding or minimizing worldwide tax payments. Countries with low tax rates, or providing the opportunity of using certain fiscal loopholes, are very attractive for the establishment of such conduits. The use of these conduits often gets front-page news coverage once they become publicly known and relate to well-known MNEs. However, it also has become less obvious to exactly pin down the economic ownership of these intangible assets. For more details on the determination of economic ownership, reference is made to UNECE (2015).

3. A proposal for an alternative way of recording

13. As explained in the above, the current international standards for national accounts clearly can have a significant impact on the allocation of output, value added (GDP) and profits across countries. The main discomfort with the current international standards is related to the fact that the allocation of multinational activities to national economies is not governed by economic substance, but that legal considerations related to minimisation of the global tax burden directly affect the macro-economic statistics including the indicators derived from them. The main problems are caused by transfer pricing and by the international allocation of IPPs and related income, with or without the involvement of SPEs.

14. The first issue to address concerns the treatment of the SPEs. It is apparent from the start that these SPEs are only considered as separate institutional units because they are resident in an economy different from their parents and/or affiliates. Were this not the case, they would be consolidated into the rest of the MNE. Similarly, assigning economic ownership of IPPs to these brass plate companies is a matter of practicality or legality, not a way to approximating economic substance. Therefore, as also proposed by Rassier (2017), a first suggestion to be considered in the future international standards for compiling national accounts is the consolidation of SPEs with their ultimate owners. Consequently, all returns, outlays, financial stocks, and positions of these SPEs would directly end up in the accounts of the country where the headquarters of the multinational are located.

15. The second problem of allocating output and value added of MNEs to national economies more generally concerns the allocation of IPPs. These assets, including the income generated through the use of them, are neither physically nor locally constrained, it is relatively easy to relocate them across countries. Instead of following the actual money flows that are primarily governed by tax considerations, allocating the IPPs and related income to the

country of the ultimate parent would be a logical alternative. Only the part of value added with a physical presence, i.e. compensation of employees and depreciation of non-IPP assets, possibly including a return on the investment, would then remain in the countries in which the affiliates are located. Leaving apart the relocation of IPPs and related depreciation, in a sense this treatment would come down to an "upward shift" of distributed and reinvested earnings from GNI to GDP.

16. From a conceptual point of view, the above treatment would also be quite justifiable. IPPs are quite different from other types of fixed assets used in the production of goods and services. Apart from having no physical and local constraints, IPPs often concern the whole value chain, not a particular part of the production process. Often, there is also no direct link to the production process, and also no direct link between today's stock of assets and today's production of goods and services. IPPs often concerns results from R&D, design, trademarks, etc., and once implemented they are easily scalable. As such, there are very good reasons to consider IPPs as corporate assets, of which the ultimate parent is indeed the true economic owner.

17. One complication may relate to the determination of the residence of the ultimate parent. This does not necessarily coincide with the country in which the formal holding of the MNE is located. Here, one can also observe the phenomenon of corporate inversions, by setting up a holding type of SPE to minimise tax burden. Having such a legalistic approach to the residence of a parent, in combination with the above proposals for the treatment of SPEs and IPPs would potentially lead to far more dramatic shifts of output and value added from one country to another. Instead, one will have to determine the "true" residence of the parent, on the basis of the location of the centre of economic decisions. This centre would typically coincide with the location from where decisions are made on e.g. global arrangements of production, R&D and other corporate investments, corporate finance, appointment at senior management level, etc. It would therefore usually be the same location as the one where the physical headquarters and the board of directors are located.[3]

4. Discussion and Conclusion

18. It is clear that all of the above options require, at least to some extent, the exchange of individual data on MNEs across countries. In the current legal circumstances, this is a major issue that would need to be resolved rather urgently. Two possible ways forward can be distinguished: (i) a top-down approach according to which data on MNEs are collected at the international

[3] In some cases, two countries may share the headquarters of an MNE. Good examples are Royal Dutch Shell and Unilever, which have part of their headquarters in the Netherlands and the United Kingdom. In such cases, one may need to apply some proportionality rule.

level, with a subsequent provision of data on the national parts of the MNEs to the countries; and (ii) a bottom-up approach according to which each country collects data on the national parts of the MNEs, which are subsequently exchanged and verified across countries. Given current circumstances, both ways forward require a paradigm shift in allowing for international exchange of individual data within the statistical community and, in the case of the top-down approach, in collecting statistical data and compiling national accounts data. One possible step would be the re-use of the data that will become available from the OECD Base Erosion and Profit Shifting (BEPS) initiative. See http://www.oecd.org/ctp/aggressive/beps-2015-final-reports.htm. Action 13 of this initiative requires MNEs to provide much more detailed country-by-country reporting on their worldwide business, with more detailed information requested for large MNEs. But it remains unclear whether this data becomes available for statistical purposes as well.

19. In this paper, options for an alternative recording of MNE-activities have been put forward. Although one might wish for a world in which national accounts fit into traditional narratives about domestic production using capital and labour, the way forward involves recognising that the world has become more complex. While one can supplement the current international standards to help in the analysis of a globalised economy, in the long term it is necessary to adapt our data collection and compilation strategies to come up with innovative ways to measure global entities that are engaged in production that is not limited by national boundaries.

References

1. N. Ahmad and J. Ribarsky. 2014. Trade in Value Added, Jobs and Investment, Paper Prepared for the 33rd General Conference of the IARIW (Rotterdam, August 24-30, 2014). Available at: http://www.iariw.org/papers/2014/AhmadPaper.pdf
2. J. Bruner, D.G. Rassier, and K.J. Ruhl. 2018. Multinational Profit Shifting and Measures throughout Economic Accounts, NBER Working Paper No. 24915. Cambridge (USA). Available at: https://www.nber.org/papers/w24915.pdf.
3. M. de Haan, and J. Haynes. 2018. R&D Capitalisation: Where Did We Go Wrong?, EURONA 1/2018. Luxembourg. Available at: https://ec.europa.eu/eurostat/cros/system/files/euronaissue1-2018-article1.pdf.
4. R. Lipsey. 2010. Measuring the Location of Production in a World of Intangible Productive Assets, FDI and Intrafirm Trade, The Review of Income and Wealth, Volume 56: S99-S110. DOI: 10.1111/j.1475-4991.2010.00385.x.
5. B.R. Moulton, and P. van de Ven. 2018. Addressing the Challenges of Globalization in National Accounts, Paper prepared for CRIW Conference

on "The Challenges of Globalization in the Measurement of National Accounts" (Washington, D.C., March 9 – 10, 2018).nAvailable at: http://papers.nber.org/conf_papers/f100570.pdf

6. D.G. Rassier. 2017. Improving the SNA Treatment of Multinational Enterprises, Review of Income and Wealth, 63(s1), (December), s287-s320. Available at: http://onlinelibrary.wiley.com/doi/10.1111/roiw.12323/abstract.
7. UNECE, Eurostat, and OECD. 2011. The Impact of Globalization on National Accounts. New York, Geneva. Available at: https://www.unece.org/fileadmin/DAM/stats/publications/Guide_on_Impact_of_globalization_on_nat ional_accounts__web_.pdf.
8. UNECE. 2015. Guide to Measuring Global Production. New York, Geneva. Available at: https://www.unece.org/publication/guide_mgp.html.
9. United Nations, European Commission, International Monetary Fund, Organisation for Economic Cooperation and Development, and World Bank. 2009. System of National Accounts 2008. New York. Available at: https://unstats.un.org/unsd/nationalaccount/docs/sna2008.pdf.

Tracking the international footprints of global firms

Mary Everett[1], Stefan Avdjiev[2], Hyun Song Shin[3], Philip Lane[4]
[1]Central Bank of Ireland
[2]Bank for International Settlements
[3]Bank for International Settlements
[4]Central Bank of Ireland

Abstract
As the global economy becomes more integrated, there is a growing tension between the nature of economic activity and the measurement system that attempts to keep up with it. Many policies are still determined by measuring economic activity at the national level. Since the typical unit of analysis is the economic area (the "island"), economic activity is measured within the island and in terms of transactions between islands. But, increasingly, companies and their ownership are global, with economic activity taking place in a geographically dispersed way. We analyse several important issues created by this tension, show how they manifest themselves in the data and draw lessons from them.

Keywords
International capital flows; financial globalisation; current account

1. Introduction

Our existing measurement framework for economic activity in national accounts and the balance of payments is based on an "islands" view of the global economy. Taking the economic area (the "island") as their unit of survey, analysts measure economic activity within the island and the transactions between islands.[1] In the simplest case, the workers, production processes, headquarters, management and owners of firms are all located in the same economic area, typically defined by a national boundary. The key concept in national accounts is that of *residence*. National accounts convey information on the activities of residents on the island. In simple cases, residence is clear-cut. For a firm producing goods in a plant located on a single island, employing workers from the same island and owned by residents on the island, the notion of "residence" for the firm is straightforward. It coincides with the physical location of the firm on the island. If such a firm exports goods, then the goods will cross the boundary of the island into another island. Thus, exports will

[1] The national accounting framework comprises the suite of macroeconomic and financial statistics on the evolution of flows and stocks for an economy and its institutional sectors, as well as their interactions with residents and non-residents. For an introduction to the system of national accounts see Lequiller and Blades (2014).

show up in the customs data for the island. However, "residence" is a legal concept denoting the relationship between an entity and a location. For a person, travelling through another country does not make the person a resident of that country. For a firm, residence is defined as "the economic territory with which it has the strongest connection, expressed as its centre of predominant economic interest".[2] But a firm resident on island A can operate elsewhere. For example, it could enter into a contract manufacturing agreement with a firm in island B, and sell the output in island C. The good is shipped from B to C, and never touches the shores of island A. The sale would nevertheless be counted as an export of island A and would enter its trade and GDP statistics. Island A's GDP would go up even if no workers are employed on the island.

Closely related to the notion of residence is that of domicile, which indicates greater permanence. For a person, domicile is a legal concept similar to residence, but which carries additional implications as a place of origin and permanent place of residence. For firms, the term is often used to denote the location of the headquarters. However, there are far-reaching implications from the designation of a particular location as its domicile, as the firm's relationship with its subsidiaries, branches, offices and subcontractors all make reference to the domicile. When a firm moves its domicile, a cascade of other changes follow. The firm's place in the world undergoes fundamental alterations, as its relationship with other jurisdictions is rearranged. The redomiciling of a firm is not just a relabeling but involves a long list of changes in bilateral relationships between jurisdictions that flow from the alteration in domicile.

In a global context, we can think of the above two perspectives, the residency view and the domicile view, as two distinct but integrated frameworks from an accounting, statistical, legal and regulatory angle. In the international statistical framework, the islands view allocates economic agents to the country in which they are deemed to reside. An alternative approach is to take a consolidated view, which assigns economic entities to the country of headquarters of the parent institution (Avdjiev et al (2016), Bénétrix et al (2017), McCauley et al (2017)). This latter approach is, therefore, more closely aligned with notion of domicile.[3] In a consolidated framework, the entire

[2] IMF (2009, p 70). According to the international statistical framework, residency is expressed as the centre of predominant economic interest. Each institutional unit is a resident of only one economic territory. An institutional unit is defined as households, corporations, non-profit institutions, government units, legal or social entities recognised by law or society or other entities that may own or control them.

[3] Strictly speaking, there are several different ways to consolidate group level information, depending on whether one assumes a supervisory, statistical or business accounting point of view (IAG (2015)).

corporate group is assigned to the country of headquarters, no matter where its constituent operating units may reside.[4]

Since the national accounting framework was developed in the 1930s and 1940s, the activities of global firms and the structure of the global economy have undergone profound changes. Balance of payments accounting has adapted to changes in economic reality, with the latest standard being the sixth edition of the IMF's balance of payments manual (known as BPM6), published in 2009 (IMF (2009)). However, the pace of globalisation has arguably outstripped the pace of innovation in the measurement rules, increasing the tension between the nature of economic activity and the measurement system that strives to keep up with it. Increasingly, companies are global, as is their ownership, with economic activity taking place in a geographically dispersed way. Understanding the impact of macroeconomic developments, financial price movements or public policies on corporate decisions requires the rearrangement of institutional units dispersed across the world into corporate groups on the basis of ownership and control. And yet measurement is still largely residence-based, classifying institutional units by attributing a location of "predominant economic interest" to each entity.[5]

As corporate activity increasingly straddles national borders, it takes place through many separate legal entities that together span the globe. A manufacturing operation and its workers can be sited far from the headquarters of the firm, and far from its other operations, such as marketing, sales, or research and development. Ownership is also global, since the investors of a listed firm are spread around the world. The jurisdiction in which a company is headquartered (its domicile) may reflect the firm's origin and history, or simply tax or corporate governance considerations. Domicile applies to a firm's assets, which need not be only physical capital but can include intellectual property used to create value.

In this article, we go over a number of the key issues raised by the tension between the traditional residence-based measurement system and the evolving nature of globalisation. In many instances, the consolidated approach has the potential to provide a useful alternative perspective. That said, given the increasingly complex nature of the global economy, there are no straightforward ways to comprehensively address many important economic questions using a single measurement framework. Instead, one needs to extract information from multiple frameworks, using an approach tailored to the specific question at hand.

[4] The country in which economic decisions are taken may be different from both the country of residence and the country of headquarters.

[5] There are several data sets that represent notable exceptions to the above pattern. We discuss those in the last section of this article.

The remainder of the article is organised as follows. In the next section, we discuss several important measurement issues associated with the way in which the activities of global firms are recorded under current national accounting rules. Next, we investigate how some of those issues manifest themselves in the data. We conclude by drawing lessons from the above issues.

2. Methodology: National accounts and global firms - measurement issues

As multinational firms, with their complex corporate structures, distribute their activities across traditional borders, they complicate the task of capturing economic activity within traditional national accounts (Tissot (2016)). A growing body of evidence suggests that the activities of global firms have outgrown some features of the existing national accounting framework.[6]

It is now well understood that net concepts such as the current account do not adequately reveal the underlying linkages across countries, which are likely to reflect gross flows to and from different national sectors. As a consequence of their growing size and complexity, gross capital flows increasingly affect the current account through their impact on primary income. That is why it is necessary to analyse the composition of both gross and net flows, by functional component and sector, even within the confines of the existing residence-based accounting framework (Lane (2013)).

In the context of international banking flows, this has already been well documented in the existing literature. For example, several authors have argued that current account balances did not reveal underlying vulnerabilities created by European banks' large-scale reinvestment of funds raised from US money market funds into US mortgage-backed securities before the 2007–09 Great Financial Crisis (GFC).[7]

In this section, we provide three hypothetical examples to illustrate some consequences of globalisation. We start with the "classical" measurement issues associated with global firms, illustrating how offshoring affects national accounts. Second, we highlight additional conceptual and measurement challenges associated with the redomiciling of global firms – that is, the change of legal domicile of a firm to another location. Third, we describe issues raised by the cross-border mobility of corporate assets, in particular intangible assets such as intellectual property. While we provide three separate examples for simplicity, these phenomena can interact in practice, further complicating the interpretation of balance of payments data.

[6] Lane (2015, 2017), Forbes et al (2017), Guvenen et al (2017).

[7] Obstfeld (2012), Borio and Disyatat (2011), Lane (2013), Shin (2012).

3. Result: Quantitative importance

The post-World War II increase in global external financial openness accelerated sharply between the mid-1990s and the GFC.[8] Spurred by financial liberalisation and innovation, external assets and liabilities surged from a combined total of less than 150% of GDP in 1995 to over 400% in 2007 (Lane and Milesi-Ferretti (2018)). The GFC seemingly brought to a halt the rapid rise in external financial openness, with the global stock of external assets and liabilities contracting to slightly under 400% of GDP in 2015 (BIS (2017b)).

Given the expansion in gross assets and liabilities, a focus on the trade balance when measuring external imbalances ignores the dynamics of international trade in financial assets (Lane (2015), Lane and Milesi-Ferretti (2018), Forbes et al (2017)). The importance of gross primary income flows (relative to gross trade flows) rose steadily between the mid-1990s and the GFC. This largely reflected the rapid pre-crisis expansion of the stocks of cross-border financial assets and liabilities (discussed above). This trend was most pronounced for financial centres (FCs), where the ratio of gross primary income flows to gross trade flows more than quadrupled from 14% in 1995 to 65% in 2007. The relative importance of primary income flows also rose considerably for (non-FC) advanced economies (AEs) – from 12% in 1995 to 23% in 2007.

The post-GFC pullback in gross external financial positions (together with the low interest rate environment) reversed this trend, but only partially. The 2015 level of the ratio of gross primary income flows to gross trade flows was still roughly three times that in 1995 for FCs. By contrast, the respective ratio remained relatively flat for EMEs both before and after the GFC. Delving deeper into the main components of primary income flows reveals that the relative importance of direct investment income has increased sharply since the 1990s. This is the case not only at the global level, but also for all major country groups. The rise is especially notable in the case of FCs. The increase was primarily driven by the growth of offshoring.

The increased relative importance of direct investment income (DII) flows suggests that global firms' foreign profits merit special attention. DII reduces the current account of the country in which it is generated and boosts those of (i) the country in which the company is headquartered and (ii) the countries in which the company's shareholders reside. In economic terms, all the benefits (abstracting from labour income) accrue to the countries in which the shareholders reside. In accounting terms, however, the positive current account impact is split between the countries in groups (i) and (ii) above because of the asymmetrical treatment of DII relative to portfolio investment

[8] The post-World War II increase in external financial openness was a part of the second major wave of globalisation. The first major globalisation wave, which lasted from the early 1800s to World War I, also saw a substantial increase in both real and financial cross-border linkages.

income (PII) in the existing balance of payments framework. The split depends on the proportion of DII which is not distributed to shareholders. Such "undistributed profits" could take the form of either DII reinvested earnings (DII_RE) or dividends paid from an affiliate to a parent, which are added to the corporate cash pile rather than paid out to shareholders.[9]

NFCs are also important providers of funding to banks. As illustrated by Aldasoro et al (2017), banks outside the United States have reported considerable increases in their US dollar-denominated deposits from non-banks since the GFC. This source of funding has more than offset the run-off of eurodollar deposits by US money market funds that took place in 2016 (BIS (2017a)). The recently enhanced counterparty sector dimension of the BIS locational banking statistics reveals that NFCs' deposits in BIS reporting banks have grown by nearly 20% between end-March 2015 and end-September 2017. Thus, the undistributed profits of NFCs have contributed to keeping global liquidity conditions relatively loose, despite a number of factors pulling in the opposite direction.

4. Discussion and Conclusion:

Given the measurement issues discussed in this special feature, policymakers should exercise caution when using rules of thumb developed for a bygone era. For example, debt/GDP and credit/GDP ratios may not be good measures of financial system vulnerabilities for some countries, as the denominator does not adequately capture the size of the domestic economy. Unfortunately, the current national accounting framework creates obstacles to the accurate interpretation of key economic indicators by stakeholders including the official sector, financial market participants and researchers. The complexity of global firms indicates that additional measures are necessary. Such measures should augment the traditional national accounting framework by looking through the "islands" with the ultimate goal of creating consolidated national accounts.

A number of data initiatives now under way point to progress in addressing these problems, as recognised in the G20 Data Gaps Initiative. These include the Legal Entity Identifier initiative to identify distinct legal entities and link them to the ultimate parent group;[10] the various data sets collected by the BIS on a consolidated basis – eg the consolidated banking statistics, the G-SIB data hub collection and the international debt securities; Ireland's concept of

[9] These undistributed profits should be captured in the financial account of the balance of payments under the reinvested earnings component of direct investment and affect the dynamics of the net international investment position (since, all else equal, the value of portfolio equity liabilities should rise in proportion to the scale of retained earnings). The net international investment position, however, receives less attention than the traditional current account balance.

[10] www.leiroc.org/

GNI*, which strips the depreciation of foreign-owned capital assets from the measurement of domestic income; and the foreign affiliate trade statistics.

The residence-based and the consolidated accounting frameworks should be considered complementary rather than mutually exclusive. The consolidated accounting framework, while newer and more suited to addressing some of the measurement issues discussed here, is not unconditionally superior to the residence based framework along all dimensions. Instead, its real benefit would be in providing a useful supplementary perspective, whose relevance would naturally depend on the question that is under investigation.

References

1. Aldasoro, I, T Ehlers, E Eren and R McCauley (2017): "Non-US banks' global dollar funding grows despite US money market reform", BIS Quarterly Review, March, pp 22–3.
2. Avdjiev, S, R McCauley and H S Shin (2016): "Breaking free of the triple coincidence in international finance", Economic Policy, vol 31, no 87, pp 409–51.
3. Bank for International Settlements (2017a): 87th Annual Report, June, Chapter V.
4. ——— (2017b): 87th Annual Report, June, Chapter VI.
5. Bénétrix, A, R McCauley, P McGuire and G von Peter (2017): "The consolidated wealth of nations: a first step", mimeo, February.
6. Beusch, E, B Döbeli, A Fischer and P Yeşin (2017): "Merchanting and current account balances", The World Economy, vol 40, no 1, pp 140–67.
7. Borio, C and P Disyatat, (2011): "Global imbalances and the global crisis: link or no link?", BIS Working Papers, no 346, May.
8. Borio, C, H James and H S Shin (2014): "The international monetary and financial system: A capital account historical perspective", BIS Working Papers, no 457, August.
9. Central Statistics Office (2015): "Redomiciled PLCs in the Irish Balance of Payments".
10. ——— (2016): "Explaining Ireland's FDI asymmetry with the United States".
11. European Commission, International Monetary Fund, Organisation for Economic Cooperation and Development, United Nations and World Bank (2009): "System of National Accounts 2008", Sales, no E 08 XVII 29.
12. Everett, M (2012): "The statistical implications of multinational companies' corporate structures", Central Bank of Ireland, Quarterly Bulletin, no 2, Box 3.
13. Fitzgerald, J (2013): "The effect of re-domiciled PLCs on Irish output measures and the balance of payments", Quarterly Economic

Commentary Research Note, Economic and Social Research Institute, no 2013/1/2.
14. ——— (2015): "Problems interpreting the national accounts in a globalised economy – Ireland", Quarterly Economic Commentary Special Article, Economic and Social Research Institute, June.
15. Forbes, K, I Hjortsoe and T Nenova (2017): "Current account deficits during heightened risk: Menacing or mitigating?", Economic Journal, vol 127, no 601, pp 571–623.
16. Gravelle, J and D Marples (2014): "Corporate expatriation, inversions, and mergers: tax issues", Congressional Research Service Report prepared for the Members and Committees of Congress, 27 May.
17. Guvenen, F, R Mataloni, D Rassier and K Ruhl (2017): "Offshore profit shifting and domestic productivity measurement", *NBER Working Papers*, no 233324, April.
18. Inter-Agency Group (IAG) on Economic and Financial Statistics (2015): *Consolidation and corporate groups: an overview of methodological and practical issues*, IAG reference document, October.
19. International Monetary Fund (2009): *Balance of payments and international investment position manual.*
20. ——— (2015): "Revisiting global asymmetries – think globally, act bilaterally".
21. Jordan, T (2017): "High Swiss current account surplus: consequences for SNB monetary policy?", speech at University of Basel, Faculty of Business and Economics, November.
22. Lane, P (2013): "Capital flows in the euro area", *European Economy Economic Paper*, no 497, April.
23. ——— (2015): "A financial perspective on the UK current account deficit", *National Institute Economic Review*, National Institute of Economic and Social Research, vol 234, no 1, pp 67–72, November.
24. ——— (2017): "Notes on the treatment of global firms in national accounts", *Economic Letter*, no 1, Central Bank of Ireland.
25. Lane, P and G Milesi-Ferretti (2017): "International financial integration in the aftermath of the global financial crisis", *IMF Working Papers*, no WP/17/115.
26. Lequiller, F and D Blades (2014): *Understanding national accounts: Second edition*, OECD Publishing, Paris.
27. McCauley, R, A Bénétrix, P McGuire and G von Peter (2017), "Financial deglobalisation in banking?", *BIS Working Papers*, no 650, June.
28. Obstfeld, M (2012): "Does the current account still matter?", *American Economic Review*, vol 102, no 3, pp 1–3.
29. Office of National Statistics (2016): "An analysis of the drivers behind the fall in direct investment earnings and their impact on the UK's current account deficit".

30. Organisation for Economic Cooperation and Development (2008): "OECD benchmark definition of foreign direct investment: Fourth edition".
31. Setser, B (2017): "Dark matter. Soon to be revealed?", Council on Foreign Relations blog, *Follow the Money*, 2 February.
32. ——— (2018): "The impact of tax arbitrage on the U.S. balance of payments", Council on Foreign Relations blog, *Follow the Money*, 9 February.
33. Shin, H S (2012): "Global banking glut and loan risk premium", *IMF Economic Review*, vol 60, no 3, pp 155–192.
34. Stapel-Weber, S and J Verrinder (2016): "Globalisation at work in statistics—Questions arising from the 'Irish case'", *EURONA: Eurostat Review on National Accounts and Macroeconomic Indicators*, no 2(2016), pp 45–72.
35. Tissot, B (2016): "Globalisation and financial stability risks: is the residency-based approach of the national accounts old-fashioned?", *BIS Working Papers*, no 587, October.

Mobile phone and credit card data: Experience from 10 years of public private partnership

Jaanus Kroon
Eesti Pank / Bank of Estonia, Tallinn, Estonia

Abstract

Since 2008, Eesti Pank, the central bank of Estonia, has been cooperating with private companies to source big data for official statistics. The central bank uses mobile phone data from Estonian mobile network operators to quantify inbound and outbound travel, and credit card payment data to calibrate expenditure figures. The combination of these two datasets and several other sets of reference data allows timely and efficient production of trade in services statistics, used in the compilation of the balance of payments and the national accounts. The cooperation has stood the test of time as part of the statistical business process at the central bank. After a full decade of cooperation between official statistics and big data providers in Estonia, we look back at the experience gained and the lessons learned:

- the evolution of mobile phone data and credit card data over the years
- the ways both types of data can be used for official statistics
- lessons from the partnership

This best practice can be used by statisticians in accessing and adopting data sources to produce official statistics in their own countries.

Keywords

big data; mobile positioning; central bank statistics; balance of payments; cooperation for official statistics.

1. Introduction

Globalisation, the blurring of borders and the complexity of measuring cross-border transactions have posed a considerable challenge for external statisticians for some time now. The rapid growth in worldwide travel, membership of the Schengen Area, where there are no border controls and so no data collection at borders, and the discontinuation of regular border surveys by Statistics Estonia because of budget cuts forced Eesti Pank as the institution responsible for external sector statistics to find an alternative way to continue the border-crossing time series. Border-crossing data are an important input in the compilation of the country's monthly and quarterly balance of payments, where exports and imports of travel services play a major role. Many alternative options were explored to meet the demand for a high quality and efficient data source at a reasonable cost and with low labour

intensity. The table 1 briefly describes these alternatives and lists the advantages and disadvantages of each, given the situation at that time.

Eesti Pank opted for mobile positioning as it was the simplest statistics instrument and relatively low-cost. The choice was largely determined by the availability of the potential partner, the Department of Geography of the University of Tartu in Estonia, whose spin-off company Positium OÜ had had regular experience in using mobile positioning data in urban and regional geography and planning since 2001. Building on the regular data exchange that was already established with the biggest mobile operator in Estonia and on the availability of related calibration surveys, remarkable scale effects were expected. In 2008, the central bank started working with researchers at the University of

Table 1. List of possible alternatives

Alternatives	*Pros*	*Cons*
Taking over the border- survey from Statistics Estonia or financing it	-Two years of experience, routine - Can be partly integrated with visitor motivation interview surveys	-Time and labour intensive -Disproportionately expensive -Insufficient reliability: coverage, sampling, grossing-up, etc.
Setting up an accommodation statistics-based assessment model	-Monthly frequency -Easy to run -Reliable geographical allocation -Low costs	-Does not cover outbound tourists -Additional costs for a regular calibration survey (private vs. hotelMstays; visitors vs. tourists, etc.) -Estimation errors
Using a Road Office sensor data-based assessment model (car- counters on the road at border-crossings)	-High periodicity -Supplements existing harbour and airport business data on border- crossings	-Very limited coverage -Additional costs for a calibration survey (geo allocation, number of passengers, length of stay)

		-Estimation errors and quality issues
Using credit and debit card data from Northern European Transaction Services (NETS Estonia)	-High periodicity -Gives estimation on total expenditures and indirect geo allocations -Coverage (only one service provider in Estonia) -No administrative burden -Low costs for compiler -Long experience with payments statistics	-Additional costs for a calibration survey (card vs. cash, expenses and visitors on "behalf of family", etc.) -Negative example from neighbouring countries -"Noise" related to e-commerce
Introducing the methodology based on mobile network roaming information	-High periodicity -Representativeness (almost everyone has a mobile phone) -Operational information in time and space (including geographical allocation) -No major administrative burden	-Lack of experience and practice -Undefined cooperation model with data providers -Additional costs for the calibration survey of mobile usage patterns -Substantial IT resources needed for data processing

Tartu and Positium OÜ to develop the new data collection methodology and models. Methods for inbound travel were fixed in 2008-2009, and those for outbound travel in 2009-2010. The complete methodology was revised and

updated in 2015. Mobile owners are a representative large sample whose spatial behaviour and characteristics in time can be extended to the entire population. The data are readily available, which makes data collection faster and more cost-effective. As mobile phones are widely used, the resulting data set is comprehensive; it minimises the human factor affecting interviewer interpretation in surveys and ensures homogeneity. So, the data are more accurate and of better quality than those collected by traditional data collection methods. Although each transaction with a payment card would give a location-based fact similar to the roaming transaction of a mobile owner, ten years ago it was obvious that the facts from card data were not as numerous as those detected by mobile positioning. That is why it was decided to continue payment card data collection on an aggregated basis for estimating travel expenditures and to use it as one of the main alternative data sources that can help to validate the border-crossing aggregates derived from the mobile positioning data (MPD).

2. Mobile positioning-based statistics of border crossing: methodological aspects

The jointly developed methodology is based on the use of readily available log files from Mobile Network Operators (MNOs), registering the information needed for billing incoming and outgoing roaming activities like voice calls, SMSs and MMSs, mobile-data usage, and mobile supported GPS usage. These activity events are called Call Detail Records (CDR). The parameters that the methodology needs for each call activity are:

- SIM card ID, replaced by a randomly generated pseudonymous ID for statistical use;
- date and time;
- antenna ID with location data;
- country ID.

In line with the Balance of Payments (BoP) methodology, mobile positioning determines the residence of a traveller using the permanent residence criterion, regardless of the resident's citizenship or nationality. As a rule people sign a contract with a mobile company in the country where the phone will be used most frequently, and so the phone owners are presumed to reside in the country where their SIM card is registered. This approach is supposed to give even more precise results under the residency concept of BoP statistics. The amount, length and nature of the visits of Estonian residents and non-residents are determined by the location-based anonymised use patterns of mobile phones derived from the roaming activities in the reporting resident operator network, and operator clients' roaming activities in networks abroad. The statistics on inbound and outbound travel reflect both same-day and overnight visits:

- The number of visitors is determined from the encoded number IDs.
- The duration of a visit is determined from the temporal distribution of the call operations of an individual mobile phone. If all the call activities are within a single day, the recorded duration of the visit is one day. If there are call activities on several days, the number of the days with call activities and any 'empty' days in between is assumed to be the number of days of the visit. If there are no call activities for seven consecutive days, the person is assumed to have left the country. For outbound travel, all the visits are compared against the activities in the same period in Estonia, and if necessary the initial visit is split into several visits.

Data processing broadly consists of the following steps, some of which are country-specific:

- Quality control of the data collected from the operator's system. Since there is a huge amount of data, filters have been developed to find and correct errors based on data characteristics.
- Filtering and evaluation of the roaming data in order to ensure representativeness and quality of data.
- Geographical and temporal interpolation, i.e. linking additional parameters to ensure administrative and chronological comparability.
- Elimination of border bias:
  - the registration of the roaming activities by Estonian residents who are incidentally in the coverage area of foreign mobile operators or
  - the registration of the roaming activities of the customers of foreign mobile operators who are incidentally in the coverage area of mobile phone masts located near Estonian borders, such as ships' crews and passengers on passing ships, or residents of the neighbouring countries.
- Elimination of transit travel:
  - For inbound travel, travellers detected in Estonia's main transit corridors if the stay in Estonia is shorter than a certain limit; there are 10 such transit areas, including the Tallinn-Ikla and Tallinn-Narva roads and the Estonian section of the Riga-Pskov road, Tallinn Airport, and the major ports.
  - For outbound travel, trips through foreign countries. Countries visited without an overnight stay, which do not comply with the criteria of a destination country, are classified as transit countries. One of the criteria for determining visiting and transit countries is the distance from Estonia and the length of stay in hours.
- Profiling and segmenting of individual trips for single-day and multiday visitors; the number of visits to and from the country, the country of origin (inbound) and destination (outbound) and the number of nights

and days spent are calculated. Certain exceptions apply to long-term students and border workers:

- Long-term stay of non-residents. Students are expected to have stayed in Estonia for over 183 days during the preceding 12 months. Two mutually independent criteria have been chosen for distinguishing non-resident border workers: either based on the number of visits to Estonia in a year or on the duration of the stay. A visitor is considered to be a nonresident worker when they have stayed in Estonia on seven or more occasions in a year or for more than 91 days during the past 12 month (derived from experience).
- Long-term visits by Estonian residents who work or study abroad, if they stayed in the destination country for over 91 days during the past 12 months.

- Grossing-up to the total population with a special penetration model which takes into account the following:
  - Market shares of the mobile network operators (MNOs) covered
  - Penetration rate of SIM cards in the country with roaming services
  - Calibration survey results for differences in phone usage between residents of different countries, seasonality, and over or under coverage of SIM cards to cover double or travel SIMs
  - Available reference data
  - Estimate using the calibration coefficients matrix

Theoretically, it would be possible to use random sampling of SIM cards instead of the whole raw data-file, which could be done with weights in grossing-up but given the small size of Estonia and the small population, it is considered impractical.

3. Card payment statistics: methodological aspects

Eesti Pank has compiled and disseminated detailed monthly payment statistics since 1998. The statistics are based on aggregation of the detailed input data reported directly by credit and other payment institutions and as a part of the reporting obligation of monetary and financial institutions, following the census principle.

The following data are reported for card payments:

- Date [Month]
- Type of payment [Cash deposit/withdrawal | Point of sales (POS) payments | E-commerce payments, which are card payments in internet-based stores or elsewhere in the online environment]
- Residency of card issuer [ISO country code]
- Country of payment [ISO country code]

- Number of transactions
- Total turnover

All payments with similar identifiers are aggregated in the data model. No granular payment data are currently reported to keep the reporting volumes low. The biggest limitation of using the card data for travel statistics is that the residency of the card issuer is not always a good proxy for the residency of the card holder, i.e. the traveller. Despite this, card payment statistics are a good data source for quantifying inbound and outbound travel and credit card payment data to calibrate expenditure figures. The dynamics of card transaction volumes and turnover correlate strongly with the dynamics of visits. Card expenditures at home and abroad correlate strongly with BoP travel exports and imports. Card payment statistics could be developed further by exploiting other information stored by the card service provider for each card payment, such as Merchant Category Code (MCC), assigned by the acquiring bank when the business applies for a merchant account, and Transaction Category Codes (TCC) groups according to ISO 18245. Such data are readily available and could give additional information needed for the estimation of BoP sub-categories and provide important detail for economic flash forecasts and other users of statistics.

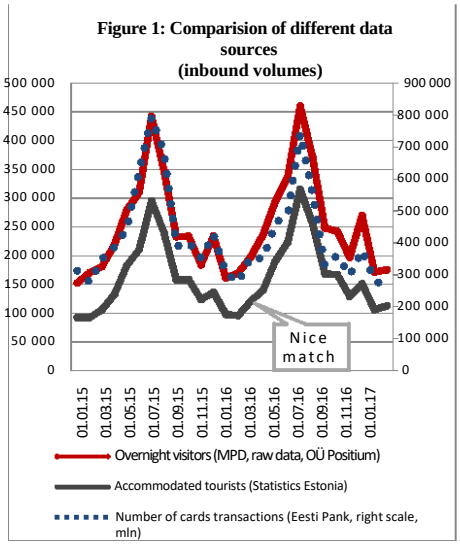
4. Data validation and cooperation model

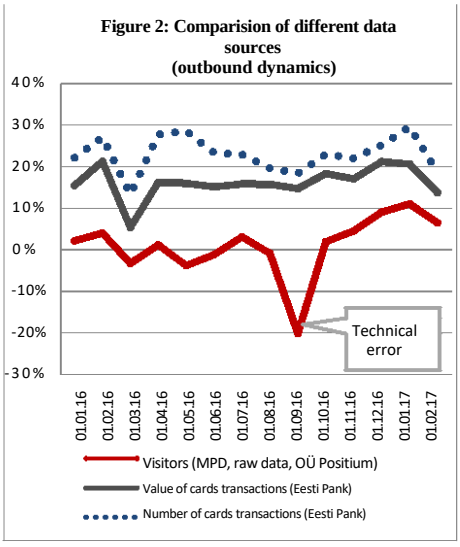
The daily cooperation takes the form of a Public Private Partnership (PPP), outsourcing data processing contracts from Positium OÜ. Three-year contracts have been announced in 2009, 2012, 2015 and 2018. The work has been arranged according to the Generic Statistical Business Process Model (GSBPM) as described in Table 2.

Table 2. Work arrangements

Positium OÜ	Eesti Pank
	- Specifying needs and defining business case
-	
- Design - Build	
- Data collection and processing	
- Calibration surveys - Revising Design and Build	- Data analysing and validation - Data dissemination - Feedback for fine-tuning design and build

Monthly results provided by OÜ Position are verified and validated by Eesti Pank. Along with mobile data, other official data sources are used (e.g. the number of passengers in the Port of Tallinn and Tallinn Airport, crossings of the Estonian/EU administrative border, official accommodation statistics, the press, etc.). The most relevant time series for data validation is provided by card payment statistics (Figures 1 and 2).





In case of doubts about the quality of the original data, an enquiry is addressed to our partner, who carries out his own analysis and, if necessary, forwards the enquiry to the mobile operators. If the data need to be corrected, each specific case will be considered separately and adjusted as appropriate. If it is not possible to identify the specific reasons for the discrepancy, the data will be corrected based on the data from the same period of the previous year, multiplied by a coefficient equal to:

- the average increase/decrease of available payment statistics and other reference data or
- the average change in the last eleven months.

The main inconsistencies in the data relate to a) errors in the source data files provided by MNOs (missing or invalid data); b) changes in MNOs market shares or roaming rates; c) under coverage caused by improved availability of alternative networks (wifi) or d) new sources of noise, incl. overflights. The data are used as an input for monthly and quarterly external sector statistics and have been disseminated as official statistics on Eesti Pank's website on a quarterly basis since 2012. A press-release on this is attached when disseminated. Border-crossing statistics is included in the list of Eesti Pank's statistical activities, which forms a part of the official statistical programme

of Estonia. Compared with earlier expenditures on similar statistics, involving regular border-crossing surveys and the related interviews, the current approach is remarkably cost-effective.

5. Conclusions

A decade of experience of using mobile data for external sector statistics has shown that the methodology offers a reliable overview of travellers crossing the Estonian border. When Eesti Pank started publishing the time series of international travel in 2012, Estonia was one of the first few countries publishing official statistics based on big data. It should be pointed out, however, that the methodological concepts cannot be uniformly applied in every country. Local conditions and the particular statistical goals should be taken into account. The biggest advantage of the method is its speed, as it uses existing information, stored by mobile operators as potential respondents for statistics. There are neither direct costs associated with a network of interviewers nor a burden for travellers as potential respondents. In comparison with earlier expenditures on regular border-crossing surveys with interviews, the current approach is remarkably cost-effective.

The current PPP-based collaboration model enables both parties to take advantage of their specialisation. While Eesti Pank has experience in producing national statistics in traditional fields, OÜ Positium as a spin-off company of the Department of Geography at the University of Tartu is better equipped to model human mobility from different facets. As border-crossing statistics is not a core activity for the central bank, this cooperation enables us to draw on OÜ Positium's long-term experience in the field of direct processing of big data (based on the Hadoop software framework).

References

1. Using mobile positioning data for tourism statistics: methodology and data model. OÜ Positium LBS, Eesti Pank; Tartu 2009 (in Estonian).
2. Ahas, R. Aasa, A., Roose, A., Mark, Ü., Silm, S. 2008. Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management 29(3): 469–486.
3. Using mobile positioning data for tourism statistics: updated methodology and data model. OÜ Positium LBS, Eesti Pank; Tartu 2015 (in Estonian).
4. Eurostat. (2014). Feasibility study on the use of mobile positioning data for tourism statistics.Reports on Eurostat Contract No 30501.2012.001-2012.452.
5. Establishment of reports on payment statistics of credit institutions. Decree of the Governor of Eesti Pank No 8 of 02/06/2014
6. Payment guides. Centus. [www]

https://centus.com/uk/a-guide-to-mcc-codes-tcc-codes-and-mcg-groups (12.11.2017)
7. International Travel Statistics.(2019) Eesti Pank. [www] http://statistika.eestipank.ee/#/en/p/MAKSEBIL_JA_INVPOS/1410 (29.03.2019)

Combining mobile phone data and survey data for the best result: Experience from Indonesia

Sarpono, Titi Kanti Lestari
BPS-Statistics Indonesia

Abstract

Data collection is complicated in Indonesia due to vast geographic distances and difficulty of travel. So is it that the inbound tourism survey for measuring tourism is under coverage. Improving the accuracy of inbound tourism statistics with mobile positioning data. Mobile positioning data (MPD) is considered as one of the most promising big data for measuring the mobility of people, including mobility of tourists based. It holds more information, is much faster and more reliable. However, as a new data source, there also challenges and limitation of big data (including mobile positioning data) that have to be taken into account in order to become a valuable and quality statistics. In MPD there is a lack of qualitative data on tourism motivation and the sample does not include non-roamers. One way to overcome the limitations of big data is by combining it with small data obtained from a survey. This paper shows that combining big and small data will provide optimal results. We used mobile positioning data about roaming activities of cross-border tourists in Indonesia at border areas and mobile usage survey conducted at the borders in order to know the motivations for cross-border movements (both roamers and non-roamers). Then, we come up with the formula that combines big and small data to obtain the best result for tourism statistics.

Keywords

Big Data, tourism statistics, remote area

1. Introduction

Indonesia has a border land with Malaysia, Timor Leste, and Papua New Guinea along 3092.8 km. While, the sea area borders with 10 countries, namely India, Malaysia, Singapore, Thailand, Vietnam Philippines, Australia,Timor Leste, Palau, and Papua New Guinea. This sea border covers 92 leading small island, starting from Miangas Island in the north to Dana Island in the south. During this time foreign tourists were calculated based on the Immigration Office based on the passport swipe, which recorded the traffic of all people entering Indonesian territory. Since the vast condition of the Indonesian territory with diverse border areas (sea and land) and the limitations of the Immigration Office, not all foreign tourists entering Indonesian territory are recorded regularly and on time. There are still many border regions of

Indonesia with neighboring countries that are traditional, so there is no recording of people entering and leaving Indonesian territory. Therefore, to obtain a complete and up-to-date data on foreign tourists, data collection was carried out to calculate the number of foreign tourists visiting these border areas. With the addition of these data, the data of foreign tourists visiting have more coverage and can describe the actual conditions of foreign tourists. To increase the data coverage on the number of foreign tourist visits, especially in border areas that have not been recorded, the BPS Statistics Indonesia and Ministry of Tourism tries to improve the methodology to calculate the number of foreign tourists visiting through the border gates using big data, namely Mobile Positioning Data (MPD) since October 2016. MPD is used in cross border posts in districts where immigration checks are not available and cross-border postal surveys are difficult due to geographical conditions.

This paper shows that combining big and small data will provide optimal results. We used mobile positioning data about roaming activities of cross-border tourists in Indonesia at border areas and mobile usage survey conducted at the borders in order to know the motivations for cross-border movements (both roamers and non-roamers). Then, we come up with the formula that combines big and small data to obtain the best result for tourism statistics.

2. Methodology

Statistical office, currently, has a challenge to adapt to the rapid technological changes, which produce big data. The main challenges for big data for official statistics is obtaining the data sources. The main and considered as one of the gold standard sources of big data is Mobile Call Data Record. Several information can be obtained from Call Detail Record (CDR) such as the location, duration, the phone type, etc. Having this broad information, the use of CDR now not only limited for mobile transaction purpose. Broad range of the CDR use from Business, politics, education to official statistics. Šćepanović et al (2015) showed how the Mobile Phone Call Data can be used as a Regional Socio-Economic Proxy Indicator. Furthermore, CDR can also be used for inferring people migration (Sniowski et.al 2016). Statisticians are stimulated to use big data whether to complement official statistics or to produce indicators. Big data offers great potential for monitoring the sustainable development goals and it has been promoted as a more timely and cheaper alternative to traditional sources of official data (Abdulkadri, Evans, and Ash, 2016). Better decision-making and real-time citizen feedback as the result of more diverse, integrated, timely and trustworthy information which in turn enables everyone to make choices that are good for them and for world they live in (Morales et.al., 2014).

Mobile positioning data (MPD) is considered as one of the most promising big data that can be used for official statistics. Statistics Netherlands studied

several uses of mobile phone data for official statistics, it confirmed that mobile phone data may be of use to statistical topics varying from economic activity, tourism, population density to mobility and road use (de Jonge, 2012). The outcomes of the Eurostat study concluded that at present mobile positioning data can be used as a supplement rather than as a replacement source of data for the current official tourism indicators. Furthermore, the study commissioned by Eurostat explained the use of mobile data as a source for tourism indicators, this new source of data can improve timeliness (in some cases up to near-real time), access to statistical information previously not available (new indicators) calibration opportunities for existing data, better resolution, and accuracy in time and space. In Estonia, MPD has been used as an official source of travel statistics since 2008 (Kroon, 2012).

As a new data source, there also challenges and limitations with big data (including mobile positioning data) that have to be taken into account in order to produce valuable and quality statistics, especially regarding its accuracy. In MPD there is a lack of qualitative data on tourism motivation such as the purpose of the trip, expenditure, type of accomodation and means of transport used (Eurostat, 2014) and the sample does not include non-roamers.

One way to overcome the limitations of big data is by combining it with small data obtained from a survey. MPD still has weakness such as related to privacy issues or confidentiality of the costumers and surveillance and also still expensive. In Indonesia, the use of MPD for Tourism Statistics have been initiated since 2016 by collaboration among Indonesia Ministry of Tourism, Statistics Indonesia, and the main MDP Company in Indonesia, Telkomsel. In order to know the ground truth of MPD, BPS Statistics Indonesia, in collaboration with Ministry of Tourism, conducted an extended Cross Border Mobile Usage Survey in 2017 to know the behavior of the border-crosser in using their mobile also the characteristics of the border crosser. The number of entry gates covered was higher than usual. The aim of the survey was also to obtain information that will be used to form ratios for the formula of additional tourism.

BPS-Statistics Indonesia started to use MPD since October 2016 for inbound tourism statistics collaboration with Indonesia Ministry of Tourism and the main Mobile Network Operators in Indonesia, Telkomsel. MPD is a method of tracking the locations of mobile devices in time and space, collected by MNOs and mobile app developers (Tiru, 2014). Prior to MPD use, Indonesia used administrative data (immigration data) and cross-border (shuttle) survey. The cross-border survey were quite expensive due to the borders areas being remote, and the transportation costs to survey locations are high. Also, the survey is only conducted in a month in selected locations, to estimate the number for a year for the entire border. So, there is a coverage problem in the tourism data in Indonesia. If we compare with the neighboring countries, in Indonesia the neighboring countries only constituted about 7

percent in 2015, while for other countries the neighboring countries constitute about 40 percent of tourism.

Mobile positioning data is used to complement tourism data at cross border posts in which Immigration Checkpoint is not available and difficult to conduct Cross-Border Survey. Before releasing the tourism data, BPS compares international visitor arrival data obtained from cross-border survey, immigration checkpoint at cross-border area, and mobile positioning data. If mobile positioning data from Telkomsel (the mobile network operator) is more than data from cross-border survey and immigration checkpoint, the excess of mobile positioning data will be added to the number of international visitor arrival, this is the number of additional international visitor arrivals from mobile positioning data.

However, the use of MPD in the calculation of visits of foreign tourists is not without obstacles. MPD has several shortcomings, among them not all foreign SIM card users are residents abroad. In addition, there is also no information about the number of SIM cards used per person. MPD also cannot provide information about the characteristics of SIM card owners such as the purpose of the visit and expenditure. Other information such as how many strangers passing by who do not carry a mobile phone or use a local sim card also cannot be known with the MPD.

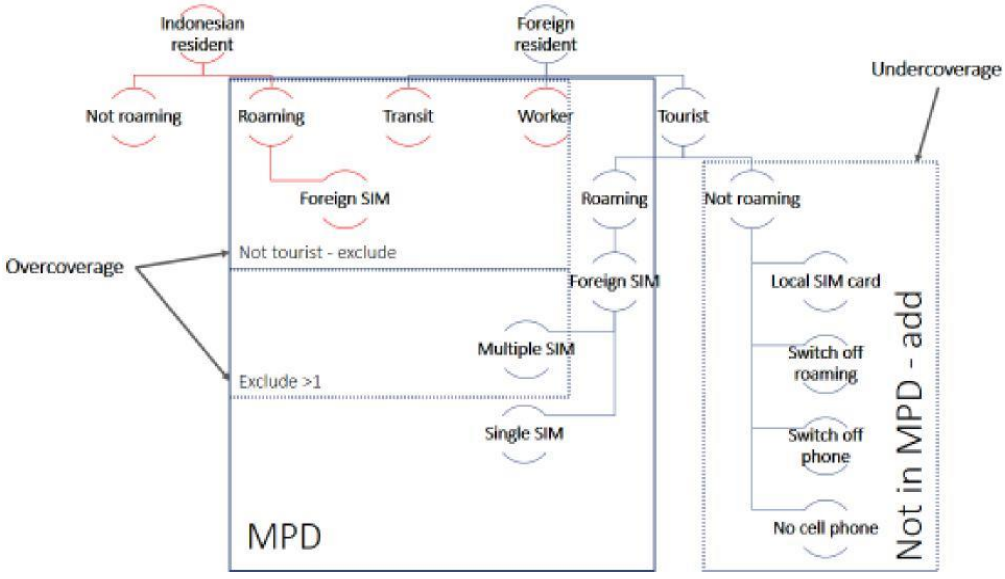


Figure 1. Over and under coverage in MPD

The data collection resulted in cross border ratio as correction factor for MPD data. These values will be used as a basis for calculating the additional number of foreign tourists visiting in the border areas. The first value is the average foreign SIM card brought by the foreign SIM card holders. This number is calculated from the total number of foreign SIM cards divided by

the number of foreign SIM card holders, regardless of whether the card holders are local residents or foreigners. It is also necessary to calculate the ratio of foreign residents who are not recorded in MPD because they do not carry mobile phones or replace their SIM cards with local SIM cards (not roaming). All the above ratios will be used as a correction factor in calculating the number of visits by foreign tourists by MPD. To facilitate the calculation of the additional number of foreign tourists from the MPD, we then made a formula that includes all the above correction factors:

$$AT = \frac{MPD}{X_{roam}} \times \frac{1}{1 - P_{NR}} \times \frac{1}{MS} - WCI$$

Where

AT = Additional Tourism

MPD = Number of SIM cards detected by MNO (Telkomsel) in the border area

X_{roam} = The ratio of foreign SIM cards per person that actively roaming;

P_{NR} = The ratio of foreign residents with foreign SIM cards who turn off their phone, roaming or switch to local SIM card to total number of foreign residents with foreign SIM cards.

MS = Market Sharing

WCI = Number of tourists entering through Immigration Post

3. Results

As mentioned above, BPS Statistics Indonesia conducted cross-border mobile usage survey for inbound tourists at the border area. This survey is conducted by BPS Statistics Indonesia, in collaboration with the Ministry of Tourism, to obtain the ground truth of mobile positioning data, to know the behavior of border crosser in using their mobile and to obtain ratio/proportion for the formula.

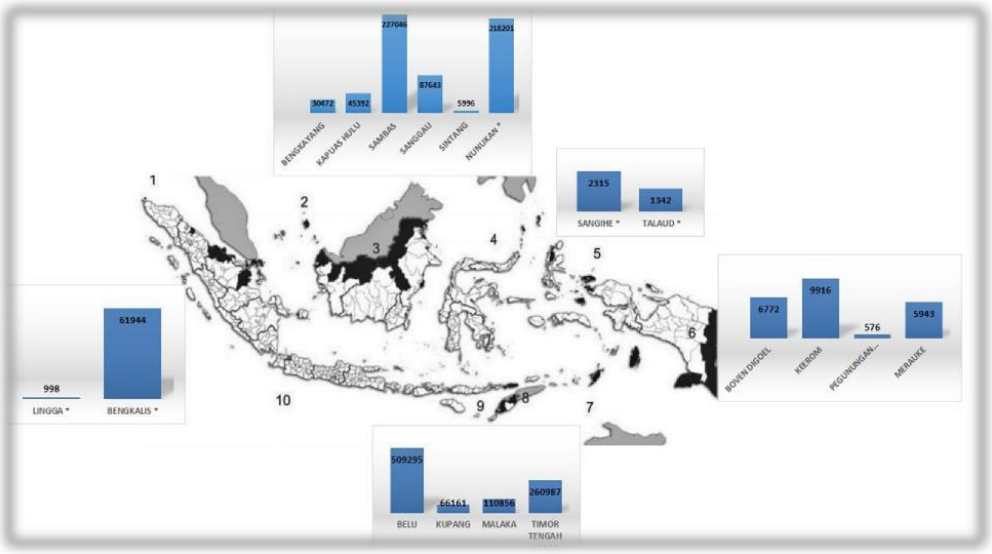


Figure 2. Number of Additional Tourism in Border Areas Jan-July 2018

Figure 2. Shows a total number of additional tourism in border area January to July 2018. The biggest additional number is 509 295 tourists in Belu regency, Nusa Tenggara Timur and the smallest is 576 in Pegunungan Bintan, Papua. Number of additional tourism in border area prove that use the MPD capture significantly and increase the coverage of inbound tourism in Indonesia.

4. Discussion and Conclusion

MPD is useful for BPS Statistics Indonesia as it gives more accurate data on the tourism arrival figures compared to cross-border survey, which can only be conducted during one month and in limited geographical area to estimate the whole year and the entire border. However, there also limitations to MPD. So, there is no data source that is superior compared to other data source. All of data sources could complete each other.

It is important to design a mobile usage survey to accompany the MPD. Moreover, from the cross-validation results, both MPD and survey had their limitations and weaknesses which were apparent once both data sources were compared and used together.

Survey can only be conducted once or twice a year, while MPD is obtained every month with quite high accuracy. After MPD is used, the proportion of cross-border tourism for Indonesia now exceeds 30 percent, which aligns to international benchmarks.

The use of mobile positioning data changes the process of design, build, data collection, data processing, and dissemination. It also changes the process of data collection so that it can enhance the cost efficiency and time efficiency Real time dissemination also can be achieved by using mobile positioning data. Big data as a part of data revolution needs to be developed,

although the verification and its validation need to take into consideration in order to prevent double counting.

This paper showed that combining MPD and survey data will provide optimal results. Mobile positioning data about roaming activities of cross-border tourists in Indonesia at border areas was combined with mobile usage survey at the borders in order to know the motivations and behavior for cross-border movements (both roamers and non-roamers). Then, the formula of additional tourism that combines MPD and mobile usage survey were proposed to obtain the best result for tourism statistics.

References

1. Abdulkadri, Abdullahi, Alecia Evans, and Tanisha Ash. (2016). An Assessment of Big Data for Official Statistics in the Caribbean: Challenges and Opportunity, ECLAC Subregional Headquarters for the Caribbean. Santiago: United Nations.
2. Ahas et.al. (2007). Mobile Positioning Data in Tourism Studies and Monitoring: Case Study in Tartu, Estonia, accessed on 25 January 2017, https://www.researchgate.net/publication/221357419_Mobile_Positioning_Data_in_ . Tourism_Studies_and_Monitoring_Case_Study_in_Tartu_Estonia.
3. Ahn, Jeong-Im and Young-Ja Hwang. (2013). Production of Official Statistics by Using Big Data, Working Paper, accessed on 16 March 2017,http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_3_ Korea.pdf.
4. BPS. (2015). International Visitor Arrival. Jakarta: Badan Pusat Statistik.
5. BPS. (2017). Emerging Challenges in Data Collection, Survey Methodology and Implications for Official Statistics (NSO Perspectives), presented in the Plenary Session, International Statistical Institute Regional Statistics Conference (ISI RSC) 2017, Bali.
6. Kroon, J. (2012) Mobile Positioning as a Possible Data Source for International Travel Service Statistics. UNECE Conference of European Statisticians
7. Landefeld, Steve. (2014). Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues, Discussion Paper, accessed on 16 March 2017, https://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf.
8. Morales et al. (2014). A World that Counts: Mobilising the Data Revolution for Sustainable Development. The United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG).

9. Sanja Šćepanović, Igor Mishkovski, Pan Hui, Jukka K. Nurminen, and Antti Ylä-Jääski. (2015). Mobile Phone Call Data as a Regional Socio-Economic Proxy Indicator.
10. Seynaeve, Gerdy and Christophe Demunter. (2016). When Mobile Network Operators and Statistical Offices Meet-Integrating Mobile Positioning Data into the Production Process of Tourism Statistics, the 14th Global Forum on Tourism Statistics.
11. Sniowski, Alessandro, Sorichetta, Ingmar Weber, and Andrew J. Tatem (2016). Inferring Migrations: Traditional Methods and New Approaches based on Mobile Phone, Social Media, and other Big Data. *Technical Report*. European Union.
12. Rein Ahasa, Anto Aasaa, Antti Roosea, Ülar Markb, Siiri Silma (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, Vol 29, Issue 3, pp 469–486.
13. Tiru, Margus. (2014). Overview of the Sources and Challenges of Mobile Positioning Data for Statistics, presented in the International Conference on Big Data for Official Statistics. Beijing: UN Trade Statistics, Positium, and University of Tartu.

The use of mobile phone data in Tourism Statistics

Ossi Nurmi; Pasi Piela
Statistics Finland

Abstract

Survey response rates are declining at an alarming rate globally. Statisticians have traditionally used imputing and recalibration of weights to deal with nonresponse. In case survey response rates are well below 50 %, these methods may often result in little more than guesswork. Alternative data sources need to be used to improve the accuracy of statistical estimates. In the context of outbound tourism, mobile positioning data can be considered as such an alternative data source as it registers 'traces' of tourism trips. These traces are CDRs (call detail records) and DDRs (data detail records) and they are generated by the activities of mobile devices. Since 2016 Statistics Finland has worked closely with Finnish national mobile network operators (MNOs) to translate the CDRs and DDRs into tourism specific monthly aggregates such as number and duration of trips by destination country. Statistics Finland has learned that it helps to be very specific when approaching MNOs with data needs. This paper provides a summary of the methodological process that the Finnish MNOs have followed to compile tourism statistics. A similar process may be used by National Statistics Institutes or other organizations who are approaching their national MNOs with the intent of obtaining data. The paper then presents the 2017 outbound tourism data from Finnish MNOs and highlights the shortcomings of the Finnish national tourism survey in light of mobile positioning data. Based on this analysis, the paper proposes a method to enrich the tourism survey using mobile positioning data. One of the main needs for outbound tourism data comes from the Balance of Payments (BoP) statistics. The debit side of BoP requires quarterly data by destination country and purpose of trip. More accurate data is needed to reliably estimate the expenditure of resident tourists abroad. The paper proposes a method where mobile positioning data is first used to estimate the number and duration of outbound trips by country and month. The role of tourism survey data is then to provide ratios such as purpose of trip, means of transport and average expenditure. In this method the tourism survey is no longer needed for estimating the absolute number of trips.

Keywords

Mobile Positioning; Tourism; Geospatial data

1. Introduction

During 2016 up to 2018, Statistics Finland carried out the work in two phases within the context of Eurostat's ESSNet Big Data –project. The focus in these projects was on three statistical domains: inbound tourism, outbound tourism and seasonal population. The first phase in 2017 and 2017 focused on negotiations with national authorities and MNOs in order to set up such a process that is feasible from a legislative and technical point of view. The second phase was to carry out the process, collect data from each MNO and analyze the results. The chosen approach relies on the operators to process the data and aggregate it for Statistics Finland. In the current Finnish legislation, only the operator is allowed to process the raw data using automatic means. The size of the raw data is also massive, with annual data consisting of several billions of events per operator.

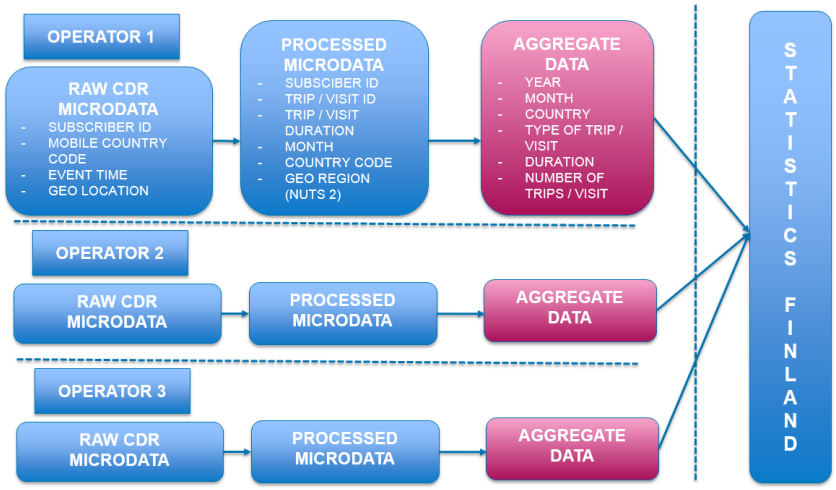


Figure 1 *Process from raw microdata to aggregated trips*

The starting point for each operator are the raw data of each of their subscribers. The subscriptions are associated with sim cards found on mobile devices. Machine-to-machine sim cards are excluded from the data as they do not represent the movement of people.

2. Methodology for raw data: processing roaming events to tourism trips

In case of outbound trips, the raw data consists of roaming events (calls, sms, mobile data) taking place outside of the subscriber's home network, in other words, a foreign country where the event took place. Based on the time gaps between these events, individual trips of each subscriber can be recognized.

The following example presents an imaginary case of the roaming events and outbound trips deduced for a single subscriber during a 30 days period. This subscriber is a particularly active traveller and five outbound trips are registered during the 30 days period.

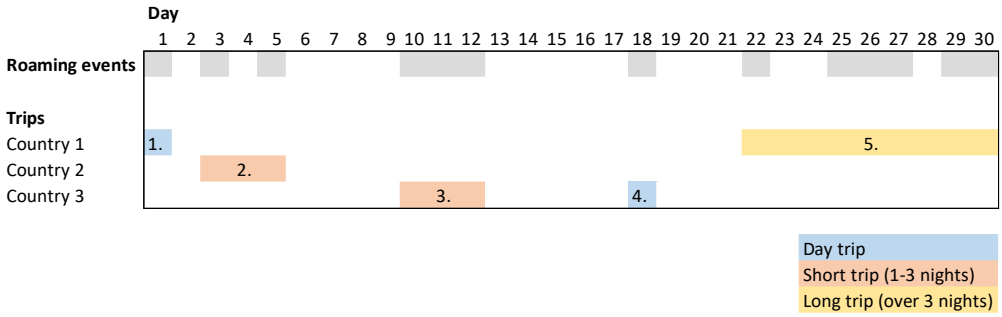


Figure 1 - Roaming events and outbound trips of a subscriber

Roaming events are registered whenever the subscriber is making calls, sending sms or using data in a mobile network abroad. The first event in a country indicates the beginning of the trip to that particular country. A total of five outbound trips were registered for this subscriber.

3. Methodology for estimating outbound trips

The target population consists of all outbound tourism trips made by Finnish residents. The reference data is collected by Statistics Finland's Finnish Travel –survey with a sample size of 28,200 persons out of which 14,700 were interviewed by phone concerning trips that ended during 2017. The data provided by the Finnish MNOs should be treated as samples of all outbound trips made by Finnish residents in 2017. Two out of three operators have provided the data and the market share of each operator is roughly one third of all subscribers. The source data thus contains the outbound trips of two thirds of the Finnish residents and one third are missing. As with every sample, the MNO data is of limited use by itself and coefficient weights are needed to estimate the target population: all outbound tourism trips. According to the Finnish Travel –survey, Finnish residents made 10.5 million outbound trips in 2017. This is annually a relatively stable figure based on 2,700 trips reported by the respondents. In order to avoid many of the known pitfalls related to over- or underestimation in MNO data, this annual figure of 10.5 million outbound trips, is used as a frame and MNO data is first simply weighted using the following formula:

$$Weight\ coefficient\ = \frac{Annual\ outbound\ trips\ in\ tourism\ survey}{\sum(MNO\ annual\ trips)_n}$$

This simple weight coefficient is thus obtained by dividing the annual outbound tourism trips by sum of trips made by the subscribers of each MNO.

In the case of Finnish data from two (out of three) operators, the coefficient for year 2017 is less than 1.3.

4. Results: simple weight coefficient for total number of outbound trips

The monthly distribution of total outbound trips is shown below separately for the survey as well as the data provided by both MNOs.

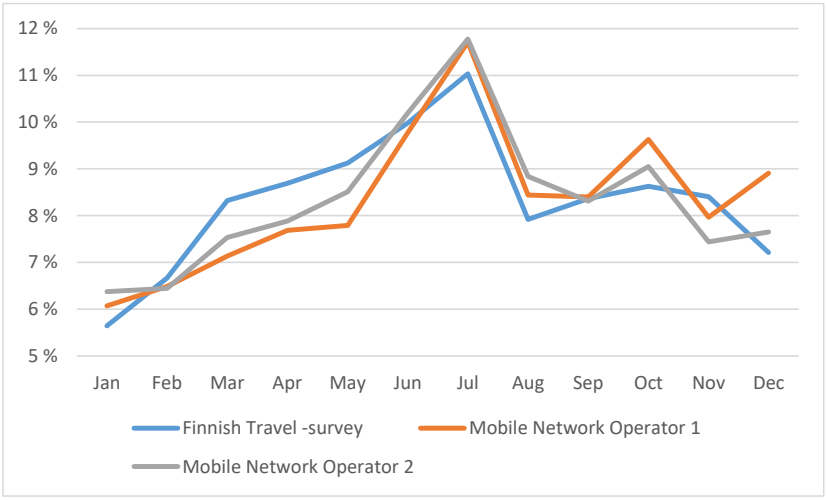


Figure 2 - Monthly seasonality of outbound trips using different sources in 2017

The monthly seasonality of outbound trips is strikingly similar for both MNOs. Both register the summer holiday months of July and June as the peak months of outbound tourism. The third biggest month for both operators is October, which is the month of autumn holidays in Finnish schools. The Finnish Travel –survey registers more outbound trips than the MNOs during the months of March through May. Concerning tourism by country, neighboring Estonia and Sweden are by far the most important outbound tourism destination countries. 45 per cent of outbound trips are made to these two countries.

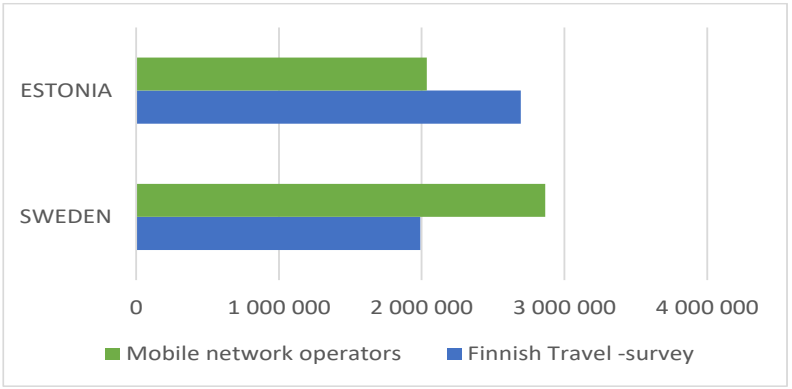


Figure 3 – *Trips to Estonia and Sweden in 2017*

There were 2.6 million trips to Estonia in 2017 according to Finnish travel –survey. Using the top-down approach for mobile positioning data, 24% less trips appear in MNO data. This indicates that this method underestimates tourism to Estonia, as the 2.6 million trips from the survey is a very stable figure with only small changes annually. In contrast, the outbound tourism to Sweden comprises of only 2.0 million trips in the survey, while the top-down approach estimates 44% more trips to Sweden in MNO data. The neighboring countries with open land borders have many possible sources of overestimation such as frequent non-tourism trips, border noise etc. There is yet another reliable data source for providing the monthly seasonality of outbound trips to Estonia, the main outbound tourism destination. Nearly all passengers to Estonia use one of the ferries that operate between the capital cities of Helsinki and Tallinn. The Finnish Transport Agency compiles statistics on the total monthly passengers departing to Estonia, including passengers of all nationalities.

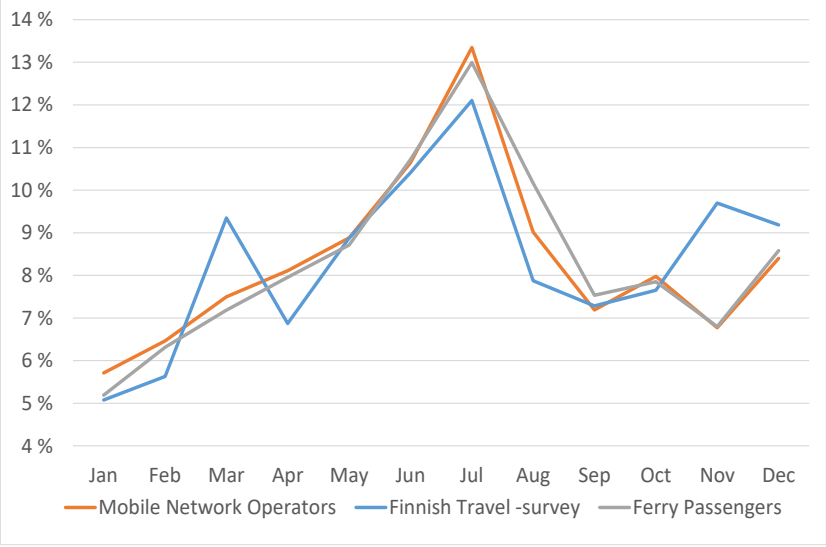


Figure 4 – Monthly seasonality of trips to Estonia in 2017

The seasonality of ferry passengers is in line with the outbound tourism data provided by the MNOs. In contrast, the seasonality of the Finnish Travel –survey seems to be affected by randomness. Some months in the survey are exceptionally high (March, November) and some too low (April, August) when compared to ferry passengers and MNO data. Although the MNO data slightly underestimates the total number of trips to Estonia, it appears to be a better data source for estimating the monthly seasonality. Unlike the survey, it's not affected by the randomness caused by a small sample size. This suggests, that MNO data should be used to adjust the monthly seasonality estimates produced the survey. The main findings from these results are the following:

1. The data from two different operators are highly correlated. The monthly seasonality is nearly identical and in line with outbound tourism based on Finnish Travel –survey.
2. Depending on country of destination, the top-down approach dramatically over- or underestimates the total number of outbound tourism trips to that country. There are many known sources of bias in mobile data: non-tourism trips, border noise, devices switched off, multiple devices, transit corridors, conceptual differences etc.
3. Mobile positioning data provides a better estimate on the monthly seasonality of outbound tourism. The monthly estimates of Finnish Travel -survey are affected by randomness due to small sample size.

5. Results: Country specific estimation

How should the survey data be enriched or recalibrated using MNO data in order to improve the accuracy? The proposed method for recalibrating the outbound trips data has to provide at least the following estimations:

1. Annual number of all outbound trips
2. Monthly seasonality of outbound trips to each country
3. Annual number of outbound trips to each country
4. Year-on-year change in the number of outbound trips

For annual number of outbound tourism trips (1.) the Finnish Travel-survey provides a solid estimate as shown earlier. For monthly seasonality (2.) the MNO data is more robust as it is not affected by survey randomness. For trips to each country (3.), the best data source depends on the country of destination. At present, it's not possible to evaluate the year-on-year change as MNO data is available only for 2017.

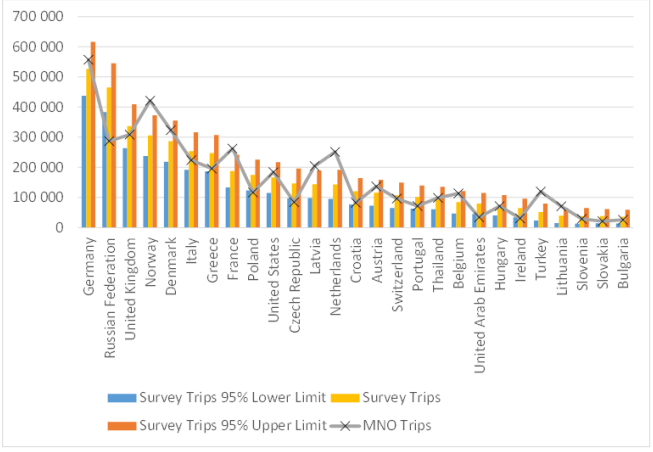


Figure 6 – 95 % confidence intervals for top 30 destination countries (excluding top 3)

The figure presents the number of outbound trips to each country based on the Finnish Travel –survey as well as the upper and lower bounds of the 95% confidence interval for the survey estimates. The number of trips based on MNO data is plotted against them as a line graph. Using this kind of method, the MNO data provides a 'second opinion' to the survey confidence intervals for each country. In case the MNO trips are outside of the confidence interval for a certain country, the MNO data most likely includes serious sources of over- or underestimation for that country. On the other hand, the survey estimates also become rather useless if the 95% confidence interval is too large. At present, countries with less than 170,000 trips have 95% confidence interval limits of plus or minus 30 per cent. For most of such small destination countries, the MNO trips can still provide a better estimate, given that the MNO trips are within the confidence interval.

There are currently only 24 destination countries where the annual estimates are considered reliable. On a monthly level, only trips to Estonia and Sweden are mostly reliable. In total there are 9 million outbound trips to these countries. The MNO data can potentially provide trips to 129 more smaller destination countries with 1,5 million trips in total.

6. Discussion and Conclusion

In the context of outbound tourism, the strengths and weaknesses of mobile positioning and survey data can be summarized as follows:

Table 1 – *Strengths and weaknesses of survey and mobile positioning data*

	Finnish Travel -survey	Mobile positioning data
Strengths	**Scope** is clean: only tourism trips are included **Provides supporting information** of the trip (ie. purpose of trip, expenditure, means of transport and accommodation)	**Granularity:** millions of observations covering nearly all destination countries **Monthly seasonality** of tourism is more accurate.
Weaknesses	**Granularity;** very few observations per year, covering only a few destination countries **Monthly seasonality** estimates are affected by randomness	**Scope** is not clean, there are many sources of over- or underestimation **No supporting information** of the trip

By recalibrating the existing survey data using data from MNOs, it's possible to combine the strengths and alleviate the weaknesses in both data sources. This results in significant improvement in geographical as well as temporal granularity of outbound tourism statistics. Geographical granularity refers to the number of countries that can be reported, while the temporal granularity is the monthly tourism seasonality to each of those countries.

This study was conducted by using data from two MNOs consisting of roughly two thirds of all Finnish mobile subscribers. The data from both operators are highly correlated. This suggests that similar results could be obtained by using data only from a single MNO, given that its subscribers are not geographically or socio-demographically biased.

References

1. Eurostat (2018). ESSNET Big Data Project. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data
2. Statistics Finland. Finnish travel statistics. Available at: http://www.stat.fi/til/smat/index_en.html
3. Eurostat (2014). Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics. Consolidated Report Eurostat Contract No 30501.2012.001 – 2012.452. Available at: https://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf

Improved robust rank-based test statistics in high-dimensional regression model

Mahdi Roozbeh
Department of Statistics, Faculty of Mathematics, Statistics and Computer Sciences, Semnan University, Iran

Abstract

In classical regression analysis, the ordinary least-squares estimation is the best estimation method if the essential assumptions such as normality and independency to the error terms as well as a little or no multicollinearity in the covariates are met. More importantly, in many biological, medical, social, and economical studies, nowadays carry structures that the number of covariates may exceed the sample size (high-dimension or wide data). In this situation, the least-squares estimator is not applicable. However, if any of these assumptions is violated, then the results can be misleading. Especially, outliers violate the assumption of normally distributed residuals in the least-squares regression. Robust ridge regression is a modern technique for analyzing data that are contaminated with outliers in high-dimensional case. When multicollinearity exists in the data set, the prediction performance of the robust ridge regression method is higher than rank regression method. Also, the efficiency of this estimator is highly dependent on the ridge parameter. Generally, it is difficult to give a satisfactory answer about how to select the ridge parameter. Because of the good properties of generalized cross validation (GCV) and its simplicity, we use it to choose optimum value of the ridge parameter. The proposed GCV function creates a balance between the precision of the estimators and the biasness caused by the ridge estimation. It behaves like an improved estimator of risk and can be used when the number of explanatory variables is larger than the sample size in high-dimensional problems. Finally, some numerical illustrations are given to support our findings for the analysis of gene expression and prediction of the riboflavin production in Bacillus subtilis.

Keywords

Generalized cross validation; High-dimension data; Multicollinearity; Rank regression; Robust ridge regression; Spare model.

1. Introduction

Consider the setting where the observed data are realizations of $\{(Xi, yi)\}$ with p-dimensional covariates $X_i \in R^p$ and univariate continuous response variables $y_i \in R$. A simple high-dimensional regression model has form

$$y_i = x_i'\beta + \varepsilon_i \text{ , } i = 1,2,...,n \tag{1.1}$$

where $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ is the vector of regression coefficients and ε_i is the i^{th} error component, having a continuous cumulative distribution function (c.d.f.), $f(.)$ And finite fisher information, $I(f)$,

$$I(f) = \int_R \left(\frac{f'(x)}{f(x)} \right)^2 f(x)dx < \infty, \; f(x) = \frac{dF(x)}{dx}, \; f'(x) = \frac{d^2F(x)}{dx^2} \tag{1.2}$$

Now, consider a regression model in the presence of multicollinearity. The existence of multicollinearity may lead to wide confidence intervals for the individual parameters or linear combination of the parameters and may produce estimates with wrong signs. For our purpose we employ the ridge regression concept due to Hoerl and Kennard (1970), to combat multicollinearity. There are a lot of works adopting ridge regression methodology to overcome the multicollinearity problem. To mention a few recent researches, see Hassanzadeh Bashtian et al. (2011), Amini and Roozbeh (2015), Arashi et al. (2015), Arashi and Valizadeh (2015) and Roozbeh (2016).

When $p < n$, a classical method to deal with this problem is the famous F-test statistic. However, it is shown that the power of F-test is adversely impacted by an increased dimension. Moreover, the F-test statistics is undefined when the dimension of data is greater than the within sample degrees of freedom since the pooled sample covariance matrices are not positive definite. In order to overcome this issue, we propose a robust test based on rank regression in case $p > n$.

We organize this article as follows: In Section 2, we propose a robust ridge test statistic based on rank regression. In Section 3, some regularity conditions are given, while approximate distribution of the test statistic is given in Section 4. Definition of a Stein-type shrinkage estimator and evaluating the ridge and shrinkage parameters using GCV criterion, are the content of Section 5. Numerical studies are the context of Section 6. We conclude our results in section 7.

2. Robust Test Statistics

The problem of our study is to find a rank-based test statistic for testing the following set of hypotheses:

$$H_o: \beta = \beta_0 \text{ vs } H_A: \beta \neq \beta_0. \tag{2.3}$$

Testing hypotheses similar to (2.3), has been considered by many authors. To be more specific about their finding and predispose our result, we first consider the following rank-based score test (see Hettmansperger and

McKean, 1998, Ch.3):

$$a^{T}(R(y))Ha(R(y)), \tag{2.4}$$

where for $y = (y_{1,\ldots,}y_n)^T, a\big(R(y)\big) = (a\big(R(y_1)\big), \ldots, (a\big(R(y_n)\big))^T$, and $R(y_i)$ is the rank of yi, $i = 1,\ldots,n, a(1) \leq a(2) \leq \ldots \leq a(n)$ is a set of scores generated as $a(i) = \psi(i/(n+1))$ for some square-integrable and non-decreasing score function $\psi(u)$ defined on the unit interval, satisfying

$$\int \psi(u)\mathrm{d}u = 0 \text{ and } \int \psi^2(u)\mathrm{d}u = 1,$$

and $H = X(X^TX)^{-1}X^T$ is the projection matrix onto the space Ω_F, the column space spanned by the columns of X and $X = (\mathrm{X}^{\mathrm{T}}, \cdots, \mathrm{X}^{\mathrm{T}})^{\mathrm{T}}$.

When the conditional distribution of y_i given X_i is normal with $p < n$, the classical test for $H_O: \beta = 0$ is the F-test. The F-statistic is a monotone function of the likelihood ratio statistic and is distributed as a noncentral F distribution under the alternative (Anderson, 2003). It is interesting to know the power implication on the F-test when $p/n \to \rho \in (0,1)$ when both p and n diverge to infinity. The F-statistic for testing H_O has form

$$F = \frac{\hat{\beta}^T X^T X \beta}{y^T (I-H) y}, \quad \hat{\beta} = (X^TX)^{-1}X^T y \tag{2.5}$$

Under H_O, F has central Fisher distribution with p and $n-p$ degrees of freedom (d.f.). Hence, an α- level F-test rejects H_O if $F > f_\alpha(p, n-p)$, the upper α-level critical value of the F-distribution with p and n- p d.f.

Under situations in which the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is ill-conditioned due to linear relationship among the regressors of X matrix (multicollinearity problem) or the number of independent variables is larger than sample size, usual estimators are not applicable, since we always find a linear combination of the columns in X which is exactly equal to one other. Mathematically, the design matrix is not full rank, $rank(X) \leq min(n,p) < p$ for $p > n$, and one may have $X\beta = X(\beta + \zeta)$ for every ζ in the null space of X. Consequently, without making further assumptions on the model characteristics, it is impossible to infer/estimate β from data. This issue is almost similar to the classical setting $p < n$ with $rank(X) < p$ (due to linear dependency among covariates) or ill-conditioned design matrix, leading to difficulties about identifiability. However, for prediction/estimation of $X\beta$, identifiability of the parameters is not necessarily needed. From a practical point of view, high empirical correlations among two or a few other covariates lead to unstable results for estimating β or for pursuing variable selection. To overcome this problem, we can use the ridge estimation. In what follows, we revisit (2.4) by considering its ridge version in case $p > n$.

To be more specific and motivate our approach, one way of controlling β not to deviate much from the origin, i.e., satisfying the null-hypothesis $H_0 : \beta = 0$, is simply not to let $||\beta||^2 = \beta^T \beta$ get larger. In other words, one may think of constrained hypothesis testing in which the penalty term $||\beta||^2 < \lambda$, for some $\lambda > 0$, is taken into account. This key element recalls the well-known ridge regression approach, where the ridge estimator of β is given by

$$\hat{\beta}(k) = (X^T X + kI_p)^{-1} X^T y, \tag{2.6}$$

for some ridge parameter $k > 0$. Concentrating on the shrinkage factor $(XTX + kI_p)^{-1}$, one may think of replacing $(XTX)^{-1}$ by $(XTX + kI_p)^{-1}$ in (2.4) rather than eliminating any component, when XTX is not invertible. Hence, we propose a rank-based test statistic for testing (2.3), when $\beta 0 = 0$ by incorporating the shrinkage factor $(XTX + kI_p)^{-1}$ as

$$R_n(k) = \sigma a^{-2} a^T(R(y)) X(n^{-1}X^T X + kI_p)^{-1} [X(k)]^{-1} (n^{-1}X^T X + kI_p)^{-1} X^T a(R(y)), \tag{2.7}$$

where $X(k) = (n^{-1}X^T X + kI_p)^{-1}$-k$(n^{-1}X^T X + kI_p)^{-2}$ is an invertible matrix formulated based on XTX (see Eq. (4.2)), $\sigma a^2 = \frac{1}{n-1}\sum_{j=1}^{n} a_j^2$ and $k > 0$ is a regularization (ridge) parameter. It will be shown that $R_n(k)$ has approximate $\chi 2$ distribution with p d.f.. Our test statistic has some advantages compared to the previously proposed results, which are listed below:

1. Our approach does not take any asymptotic assumption for $\lim_{n\to\infty} n^{-1} X^T X = \Sigma$, i.e., of small $o(.)$ concept.
2. There is no need to have many knowledges about asymptotic theory in high-dimension to derive the approximate distribution of the test statistic.
3. The estimate of Σ can be easily achieved.
4. The shrinkage estimator based on $R_n(k)$ performs much better than the ridge estimator in the sense of having smaller risk values.

Theorem 1 ***Under the specified model*** **(1.1)**, ***reject*** $\mathbf{H_o}$ ***in favor of*** $\mathbf{H_A}$ ***at approximate level*** α ***iff*** $R_n(k) \geq \chi^2(\alpha)$, ***where*** $\chi^2(\alpha)$ ***denotes the upper level*** α ***critical value of*** χ^2 ***distribution with p d.f.***

3. Shrinkage Estimator

In this section, we define a Stein-type shrinkage estimator for β based on the rank-statistic Rn(k). Further, we evaluate the unknown parameter of our estimator making use of a generalized cross validation criterion. According to the result of Hettmansperger and McKean (1998, Ch.3), under the assumptions of Section 1, for p < n case, the rank-estimate of β is given by

$$\hat{\beta}_{\psi} = (X^T X)^{-1} X^T \hat{y}_{\psi}, \tag{3.1}$$

where $\hat{y}_{\psi}$ is the minimizer of $D_{\psi}(\eta) = ||y-\eta||_{\psi}$ over $\eta \in \Omega_F$ and $||v||_{\psi} = \sum_{i=1}^{n} a(R(v_i))v_i$. Thus, $\hat{\beta}_{\psi}$ is the solution to the rank-normal equations $X^T a(R(y-X\beta)) = 0$. In the same fashion as in formulating the ridge estimator, we define a robust ridge estimator as

$$\hat{\beta}_{\psi}(k) = (n^{-1} X^T X + k I_p)^{-1} X^T \hat{y}_{\psi}, \tag{3.2}$$

where $k > 0$ is the ridge parameter.

One way of improving upon the robust ridge estimator, is to incorporate the information consists in $\beta = 0$. Preliminary test estimator, emanating from testing the null-hypothesis H_0, which leads to select one of the extremes 0 or $\hat{\beta}_{\psi}(k)$ depending on the output of test, is one way of improvement. However, this estimator is heavily dependent on the size of the test and has discrete nature. Hence, we consider its continuous version, and shrink the ridge estimator toward the origin by proposing the following Stein-type shrinkage estimator (SSE)

$$\begin{aligned} \hat{\beta}_{\psi}(k,d) &= \left(1-\frac{d}{R_n(k)}\right)\hat{\beta}_{\psi}(k) \\ &= \hat{\beta}_{\psi}(k) - d R_n(k)^{-1} \hat{\beta}_{\psi}(k),\ d > 0. \end{aligned} \tag{3.3}$$

The SSE depends on ridge parameter k and shrinkage parameter d that must be evaluated in practice. To this end and finding the optimal values, we use the generalized cross validation (GCV) criterion for selecting the optimal values of both parameters, simultaneously. The GCV has been applied for obtaining the optimal ridge parameter in a ridge regression model (Golub *et al.*, 1979) and for obtaining the optimal ridge parameter and bandwidth of the kernel smoother in semiparametric regression model (Amini and Roozbeh, 2015) as well as partial linear models (Speckman, 1988). Our proposed GCV criterion creates a balance between the precision of the estimators and the biasedness caused by the ridge and shrinkage parameters. The GCV function is then defined as

$$GCV(k,d) = \frac{\frac{1}{n}||(I-L(k,d)y||^2}{\left(1-\frac{1}{n}tr(L(k,d))\right)^2} \tag{3.4}$$

where $L(k,d) = \left(1-\frac{d}{R_n(k)}\right)X(n^{-1}X^T X + kI_p)^{-1}X^T$.

Theorem 2 *For the GCV function in (3.4)*

$$\lim_{n\to\infty} R(\hat{\beta}_{\psi}(k,d)) = \sigma^2 + \lim_{n\to\infty} GCV(k,d). \quad (3.5)$$

4. Real Data Study

To illustrate the usefulness of the suggested strategies for high-dimensional data in the regres- sion model, we consider the data set about riboavin (vitamin B2) production in Bacillus subtilis, which can be found in R package "hdi". There is a single real valued response variable which is the logarithm of the riboavin production rate. Furthermore, there are p = 4088 explanatory variables measuring the logarithm of the expression level of 4088 genes. There is one rather homogeneous data set from n = 71 samples that were hybridized repeatedly during a fed batch fermentation process where different engineered strains and strains grown under different fermentation conditions were analyzed. Table 1 shows a summary of the results. In this Table, the *RSS* and R^2 respectively are the residual sum of squares and coefficient of determination of the model. The 3D diagram of GCV versus k and d is plotted in Figure 1 for real data set. The minimum of GCV approximately occurred at k_{opt} = 0.468759 and d_{opt} = 0.002332.

Estimator	$\widehat{\boldsymbol{\beta}}$(k)	$\widehat{\boldsymbol{\beta}}$ (k)	$\widehat{\boldsymbol{\beta}}^{(s)}$(k,d)
RSS	11.440721	1.216623	0.068050
R^2	0.807080	0.979485	0.998853

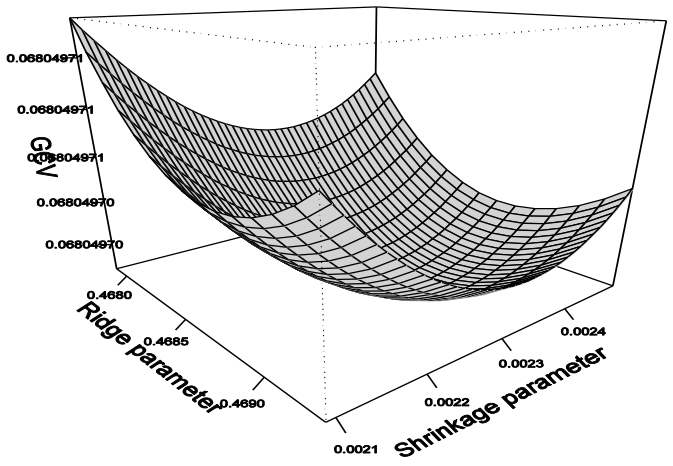


Figure 1: The diagram of GCV versus k and d for real data set

References

1. Amini, M. and Roozbeh, M. (2015). Optimal partial ridge estimation in restricted semipara-metric regression models. Journal of Multivariate Analysis, 136, 26-40.
2. Arashi M, Janfada M, Norouzirad M. (2015). Singular ridge regression with stochastic con- straints, Comm. Statist. Theo. Meth., 44, 1281-1292.
3. Arashi M, Valizadeh T. (2015). Performance of Kibria's methods in partial linear ridge regression model. Stat. Pap., 56, 231-246.
4. Golub, G., Heath, M. and Wahba, G. (1979). Generalized cross validationas a method for choosing a good ridge parameter. Technometrics 21, 215-223.
5. Hassanzadeh Bashtian, M., Arashi, M. and Tabatabaey, S. M. M. (2011). Using improved estimation strategies to combat multicollinearity. J. Statist. Comp. Sim., 81, 1773-1797.
6. Hettmansperger, T. P. and McKean, J. W. (1998). Robust Nonparametric Statistical Methods, London: Arnold.
7. Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for non-orthogonal problems, Thechnometrics, 12, 69-82.
8. Roozbeh, M. (2016). Robust ridge estimator in restricted semiparametric regression models. J. Mul. Anal., 147, 127-144.
9. Speckman, P. (1988). Kernel somoothing in partial linear models. Journal of the Royal Statis- tical Society, Series B, 50, 413-436.

Perfection of volatility prediction with time scale information using wavelet transformation

Md. Sabiruzzaman, Md. Kamrul Islam
Department of Statistics, University of Rajshahi

Abstract

Volatility of stock market returns is one of the major concerns among investors, industrialists and policymakers. GARCH family model is a very popular tool for describing existence of volatility clustering in such time series data. However, it lacks from interpreting time-scale variation that is an important issue for different levels of investors. An efficient way of representing a time series with such complex dynamic is given by wavelet methodology. With the help of a wavelet basis, the Maximal Overlap Discrete Wavelet Transform (MODWT) is able to break a time series with respect to a time scale while preserving the time dimension and energy. Time scale specification information is necessary if one accepts the view that stock market consist of heterogeneous investors operating at different time scales. In that case, considerable more insight in to the volatility dynamic can be gained by looking at the data at several time scales. Wavelet transformations are also fast to calculate and are ideally suited for analyzing large data set. This paper provides an improved alternative to the classical econometric tools in the financial markets prediction. Forecasting stock market volatility with wavelet analysis is the central element of this paper. A novel algorithm, where wavelet transformation is incorporated to an econometric model, is implemented in order to improve the performance of volatility prediction. On the analyzed data we showed that our forecasting algorithm has achieved better results compared with the approach which not using the wavelet transform.

Keywords

Volatility prediction; Wavelet transformation; Time scale variation; Stock index

1. Introduction

In the financial domain of stock market estimation, prediction of the risk of holding assets is a challenging job. In stock market, there are different level of investors who are concerned about the market series and volatility on different time horizon. Therefore, the researcher pays much attention to give proper information for the different level of investors. They need more time to collect and analyze the data at different time horizon. For example, if one wants to estimate the risk for daily, weekly and monthly investor. And he has a daily data, when the return interval is increased in a given sample period, the number of sample points will decrease which result in loss of information. It is

not only the time interval which makes a difference in risk estimation, but also the sampling rule employed to construct a particular time series. As for example, in constructing a monthly time series from daily data, the last business day of each month might be accepted as a representative of that month. However, there is no reason why the day before the last business day of each month should be a representative day of each month should be a representative day or two business days before the last business day and so on. Wavelet analysis works as a magic tool to analyze the data at different time horizon. The wavelet analysis decomposes the time series at the highest possible frequency into different time scale. Hence it provides a natural platform to investigate the risk at different time scale without losing any time point. Therefor it gives benefit to the researcher to analyze the time series at different time horizon at a time without giving much effort.

The maximum overlap discrete wavelet transformation (MODWT), a modified version of discrete wavelet transformation (DWT), is applicable to both dyadic and non-dyadic time series, capable of preserving information of the original signal, shift invariant and more efficient than DWT for variance analysis (Gencay et al., 2002). Applications of MODWT for time scale decomposition of time series are found in number of recent literature (Al.Wadi et al., 2013; Alves et al., 2014; Gallegati et al., 2014; Reboredo et al., 2014). For more on MODWT, we refer Percival and Walden (2000).

This study proposes a new algorithm for volatility prediction containing multi-scale information and illustrates with weekly index of Dhaka Stock Exchange (DSEX). Multiresolution analysis with MODWT is performed to obtain variability at different time scale corresponding to different level of investor and estimate risk at different scale of time using GARCH model. Time scale variations are then incorporated in volatility prediction of the original series through wavelet reconstruction.

2. Volatility prediction with Wavelet-GARCH Approach

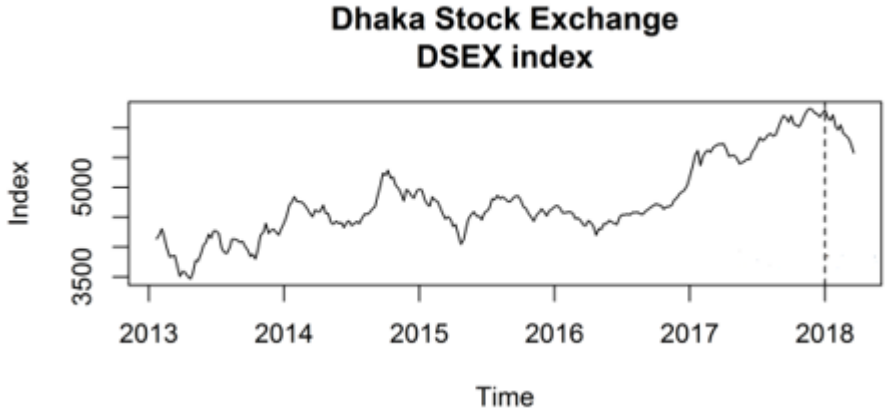
In this study, we proposed wavelet-GARCH approach to predict volatility and compare it with econometric approach. The prediction scheme capture time scale variation at different levels from the wavelet domain instead of just applying a forecasting algorithm directly on the raw data as many econometric models do. Thus, having information of multiple scales and using an adequate model for financial time series, the prediction accuracy is improved. In our proposed approach, we first decompose the return series using MODWT decomposed to obtain approximations and detail coefficients. While approximations represent the location, detail coefficients represent variability of the series at different scales. For example, detail at level j, dj represents the variation at scale 2j. The detail coefficients are, therefore, modelled with an appropriate GARCH equation to estimate volatility at different scales. Detail coefficients at each level are then replaced with estimated GARCH volatility

with an error distribution (e.g., normal, ged and t). The return series are reconstructed with the new detail coefficients. The reconstructed return series is then modeled with a GARCH equation and forecasted as well.

3. Illustration

For illustration of the proposed method, we use the weekly index of Dhaka Stock Exchange (DSEX) spanned from Feb 2013 to Feb 2018 is obtained from the DSE website (http://www.dsebd.org/recent_market_information.php). DSEX is the main index of Dhaka Stock Exchange, which reflects around 97% of the total equity market capitalization. The weekly prices are the divide in to two sets: training data (07-02-2013 to 28-12-2017) and test data (04-01 -2018 to 15-02-2018) (see Fig 1). The historical volatility of the test period is computed with exponentially weighted moving average (EWMA). The algorithm is run with two mother wavelets, the Haar and the Symlets. Although the Haar wavelet is considered to be the simplest mother wavelet function, the choice of this transform was motivated by the fact that its shape and analytical definition is similar to the financial time series patterns. The Symlets is also a good alternative in this specific wavelet analysis since it captures the asymmetry of financial data.

The wavelet-GARCH forecasts are compared with traditional GARCH forecasts with a simulation study by means of two different criteria: root mean square error (RMSE) and Dynamic Time Wrapping (DTW) distance. RMSE is a well-accepted and widely used measure of predictability in the field of econometrics. On the other hand, DTW distance is popular tool for measuring the similarity of simulated time series with original or referenced series. It is a window based algorithm which considers the trend of time series data. DTW distance does not require the sampling time of two time series are synchronous, not be sensitive to abnormal points, furthermore, it is able to measure the similarity of time series with different lengths or distorted timeline.



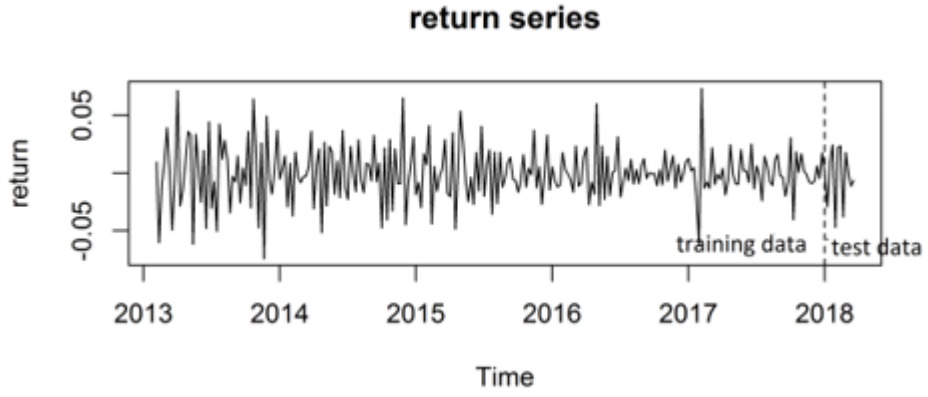


Fig 1. DSEX index and the return series.

To compare the proposed algorithm with standard econometric approach, we conduct a simulation study which consists of the following steps:

(i) The return series of the training period are decomposed with MODWT to obtain approximations and detail coefficients up to level 2.

(ii) Details at each level are modeled with the GARCH equations.

(iii) The residuals of GARCH model are simulated using a Monte Carlo method from either normal, GED or t.

(iv) Estimated GARCH volatility and the simulated random error are used to re-estimate details at each level.

(v) The return series is reconstructed with the new details using the inverse MODWT.

(vi) The reconstructed series is modeled and forecasted with GARCH equation.

(vii) Forecasted volatility is evaluated with referenced to historical EWMA volatility in the test period using some forecasting evaluation criteria.

The simulation outputs of the proposed algorithm for Haar and Symlets wavelet basis and for different error distribution together with standard GARCH results are reported in Table 1. We observed that irrespective of wavelet basis and error distribution, wavelet-GARCH approach produces lower RMSE and DTW distance than those produced by standard GARCH model. Forecast error is much lower when error distribution is considered as t. This is very much natural since most of the financial time series used to have heavier tail than normal. It also should be noted that the prediction accuracy increased if Symlet wavelet basis is used instead of Haar. This support another stylized fact that financial time series possess some asymmetry. The results can be summarized by saying that wavelet-GARCH approach outperforms the standard econometric approach for volatility prediction.

Table 1: Forecast errors of Wavelet-GARCH and GARCH forecasts

			RMSE			DTW distance		
Standard GARCH Forecast			**0.008667**			**0.1528**		
Wavelet-GARCH Simulation								
Wavelet basis	**Error Dist.**	**N**	100	500	1000	100	500	1000
Haar	**Normal**	**mean**	**0.008058**	**0.008043**	**0.008015**	**0.1390**	**0.1364**	**0.1367**
		sd	0.00232	0.00219	0.00227	0.06168	0.05819	0.05845
	GED	**mean**	**0.007886**	**0.007874**	**0.007887**	**0.133332**	**0.13484**	**0.13522**
		sd	0.00236	0.00242	0.00244	0.06222	0.06091	0.06116
	t	**mean**	**0.006251**	**0.005824**	**0.005840**	**0.11593**	**0.1064**	**0.10490**
		sd	0.00278	0.00271	0.00263	0.0593	0.0581	0.0564
Sym8	**Normal**	**mean**	**0.005862**	**0.005786**	**0.005716**	**0.09600**	**0.09589**	**0.09469**
		sd	0.002445	0.002343	0.002368	0.05486	0.05301	0.05301
	GED	**mean**	**0.005691**	**0.00535**	**0.005639**	**0.09715**	**0.09044**	**0.09301**
		sd	0.00250	0.00239	0.00233	0.0555	0.0494	0.05055
	t	**mean**	**0.003734**	**0.00410**	**0.004138**	**0.07015**	**0.074415**	**0.07458**
		sd	0.002431	0.00237	0.00234	0.03997	0.04528	0.04524

4. Conclusion

This study proposed a new algorithm for volatility prediction by incorporating multi-scale information. It is demonstrated that, the wavelet decomposition can be used to obtain the volatility change at different time scale for different level of investor. From the simulation study, it is evident that inclusion of time scale variation can improve the volatility prediction. Use of wavelet transformation in analyzing financial time series is now a day frequently practiced by academicians and business analysts. However, integration of wavelet transformation with GARCH modeling is yet rarely found. Application of this new approach to a wide range of time series data would carry out its credibility and pitfall as well.

References

1. Al Wadi, S., Hamarsheh, A., & Alwadi, H. (2013). Maximum overlapping discrete wavelet transform in forecasting banking sector. *Applied Mathematical Sciences*, *7*(80), 3995-4002.
2. Alves, D. K., Neto, C. M. S., Costa, F. B., & Ribeiro, R. L. A. (2014, December). Power measurement using the maximal overlap discrete wavelet transform. In *Industry Applications (INDUSCON), 2014 11th IEEE/IAS International Conference on* (pp. 1-7). IEEE.
3. Gallegati, M., Ramsey, J. B., & Semmler, W. (2014). Interest rate spreads and output: A time scale decomposition analysis using wavelets. *Computational Statistics & Data Analysis*, *76*, 283-290.
4. Reboredo, J. C., & Rivera-Castro, M. A. (2014). Wavelet-based evidence of the impact of oil prices on stock returns. *International Review of Economics & Finance*, *29*, 145-176.

5. Percival, D. B., & Walden, A. T. (2000). Wavelet methods for time series analysis (Cambridge Series in Statistical and Probabilistic Mathematics).

Stochastic evolutionary system on multidimensional lattices

Elena Yarovaya
Lomonosov Moscow State University, Moscow, Russia

Abstract

For the study of stochastic evolution of particle systems on a non-compact phase space we apply an approach focused on continuous-time branching random walks on multidimensional lattices. The main object of study is the limit distribution of particles on the lattice. Special attention is paid to branching random walks with large deviations. The limit theorems on asymptotic behavior of the Green function for transition probabilities were established for random walks with both a finite and infinite variance of jumps. The obtained results allow to study the front of branching random walk and the structure of the particle population inside of the front and near to its boundary. For supercritical branching random walks, it is shown that the amount of positive eigenvalues of the evolutionary operator, counting their multiplicity, does not exceed the amount of branching sources on the lattice, while the maximal of these eigenvalues is always simple. We demonstrate that the appearance of multiple lower eigenvalues in the spectrum of the evolutionary operator can be caused by a kind of 'symmetry' in the spatial configuration of branching sources. The presented results are based on Green's function representation of transition probabilities of an underlying random walk and cover not only the case of the finite variance of jumps but also a less studied case of infinite variance of jumps.

Keywords

branching random walks; limit distributions of particles; the spectrum of the evolutionary operator; Green's function; large deviations.

1. Introduction

We offer to use models of branching random walks (BRWs) for a study of the dynamics of stochastic lattice systems. Continuous-time BRWs on multidimensional lattices provide an important example of stochastic multicompartment systems in which the evolutionary processes depend on the spatial dynamics and the structure of a medium. The dynamics of such processes is usually described in terms of birth, death and walks of particles on the lattice $\mathbb{Z}^d, d \geq 1$. The structure of a medium is defined by the particle offspring reproduction law at the lattice points called branching sources. Such a description covers various applications of BRWs (Zel'dovich et al., 1988; Cranston et al., 2009; Ermakova et al., 2019).

One of the principal problems in BRW models is a study of the evolution of the field of particles on the entire lattice. The methods of the spectral theory of operators with multipoint perturbations, see (Yarovaya, 2017a), will be applied to the study of its evolution. It will be developed the results for weakly supercritical BRWs on $\mathbb{Z}^d, d \geq 1$, obtained in (Yarovaya, 2017c). The methods of the theory of large deviations for BRWs, see (Molchanov and Yarovaya, 2013; Agbor et al., 2015), will be used to study the distribution of population inside the front of propagation of the weekly supercritical BRW. In the frame of the proposed models, it will be undertaken the spatio-temporal analysis of the system. These results help to analyze the distribution of the population inside the propagation front of particles for a weekly supercritical BRW.

2. Methodology

A BRW is a stochastic process combining in itself random walk of particles on $\mathbb{Z}^d$ with their branching at some lattice points of $\mathbb{Z}^d, d \geq 1$. A random walk of particles on $\mathbb{Z}^d$ is defined in terms of the matrix of transition intensities $A = a(x,y))_{x,y\in\mathbb{Z}^d,}$ which features the regularity property: $\sum_{y\in\mathbb{Z}^d} a(x,y) = 0$ for all x, where $a(x,y) \geq 0$ for $x \neq y$ and $a(x,x) < 0$. We assume that the intensities $a(x,y)$ are symmetric and spatially homogeneous; that is, $a(y-x) \coloneqq a(x,y) = a(y,x) = a(0,y-x)$ and a random walk is irreducible: for each $z \in \mathbb{Z}^d$ there exists a set of vectors $z_1, \dots, z_k, \in \mathbb{Z}^d$ such that $z = \sum_{i=1}^{k} z_i$ and . $a(z_i) \neq 0$ for $i = 1, \dots, k$. Birth and death of particles may occur at some points of the lattice $x_1, \dots, x_N$. The branching mechanism at each source $x_i, i = 1, \dots, N$, is controlled by a Galton Watson continuous-time process, which is defined by the infinitesimal generating function $f(u, x_i) = \sum_{n=0}^{\infty} b_n(x_i)u^n, 0 \leq u \leq 1$, where $b_n(x_i) \geq 0$ for $n \neq 1, b_1(x_i) < 0$ and $\sum_n b_n(x_i) = 0$. It is assumed that each of the particles evolves independently of the rest of particles. We note that the condition for finiteness of all moments, that is, $f^{(r)}(1, x_i) < \infty$ for all $r \in \mathbb{N}$ is essentially used in some proofs of the limit theorems on behavior of the numbers of particles in BRW (see, for example, (Yarovaya, 2007)). Put $\beta_i \coloneqq f^{(r)}(1, x_i)$ for every x_i.

In the BRW models (Yarovaya, 2012), multipoint perturbations of the generator of symmetrical random walk .0' arise which in the case of identical intensity of the sources are given by

$$\mathscr{H}_{\beta_1,\dots,\beta_N} = \mathscr{A} + \sum_{i=1}^{N} \beta_i \Delta_{x_i}. \tag{1}$$

where $x_i \in \mathbb{Z}^d, \mathcal{A}: l^p(\mathbb{Z}^d) \to l^p(\mathbb{Z}^d), p \in [1, \infty]$, is a symmetrical operator generated by the matrix A and obeying the formula

$$(\mathscr{A}u)(z) := \sum_{z'\in\mathbb{Z}^d} a(z - z')u(z'),$$

$\Delta_x = \delta_x \delta_x^T$, and $\delta_x = \delta_x(\cdot)$ denotes the column vector on the lattice assuming unit value at the point x and zero value at the rest of points. The perturbation $\sum_{i=1}^{N} \beta_i \Delta_{x_i}$ of the linear operator $\mathcal{A}$ may give rise to occurrence in the spectrum of the operator $\mathcal{H}_{\beta_1,\dots,\beta_N}$ of positive eigenvalues, the number of such eigenvalues not exceeding the number of the summands N in the last sum counted with their multiplicity (Yarovaya, 2012; Yarovaya, 2017b).

The multipoint perturbations of the generator of symmetrical random walk $\mathcal{A}$ like (1) occur in the operator equations for the moments of particle numbers. For example, let $u_t(y)$ be the number of particles at the time instant t at the point y. Then, the condition that at the initial time instant $t = 0$ the system consists of a single particle situated at the point x is equivalent to the equality $u_0(y) = \delta(x - y)$. At that, the total number of particles on the lattice obeys the equality $u_t = \sum_{y \in \mathbb{Z}^d} u_t(y)$. Denote by $m_1(t, x, y) = E_x\big(u_t(y)\big)$ the expectation of the number of particles at the time instant t at the point y, provided that $u_0(y) \equiv \delta(x - y)$., that is, at the initial time instant the system had one particle at the point x. As was shown in (Yarovaya, 2012; Yarovaya, 2013), the evolution of $m_1(t, x, y)$ obeys the operator equation in the space $l^2(\mathbb{Z}^d)$:

$$\frac{d\, m_1(t,x,y)}{d\, t} = (\mathscr{H}_{\beta_1,\dots,\beta_N} m_1(t,\cdot,y))(x), \qquad m_1(0,x,y) = \delta(x-y).$$

Evolution of the mean number of particles $m_1(t, x) = E_x\big(u_t(y)\big)$ (total size of the population) over the entire lattice (see, for example, (Yarovaya, 2007)) satisfies the operator equation in the corresponding space $l^\infty(\mathbb{Z}^d)$:

$$\frac{d\, m_1(t,x)}{d\, t} = (\mathscr{H}_{\beta_1,\dots,\beta_N} m_1(t,\cdot))(x), \qquad m_1(0,x) = 1.$$

Now we notice that the issue of the rate of growth or decrease of the mean number of particles $m_1(t, x, y)$ is tightly bound to the spectral properties of the operator $\mathcal{H}_\beta$. For example, if the operator $\mathcal{H}_\beta$ has the maximal eigenvalue $\lambda > 0$, then $m_1(t, x, y)$ grows at infinity as $e^{\lambda t}$, see (Khristolyubov and Yarovaya, 2019).

Theorem 1 *Let $\beta_i > 0$ for $i = 1,2,\dots,N$ and the operator H has an isolated eigenvalue $\lambda_0 > 0$. Moreover, the remaining part of its spectrum be located on the halfline $\{\lambda \in R : \lambda \le \lambda_0 - \epsilon\}$, where $\epsilon > 0$. If $\beta_i^{(r)} = O(r!\, r^{r-1})$ for $i = 1,2,\dots,N$ and $r \in \mathbb{N}$, then in the sense of convergence in distribution the following statements hold:*

$$\lim_{t\to\infty} \mu_t(y) e^{-\lambda_0 t} = \psi(y)\xi, \quad \lim_{t\to\infty} \mu_t e^{-\lambda_0 t} = \xi,$$

where $\psi(y)$ is a non-negative non-random function and ξ is a proper random variable.

The behavior of processes whose parameters are situated near the critical point is crucial for many applications. Previously, such problems were considered for branching processes in non-random and random environments, see (Limnios and Yarovaya, 2019) and the bibliography therein. The concept of weakly supercritical BRW was introduced in (Yarovaya, 2015) for the equal intensities $\beta := \beta_1 = \cdots = \beta_N$.

Definition 1 *If there exists* $\varepsilon_0 > 0$ *such that for* $\beta \in (\beta_c, \beta_c + \varepsilon_0)$ *the operator* $\mathcal{H}_\beta$ *has one (counting multiplicity) positive eigenvalue* $\lambda_0(\beta)$ *satisfying the condition* $\lambda_0(\beta) \to 0$ *for* $\beta \downarrow \beta_c$, *then the supercritical BRW is called weakly supercritical for* β *close to* β_c.

As was established in (Yarovaya, 2015; Yarovaya, 2017b), for $\beta \downarrow \beta_c$ each supercritical BRW is weakly supercritical. Denote now by $p(t,x,y)$ the transition probability of the random walk. Clearly, the function $p(t,x,y)$ is determined by the transition intensities $a(x,y)$ (see, for example, (Gikhman and Skorokhod, 2004; Yarovaya, 2007)). Then the Green function of the operator $\mathcal{A}$ is representable as the Laplace transform of the transition probability $p(t,x,y)$:

$$G_\lambda(x,y) := \int_0^\infty e^{-\lambda t} p(t,x,y)\, dt, \qquad \lambda \geq 0. \tag{2}$$

Of special interest for the weakly supercritical BRWs are the asymptotics of the Green function (2) and the eigenvalue $\lambda_0(\beta)$ for the evolutionary operator (1) for $\beta \downarrow \beta_c$, that is, for$\beta \to \beta_c, \beta > \beta_c$.

3. Results

In this paper, we generalize the notion of a weakly supercritical BRW for unequal intensities.

Definition 2 If there exists $\varepsilon_0 > 0$ and a set $(\beta_1, \beta_2, \ldots, \beta_N)$ of the branching source intensities such that for $\beta_1 \in (\beta_{c1}, \beta_{c1}+\varepsilon_0)$, $\beta_2 \in (\beta_{c2}, \beta_{c2}+\varepsilon_0), \ldots, \beta_N \in (\beta_{cN}, \beta_{cN} + \varepsilon_0)$ the operator $\mathcal{H}_{\beta_1,\beta_2\ldots,\beta_N}$ has at least one (counting multiplicity) positive eigenvalue $\lambda_0(\beta_1, \beta_2, \ldots, \beta_N)$ satisfying the condition $\lambda_0(\beta_1, \beta_2, \ldots, \beta_N) \to 0$ for $\beta_i \downarrow \beta_{c,i}$, $i = 1, 2, \ldots, N$, then the supercritical BRW is called weakly supercritical for the branching source intensities $(\beta_1, \beta_2, \ldots, \beta_N)$ close to $(\beta_{c1}, \beta_{c2}, \ldots, \beta_{cN})$.

One can obtain that the theorems for the Green function (2), see (Yarovaya, 2017c), remain valid. For studying (2) the key role plays the asymptotic behavior of transition probabilities $p(t,x,y)$ for underlying random walk based on the properties of $a(z), z \in \mathbb{Z}^d$. As was shown in (Molchanov and Yarovaya, 2012), the following assertion for $G_\lambda := G_\lambda(0, 0)$ is valid under the condition of a finite variance of underlying random walk jumps.

Theorem 2 *Let the condition*

$$\sum_z |z|^2 a(z) < \infty \tag{3}$$

be valid. If $\lambda \downarrow 0$*, then the following asymptotic equalities take place:*

- $G_\lambda \sim \gamma_1 \sqrt{\pi} (\sqrt{\lambda})^{-1}$ *for* $d = 1$,
- $G_\lambda \sim -\gamma_2 \ln \lambda$ *for* $d = 2$,
- $G_\lambda - G_0 \sim -2\sqrt{\pi}\gamma_3 \sqrt{\lambda}$ *for* $d = 3$,
- $G_\lambda - G_0 \sim \gamma_4 \lambda \ln \lambda$ *for* $d = 4$,
- $G_\lambda - G_0 \sim -\gamma_d \lambda$ *for* $d \geq 5$,

where γ_i, $i \in N$*, are some positive constants.*

The next theorem, see (Yarovaya, 2017c), is valid under the condition of infinite variance of underlying random walk jumps.

Theorem 3 *Let the condition*

$$a(z) \sim \frac{H\left(\frac{z}{|z|}\right)}{|z|^{d+\alpha}}, \quad \alpha \in (0,2), \tag{4}$$

be valid. If $\lambda \downarrow 0$*, then there are the following asymptotic equalities:*

- $G_\lambda \sim \gamma_{d,\alpha} \lambda^{\frac{1-\alpha}{\alpha}}$ *for* $d = 1$, $\alpha \in (1,2)$,
- $G_\lambda \sim -\gamma_{d,\alpha} \ln \lambda$ *for* $d = 1$, $\alpha = 1$,
- $G_\lambda - G_0 \sim -\gamma_{d,\alpha} \lambda^{\frac{d-\alpha}{\alpha}}$ *for* $d = 1$, $\alpha \in (\frac{1}{2}, 1)$ *or* $d = 2$, $\alpha \in (1,2)$ *or* $d = 3$, $\alpha \in (\frac{3}{2}, 2)$,
- $G_\lambda - G_0 \sim \gamma_{d,\alpha} \lambda \ln \lambda$ *for* $d = 1$, $\alpha = \frac{1}{2}$ *or* $d = 2$, $\alpha = 1$ *or* $d = 3$, $\alpha = \frac{3}{2}$,
- $G_\lambda - G_0 \sim -\gamma_{d,\alpha} \lambda$ *for* $d = 1$, $\alpha \in (0, \frac{1}{2})$ *or* $d = 2$, $\alpha \in (0,1)$ *or* $d = 3$, $\alpha \in (0, \frac{3}{2})$ *or* $d \geq 4$, $\alpha \in (0,2)$,

where $\gamma_{d,a}$ *is some positive constant for each dimension* d *of the lattice* $\mathbb{Z}^d$*.*

Let $\beta_i > 0$ for $i = 1, 2, \ldots, N$ and the operator $\mathcal{H}_{\beta_1, \beta_2 \ldots, \beta_N}$ has a finite number of positive eigenvalues. We denote the largest of them by λ_0, and the corresponding normalized vector by f. Then for all $n \in N$ and $t \to \infty$, see (Khristolyubov and Yarovaya, 2019), the limit statements hold:

$$m_1(t,x,y) = C_1(x,y) e^{\lambda_0 t}(1 + o(e^{\lambda_0 t})), \quad m_1(t,x) = C_1(x) e^{\lambda_0 t}(1 + o(e^{\lambda_0 t})),$$

where

$$C_1(x,y) = f(y)f(x), \qquad C_1(x) = f(x)\frac{1}{\lambda_0}\sum_{j=1}^N \beta_j f(x_j), \quad f(x) = \sum_{j=1}^N \beta_j f(x_j) G_\lambda(x_j - x, 0), \qquad i = 1, \ldots, N.$$

For $x = 0$ we have

$$m_1(t,0,y) = C_1(0,y)e^{\lambda_0 t}(1+o(e^{\lambda_0 t})), \quad m_1(t,0) = C_1(0)e^{\lambda_0 t}(1+o(e^{\lambda_0 t})), \tag{5}$$

where

$$C_1(0,y) = f(y)f(0), \qquad C_1(0) = f(0)\frac{1}{\lambda_0}\sum_{j=1}^{N}\beta_j f(x_j), \quad f(0) = \sum_{j=1}^{N}\beta_j f(x_j)G_\lambda(x_j,0), \qquad i = 1,\dots,N.$$

The case of weekly supercritical BRWs with a few branching sources of various intensities is quite complicated. That is why below we consider the case of one source of branching, that is, $N = 1$. If $N = 1$ and the spectrum $\sigma_d(\mathcal{H}_{\beta_1})$ of the operator $\mathcal{H}_{\beta_1}$ contains for $\beta_1 \downarrow \beta_c$ a leading eigenvalue $\lambda_0(\beta_1) \to 0$, the asymptotic behavior of $\lambda_0(\beta_1)$, as $\beta_1 \downarrow \beta_c$, see (Yarovaya, 2017c), has the following form.

Theorem 4 *Under Condition* (3) *the eigenvalue* $\lambda_0(\beta)$ *of the operator* H_β *for* $\beta \downarrow \beta_c$ *has the following asymptotic behavior:*

(i) $\lambda_0(\beta) \sim c_1\beta^2$ *for* $d = 1$,
(ii) $\lambda_0(\beta) \sim e^{-c2/\beta}$ *for* $d = 2$,
(iii) $\lambda_0(\beta) \sim c_3(\beta - \beta_c)^2$ *for* $d = 3$,
(iv) $\lambda_0(\beta) \sim c_4(\beta - \beta_c)\ln^{-1}((\beta - \beta_c)^{-1})$ *for* $d = 4$,
(v) $\lambda_0(\beta) \sim c_d(\beta - \beta_c)$ *for* $d \geq 5$,

where c_i, $i \in N$, *are some positive constants.*

Theorem 5 *Under Condition* (4) *the eigenvalue* $\lambda_0(\beta)$ *of the operator* H_β *for* $\beta \downarrow \beta_c$ *has the following asymptotic behavior:*

(i) $\lambda_0(\beta) \sim c_{d,\alpha}(N\beta)^{\alpha/\alpha-1}$ $for\ d = 1, \alpha \in (1,2)$,
(ii) $\lambda_0(\beta) \sim e^{-cd,\alpha/(N\beta)}$ $for\ d = 1, \alpha = 1$,
(iii) $\lambda_0(\beta) \sim c_{d,\alpha}(\beta - \beta c)^{\alpha/d-a}$ $for\ d = 1, \alpha \in (1/2,1)\ or\ d = 2, \alpha \in (1,2)\ or\ d = 3, \alpha \in (3/2,2)$,
(iv) $\lambda_0(\beta) \sim e^{W(-cd,\alpha(\beta-\beta c))}$ $for\ d = 1, \alpha = 21\ or\ d = 2, \alpha = 1\ or\ d = 3, \alpha = 3/2$,
(v) $\lambda_0(\beta) \sim c_{d,\alpha}(\beta - \beta_c)$ $for\ d = 1, \alpha \in (0,1/2)\ or\ d = 2, \alpha \in (0,1)\ or\ d = 3, \alpha \in (0,3/2)\ or\ d \geq 4, \alpha \in (0,2)$,

where $c_{d,a}$ *is some positive constant (for each fixed values of the parameter* α *and dimension* d *of the lattice* $\mathbb{Z}^d$*), and* $W(x)$ *is the lower branch of the Lambert W -function satisfying the condition* $W(x) \to -\infty$ *for* $x \uparrow 0$.

Based on (Yarovaya, 2007), for $N = 1$ and $\theta > \theta_c$ we get that equations (5) have the form

$$m_1(t,0,y,\beta) \sim C_1(0,y,\beta)e^{\lambda_0(\beta)t}, \quad m_1(t,0) \sim C_1(0,\beta)e^{\lambda_0(\beta)t}, \quad t \to \infty, \tag{6}$$

where

$$C_1(0,y,\beta) = \frac{G_{\lambda_0(\beta)}(0,0)\,G_{\lambda_0(\beta)}(0,y)}{\|G_{\lambda_0(\beta)}(0,y)\|^2}, \; C_1(0,\beta) = \frac{G_{\lambda_0(\beta)}(0,0)}{\lambda_0(\beta)\,\|G_{\lambda_0(\beta)}(0,y)\|^2}.$$

4. Discussion and Conclusion

Based on Theorems 2–5 we can derive some assertions for a weekly supercritical BRW under the conditions (3) or (4). Note that the related proofs for a weekly supercritical BRW are essentially based on how the higher terms of the asymptotic representations (6) depend on β.

Further in the section, we shortly discuss the results on the structure of the population inside the propagating front for recurrent underlying random walk, that is, when $G_0 = \infty$, where by the *population front* we mean the set

$$\Gamma t = \{y : m_1(t, 0, y) \leq C\}.$$

The following theorem gives the description of the population inside the front and near to its boundary for $d = 1$ and $d = 2$ for $\beta := \beta_1 = \cdots = \beta_N$. It was shown in (Yarovaya, 2017a) that there exists $\epsilon_0 > 0$ such that for $\beta \in (\beta_c, \beta_c + \epsilon_0)$ the operator H_β has a unique eigenvalue $\lambda_0(\beta)$.

Theorem 6 *Let $\varepsilon_0 > 0$, $\beta_c < 0 < \beta_c + \varepsilon_0$, and $x \in \mathbf{Z}^d$, $d = 1$ or $d = 2$. If for some $c_1, c_2 > 0$ we have $c_1 t \leq |y| \leq c_2 t$, as $t \to \infty$, and then*

$$\frac{\mu_t(x)}{m_1(t, 0, y)} \xrightarrow{law} \mu_\infty,$$

where the distribution of μ_∞ is independent on x and obeyed the relation $P\{\mu_\infty > 0\}$. Moreover,

$$\frac{\mu_t}{m_1(t, 0)} \xrightarrow{law} \mu_\star.$$

References

1. Agbor, A., Molchanov, S., and Vainberg, B. (2015). Global limit theorems on the convergence of multidi-mensional random walks to stable processes. Stoch. Dyn., 15(3):1550024, 14.
2. Cranston, M., Koralov, L., Molchanov, S., and Vainberg, B. (2009). Continuous model for homopolymers. J. Funct. Anal., 256(8):2656–2696.
3. Ermakova, E., Makhmutova, P., and Yarovaya, E. (2019). Branching random walks and their applications for epidemic modeling. Stochastic Models, 1–18, DOI: 10.1080/15326349.2019.1572519.
4. Gikhman, I. and Skorokhod, A. (2004). The theory of stochastic processes. II. Classics in Mathematics. Springer-Verlag, Berlin. Translated from the Russian by S. Kotz, Reprint of the 1975 edition.
5. Khristolyubov, I. and Yarovaya, E. (2019). A limit theorem for supercritical branching random walks with branching sources of varying intensity. ArXiv.org e-Print archive.

6. Limnios, N. and Yarovaya, E. (2019). Diffusion approximation of near critical branching processes in fixed and random environment. Stochastic Models, 1–12, DOI: 10.1080/15326349.2019.1578240.
7. Molchanov, S. and Yarovaya, E. (2012). Branching processes with lattice spatial dynamics and a finite set of particle generation centers. Dokl. Akad. Nauk, 446(3):259–262.
8. Molchanov, S. and Yarovaya, E. (2013). Large deviations for a symmetric branching random walk on a mul¬tidimensional lattice. Tr. Mat. Inst. Steklova, 282 (Vetvyashchiesya Protsessy, Sluchainye Bluzhdaniya, i Smezhnye Voprosy):195–211.
9. Yarovaya, E. (2017a). Positive discrete spectrum of the evolutionary operator of supercritical branching walks with heavy tails. Methodol. Comput. Appl. Probab., 19(4):1151–1167.
10. Yarovaya, E. (2007). Branching random walks in a heterogeneous environment. Center of Applied Inves-tigations of the Faculty of Mechanics and Mathematics of the Moscow State University, Moscow. In Russian.
11. Yarovaya, E. (2012). Spectral properties of evolutionary operators in branching random walk models. Math. Notes, 92(1-2):115–131. Translation of Mat. Zametki 92 (2012), no. 1, 123–140.
12. Yarovaya, E. (2013). Branching random walks with several sources. Math. Popul. Stud., 20(1):14–26.
13. Yarovaya, E. (2015). The structure of the positive discrete spectrum of the evolution operator arising in branching random walks. Doklady Mathematics, 92(1):507–510.
14. Yarovaya, E. (2017b). Positive discrete spectrum of the evolutionary operator of supercritical branching walks with heavy tails. Methodology and Computing in Applied Probability, 19(4):1151–1167.
15. Yarovaya, E. (2017c). Spectral asymptotics of a supercritical branching random walk. Teor. Veroyatn. Primen., 62(3):518–541.
16. Zel'dovich, Y., Molchanov, S., Ruzmaĭkin, A., and Sokoloff, D. (1988). Intermittency, diffusion and generation in a nonstationary random medium. In Mathematical physics reviews, Vol. 7, volume 7 of Soviet Sci. Rev. Sect. C Math. Phys. Rev., pages 3–110. Harwood Academic Publ., Chur.

Simulation of branching random walks with different intensity of branching sources

Daria Balashova
Lomonosov Moscow State University, Moscow, Russia

Abstract

It is a common practice to describe branching random walks in terms of birth, death and walk of particles, which makes it easier to use them in different applications. We consider a continuous-time symmetric supercritical branching random walk on a multidimensional lattice with a finite set of particle generation centers, i.e. branching sources. It is useful in different applications such as statistical physics, population dynamics and chemical kinetics. The branching of particles occurs at lattice points called branching sources and is determined by the continuous-time Galton-Watson process. The intensity of the source as the first derivative of infinitesimal generating function of the Galton-Watson process is the quantitative characteristic of the average number of particle descendants that are born in it. Existence of a positive eigenvalue of the evolutionary operator of the average number of particles involves the exponential growth of the first moment of the total number of particles both at an arbitrary point and on the entire lattice. Main attention is paid to the case when sources with positive and negative intensities are in an arbitrary configuration and, as an example, in a simplex. For applied research, behavior at finite time intervals is required. In addition to the limit theorems, the approach based on simulations by the Monte Carlo method is considered in the talk.

Keywords

branching processes; random walks; multidimensional lattices; simulation.

1. Introduction

A branching random walk (BRW) with continuous time on a multidimensional lattice $\mathbb{Z}^d, d \geq 1$, with a finite number of branching sources on it is considered. We assume that the walk is homogeneous in time and space, symmetrical and irreducible. Branching processes are used to describe the population dynamics of objects with non-overlapping generations, for example, to describe the spread of viral infections, see [1], [4], modeling the development of the epidemic and the effects of vaccination, [3] and [7], as well as biological and genetic systems [8]. The introduction of a walk will allow to describe the spatial distribution of such processes.

It is convenient to represent the evolution equations for the transition probabilities and the numbers of particles in the form of linear differential equations in Banach spaces. By virtue of their linearity, the study of the asymptotics of solutions as $t \to \infty$ leads to the study of the spectrum of the corresponding operators. In particular, the presence of a positive eigenvalue in the spectrum of an evolutionary operator ensures an exponential growth in the number of particles both at each point and in the entire lattice [10]. BRW with an exponential increase in the number of particles are called supercritical.

Analysis of the evolutionary operator of BRW with several sources in general was carried out in [10], where in particular it was noted that the presence of branching sources can lead to the appearance of positive operator eigenvalues. In [2] it was proved that for the case of equal source intensities and finite variance of jumps, the number of eigenvalues (including multiplicity) does not exceed the number of sources N and the multiplicity of each eigenvalue does not exceed $N-1$.

In this paper, we consider BRW in which the birth and death of particles can be specified by a subcritical. The sources can be of different intensities, positive where birth prevails over death, and negative where the opposite is true. Analyzing the corresponding equations is analytically quite difficult. Therefore, steps have been taken in modelling such processes, which make it possible to estimate numerically the values for critical boundaries for source intensities.

2. Methodology

We consider BRW on the multidimensional lattice $\mathbb{Z}^d, d \geq 1$, in which branching — birth or death — occurs in the sources $x_1, x_2, \dots, x_N$. We assume that random walk is given by a matrix of transient intensities

$$A = (a(x,y))_{x,y \in \mathbb{Z}^d},$$

with the properties $a(x,y) = a(y,x) = a(0,y-x) = a(y-x)$ for all x and y. Thus, a random walk is symmetric and spatially homogeneous. Moreover, we assume the regularity properties $\sum_{z \in Z^d} a(z)$ and irreducibility are fulfilled, i.e. for all $z \in \mathbb{Z}^d$ there exists a set of vectors $z_1, z_2, \dots, z_k \in \mathbb{Z}^d$ such that $z = \sum_{i=1}^{k} z_i$ and $a(z_i) \neq 0$ for $i = 1,2,\dots,k$.

The transition probability $p(t,\cdot,y)$ is conveniently considered as a function $p(t)$ in $l^2(\mathbb{Z}^d)$, depending on the time t and parameter y. For the time $h \to 0$ the following equalities hold:

$$\begin{aligned} p(h,x,y) &= a(x,y)h + o(h) \text{ for } y \neq x, \\ p(h,x,x) &= 1 + a(x,x)h + o(h). \end{aligned} \tag{1}$$

As is known from [6], the transition probabilities satisfy the system of inverse Kolmogorov equations:

$$\frac{\partial p(t,x,y)}{\partial t} = \sum_{x'} a(x,x')p(t,x',y),\ p(0,x,y) = \delta(x-y),$$

where δ — discrete δ-function of Kronecker on $\mathbb{Z}^d$.
We assume that branching occurs in the sources x_i and is determined by the infinitesimal generating functions

$$f_i(u) = \sum_{n=0}^{\infty} b_{i,n}u^n, \quad 0 \le u \le 1,$$

where $\sum_n b_n(x_i) = 0, b_n(x_i) \ge 0$ for $n \ne 1$ and $b_1(x_i) < 0, f_i^{(r)}(1) < \infty$ for all $r \in \mathbb{N}$. Denote the intensity of the source x_i:

$$\beta_i = f'(1,x_i) = \sum_n n b_n(x_i) = -(b_1(x_i))(\sum_{n\ne 1} n\frac{b_n(x_i)}{-(b_1(x_i))} - 1),$$

characterising the average number of descendants that are born in it.
Let $\mu_t(y)$ — is the number of particles at time t at the point y and $m_1(t,x,y) \coloneqq \mathrm{E}_x\mu_t(y)$ — is the mathematical expectation of the number of particles at the point y at time t under the condition, that at the initial instant of time $t = 0$ there was one particle in the system located at the point x. According to [6],

$$\frac{\partial m_1(t,x,y)}{\partial t} = \sum_{x'} a(x,x')m_1(t,x',y) + \sum_{i=1}^{N} \beta_i\delta(x-x_i)m_1(t,x,y), \qquad (2)$$
$$m_1(0,x,y) = \delta(x-y).$$

On the set of functions $u(x), x \in \mathbb{Z}^d$ we consider the operator

$$(\mathcal{A}u)(x) = \sum_{x'\in\mathbb{Z}^d} a(x-x')u(x')$$

and for each of the sources $x_i \in \mathbb{Z}^d$ the operators

$$(\Delta_{x_i}u)(x) = \delta(x-x_i)u(x),$$

where δ(·) — discrete δ-function of Kronecker on $\mathbb{Z}^d$. The operator A, as an operator in the Hilbert space $l^2(\mathbb{Z}^d)$, is self-adjoint, the operators Δx_i act in each of the function spaces $l^p(\mathbb{Z}^d), p \in [1,\infty]$, see [10].

The behavior of the mean number of particles both at an arbitrary point and on the entire lattice can be described in terms of an evolutionary operator of a special type

$$\mathcal{H}_{\beta_1,\dots,\beta_N} = \mathcal{A} + \sum_{i=1}^{N} \beta_i \Delta_{x_i}, \quad x_i \in \mathbb{Z}^d, \tag{3}$$

which is a perturbation of the generator A of symmetric random walk. This operator can be treated as a linear bounded operator acting in each of the function spaces spaces $l^p(\mathbb{Z}^d), p \in [1,\infty]$, see [10].

According to [10] the evolution equations for the transition probabilities (1) and the moments of particle numbers (2) can be represented as the following differential equation in the space $l^2(\mathbb{Z}^d)$ and spaces $l^p(\mathbb{Z}^d), p \in [1,\infty]$, respectively:

$$\frac{dp(t,x,y)}{dt} = (\mathcal{A}p(t,\cdot,y))(x), \quad p(0,x,y) = \delta(x-y),$$
$$\frac{dm_1(t,x,y)}{dt} = (\mathcal{H}_{\beta_1,\dots,\beta_N} m_1(t,\cdot,y))(x), \quad m_1(0,x,y) = \delta(x-y).$$

The Green's function of the operator A can be represented as the Laplace transform of the transition probability $p(t,x,y)$:

$$G_\lambda(x,y) := \int_0^\infty e^{-\lambda t} p(t,x,y)dt = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \frac{e^{i(\theta,y-x)}}{\lambda - \phi(\theta)} d\theta, \quad \lambda \geq 0,$$

where $\emptyset(\theta) = \sum_{z\in Z^d} a(z)\, e^{i(\theta,z)}$ for $\theta \in [-\pi,\pi]$. For further research, the value $G_\lambda := G_\lambda(0,0)$ plays an important role. If inequality

$$\sum_{z\in\mathbb{Z}^d} |z|^2 a(z) < \infty,$$

where $|z|$ is the Euclidean norm of the vector z, is fulfilled then the variance of jumps is finite and $G_0 < \infty$ for $d \geq 3$ [11].

We are interested in the description of the behavior of particles on $\mathbb{Z}^d$ in terms of the total number of particles $m(t,x) = \sum_{y\in Z^d} m_1(t,x,y)$ on the lattice.

3. Results

We consider the BRW with p positive intensity sources $\beta > 0$ and n negative intensity sources $(-\beta) < 0$, located at the vertices of the simplex, $|x_i - x_j| = const$ for $i \neq j$.

In this section, we consider an algorithm similar to the one described in [5]. We call the state of a BRW system the set of pairs $\{x, t\}$, each of which corresponds to a particle located at the point $x \in \mathbb{Z}^d$, that appeared at the point x at the time t. By evolution we mean a jump to another point, splitting or death of a particle. In the process of modelling the transition from one state of the BRW system to another will be carried out by excluding one pair from the set of states and adding to the seat of states of one or several pairs corresponding to the result of the simulated particle evolution.

Initialization. First, we set the characteristics of the simulated BRW: choose the dimension d of the integer lattice, the functions defining the distribution of the jumps matrix A, location of sources, their intensities and the execution time T. At the initial moment of time, the state of the system is determined by the presence of a single particle at zero coordinate point of a given space.

Step of algorithm. We model the evolution of the particle: the exponential time dt of staying it in x, then the jump or birth/death event. In the case of a jump, the transition state is simulated according to the matrix A, in the next state the current pair $\{x, t\}$ disappears and appears new pair $\{x', t + d\}$. In the case of death, the current pair $\{x, t\}$ disappears, and in the case of birth (dividing into several offsprings) the current pair $\{x, t\}$ disappears and two new pairs $\{x, t + dt\}$ are added to the set. After some time we may have several pairs {x, t} waiting to be processed. We select an arbitrary pair $\{x, t\}$ (due to the independence of the particles) and model the evolution of the corresponding particle.

Stop condition. The algorithm terminates when all the values of t of all pairs in the system exceed the specified time T, or when the set of states becomes empty (the process has degenerated).

Finally, we count the total of particles at each time point from 0 to T. After a certain number of runs (determined from statistical considerations) of a simulation program with the same BRW parameters, the collected information is processed by the Monte Carlo method.

Fig. 1 corresponds to the simulations on $\mathbb{Z}^3$ with 3 sources (1,0,0), (0,1,0) and (0,0,1). In cases of $\beta_1 = \beta_2 = \beta_3 = 0.4$ and $\beta_1 = \beta_2 = 0.5$, $\beta_3 = -0.5$ we can observe subcritical processes. While in cases of $\beta_1 = \beta_2 = \beta_3 = 0.5$ and $\beta_1 = \beta_2 = 0.6$, $\beta_3 = -0.6$ the processes are supercritical.

The graphs corresponding to the simulations on $\mathbb{Z}^4$ with 4 sources (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1) are shown in the fig. 2. The processes are subcritical for $\beta_1 = \beta_2 = 0.7$, $\beta_3 = \beta_4 = -0.7$ and $\beta_1 = 0.8$, $\beta_2 = \beta_3 = \beta_4 = -0.8$. In cases of $\beta_1 = \beta_2 = 0.8$, $\beta_3 = \beta_4 = -0.8$ and $\beta_1 = 0.9$, $\beta_2 = \beta_3 = \beta_4 = -0.9$ we can observe supercritical processes.

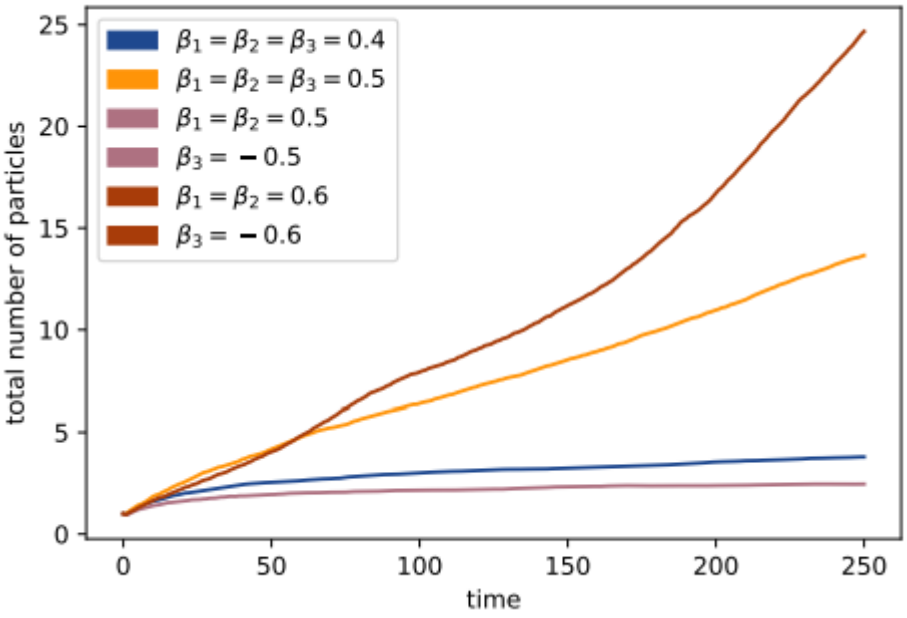


Figure 1: Mean total number of particles on $\mathbb{Z}^3$

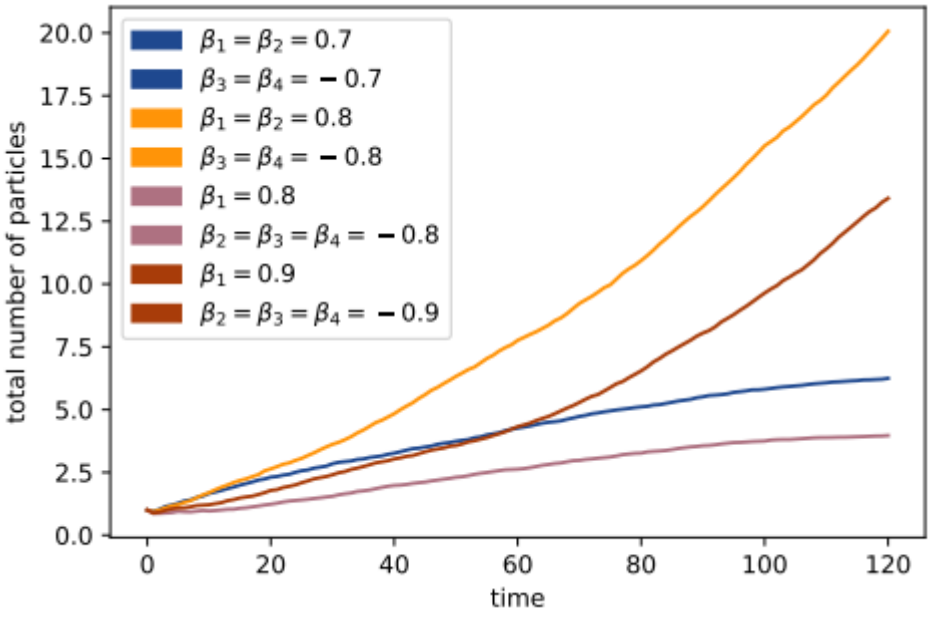


Figure 2: Mean total number of particles on $\mathbb{Z}^4$

4. Discussion and Conclusion

Denote $\tilde{G}_\lambda := G_\lambda(x_i, x_j) = G_\lambda(0, |x_i - x_j|)$ for $i \neq j$. Sources with positive intensities indicate points where the birth rate prevails over the degree of death and in sources with negative intensity — on the contrary. The number of eigenvalues $\lambda > 0$ counting their multiplicity of the evolution operator $\mathcal{H}_{\beta_1,\dots,\beta_{p+n}}$, where $\beta_1 = \dots = \beta_p = \beta$ and $\beta_{p+1} = \dots = \beta_{p+n} = -\beta$ does not

exceed the number of branching sources with positive intensity, the maximum of these eigenvalues is simple.

In addition, we can calculate critical values β_c for the intensity of sources, depending on which process will be subcritical, critical or supercritical:

$$\beta_c = \frac{(n-p)\tilde{G}_0 + \sqrt{(n-p)^2(\tilde{G}_0)^2 + 4D_0}}{2D_0}, \text{ where}$$
$$D_\lambda := (G_\lambda - \tilde{G}_\lambda)(G_\lambda + \tilde{G}_\lambda(n+p-1)).$$

Consider BRW on $\mathbb{Z}^3$ with 3 sources: (1,0,0), (0,1,0) and (0,0,1).
We find the values G_0 = 1.514 and $\tilde{G}_0 := G_0((1, 0, 0), (0, 1, 0)) =\ldots = G_0((0, 1,0),(0,0,1)) = 0.328$. Let β_c is a smaller positive root of the equation

$$\beta^3(3\tilde{G}_0^2 G_0 - 2\tilde{G}_0^3 - G_0^3) + 3\beta^2(G_0^2 - \tilde{G}_0^2) - 3\beta G_0 + 1 = 0,$$

for

$$\beta_1 = \beta_2 = \beta_3 > \beta_c = 0.461$$

and for

$$\beta_1 = \beta_2 = -\beta_3 > \beta_c = \frac{\sqrt{\tilde{G}_0^2 + 4(G_0 - \tilde{G}_0)(G_0 + 2\tilde{G}_0)} - \tilde{G}_0}{2(G_0 - \tilde{G}_0)(G_0 + 2\tilde{G}_0)} = 0.563$$

the $BRWs$ are supercritical.
Consider BRW on $\mathbb{Z}^4$ with 4 sources: (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1).
We find the values G_0 = 1.234 and $\tilde{G}_0 := G_0((1, 0, 0, 0), (0, 1, 0, 0)) =\ldots =$ $G_0((0, 0, 1, 0), (0, 0, 0, 1)) = 0.102$.
The BRW is supercritical for

$$\beta_1 = \beta_2 = -\beta_3 = -\beta_4 > \beta_c = \frac{1}{\sqrt{(G_0 - \tilde{G}_0)(G_0 + 3\tilde{G}_0)}} = 0.758$$

and for

$$\beta_1 = -\beta_2 = -\beta_3 = -\beta_4 > \beta_c = \frac{2\tilde{G}_0 + \sqrt{4\tilde{G}_0^2 + 4(G_0 - \tilde{G}_0)(G_0 + 3\tilde{G}_0)}}{2(G_0 - \tilde{G}_0)(G_0 + 3\tilde{G}_0)} = 0.838.$$

The research was supported by the $RFFR$, project no. 17-01-00468.

References

1. Antonelly F., Bosco F. (2012) Viral evolution and adaptation as a multivariate branching process. Biomat., pp. 217-243.
2. Antonenko E.A., Yarovaya E.B. (2016) On the Number of Positive Eigenvalues of the Evolutionary Operator of Branching Random Walk. Branching Processes and their Applications, Book. Lecture Notes in Statistics, Spriger, vol. 219, pp. 41-55.
3. Ball F., Gonzalez M., Martinez R. and Slavtchova-Bojkova M. (2014) Stochastic monotonicity and continuity properties of functions defined on Crump-Mode-Jagers branching processes, with application to vaccination in epidemic modelling. Bernoulli 20(4), 2076-2101.
4. Claus O. Wilke (2003) Probability of Fixation of an Advantageous Mutant in a Viral Quasispecies. GENETICS vol. 163 no. 2, pp. 467-474.
5. Ermishkina E.M., Yarovaya E.B. (2018) Simulation of Branching Random Walks on Multidimensional Lattices. Abstracts of the Ninth Workshop on Simulation.
6. Gikhman I.I., Skorokhod A.V. (1973) The Theory of Stochastic Processes. Book 2, Science.
7. Gonzalez M., Martinez R. and Slavtchova-Bojkova M. (2010) Stochastic monotonicity and continuity properties of the extinction time of Bellman-Harris branching processes: an application to epidemic modelling. J. App. Prob., 47, pp. 58-71.
8. Kolmogorov A., Petrovskii I. and Piskunov N. (1937) Etude de l'equation de la diffusion avec crois-sance de la quantite de matiere et son application a un probleme biologique. Bull. Univ. Moscou Serie internationale, Section A, Mathematiques et mecanique 1, pp. 1-25.
9. Rytova A.I., Yarovaya E.B. (2016) Multidimensional Watson lemma and its applications. Math Notes 99(3), pp. 395-403.
10. Yarovaya E.B. (2012) Spectral properties of evolutionary operators in branching random walk models. Math Notes 92(1), pp. 115–131.
11. Yarovaya E.B. (2017) Spectral asymptotics of supercritical branching random process. Teor. Veroyatnost. i Primenen., 62(3), pp. 518-541.

Survival analysis of particle populations in branching random walks

Anastasiia Rytova, Elena Yarovaya
Lomonosov Moscow State University, Moscow, Russia

Abstract

Application of the branching random walk models in the population studies is discussed. The main results obtained for the models of symmetric continuous-time branching random walks on a multidimensional lattice with a few sources of particle birth and death at lattice points. We will be mainly interested in studying the problems related to the limiting behavior of branching random walks such as existence of phase transitions under change of various parameters, the properties of the limiting distribution and the survival ability of the particle population. The survival analysis of such particle system is related with the notions of local extension probability of the branching random walk at every lattice point and of the survival probability of the particle population. Emphasis is made on the survival analysis and study of branching random walk properties depending on the configuration of the sources and their intensities. The answer to these and other questions heavily depend on numerous factors which affect the properties of a branching random walk. Therefore, we will try to describe how the properties of a branching random walk depend on such characteristics of an underlying branching walk as finiteness or infiniteness of the variance of jumps.

Keywords

population dynamics; asymptotic behavior; heavy-tailed distribution; survival probability

1. Introduction

The evolution of systems with several elements that able to move, produce descendants and die can be described by a random walk and branching process. The corresponding mathematical models were used to explore the genetic patterns in Haldane (1927), axon growing in Zhizhina, Komech, & Descombes (2015), reliability for servers system in Vatutin, Topchii & Yarovaya (2004), biological populations in Bolker, Pacala, & Neuhauser (2003), human population in Molchanov & Whitmeyer (2017), epidemic spread in Ermakova, Makhmutova, & Yarovaya (2019). It can be interesting to determine the conditions that lead to a special state of the system, such as the degeneration or exponential growth of the population, the form and stability of the spatial distribution. For example, if there is an area in space in which the particle is rather die than give descendants, the strategy increases the population

survival probability may be avoiding this area due to the complication of the return conditions for a particle, such as an increase the dimension of space or an increase in the length of the particle displacement. An example of the relevant model and conditions will be given in sections 3 and 4.

There are many combinations of assumptions on branching random walks (BRW) models: discreteness or continuity of time, discreteness, compactness and dimension of space, number and configuration of branching sources, number and location of initial particles, random walk and branching process properties. In this paper we consider a symmetric continuous time BRWs on the lattices $\mathbb{Z}^d, d \geq 1$ with a finite number of branching sources and a single initial particle. Suppose that particle performs a random walk on points of $\mathbb{Z}^d$ until reaching one of the branching sources, where it can die or give a random number of descendants, and then each one evolves according to the same rules independently of each other. The underlying random walk is assumed to be symmetric, spatially homogeneous and irreducible. As a rule, such models were considered under the assumption, that the variance of random walk jumps is finite, therefore the new effects of infinite variance condition are expected. One of the first works that investigated a model for a simple random walk with one source and pure birth was, apparently, the work Yarovaya (1991). More general cases of symmetric BRWs with finite variance of jumps and one branching source have been studied by many authors (e.g., Albeverio, Bogachev, & Yarovaya (1998)). As was shown in Yarovaya (2013a), rejection of jump variance finiteness assumption leads to new effects for BRW. A number of publications were devoted to random walks with an infinite variance of jumps (see Borovkov & Borovkov (2008) and detailed bibliography therein). In Agbor, Molchanov & Vainberg (2015), for random walks on $\mathbb{Z}^d, d \geq 1$, with heavy tails and appropriate regularity conditions, a global limit theorem on the behavior of the transition probabilities with simultaneous growth of time and lattice coordinates was obtained.

2. Methodology

Let us give a brief description of the model proposed in Yarovaya (2013a). The underlying random walk is specified by a matrix $A = (a(x,y))_{x,y\in\mathbb{Z}^d}$ of transition intensities satisfying the conditions $a(x,y) \geq 0$ if $x \neq y$, $-\infty < a(x,x) < 0, \sum_{y\in\mathbb{Z}^d} a(x,y) = 0$ symmetry and spatially homogeneity $a(x,y) = a(y,x) = a(0,y-x)$. Then the values of $a(x,y)$ can be expressed by the function of one argument $a(z) \coloneqq a(0,z)$. Assume irreducibility of the random walk, which means that for every $z \in \mathbb{Z}^d$, there exists a set of vectors $z_1, \dots, z_k \in \mathbb{Z}^d$ such that $z = \sum_{i=1}^{k} z_i$ and $a(z_i) \neq 0$ for $1 \leq i \leq k$. We denote by σ^2 the variance of jumps of a random walk, then

$$\sigma^2 \coloneqq \sum_{z\neq 0} |z|^2 \frac{a(z)}{-a(0)}. \quad (1)$$

As shown, for example, in Yarovaya (2007), for this model, the probabilities $p(t,x,y)$ of a particle transition from the point x to a point y in time t satisfy the Kolmogorov's backward equation

$$\frac{dp(t,x,y)}{dt} = \sum_{x'} a(x,x')p(t,x,y), \qquad p(0,x,y) = \delta(y-x), \tag{2}$$

where δ(·) is the discrete Kronecker delta-function on $\mathbb{Z}^d$.

Consider the *heavy-tailed* random walk, when the following asymptotic condition as $|z| \to \infty$ on the transition intensities holds

$$a(z) \sim \frac{H(z/|z|)}{|z|^{d+a}}, \tag{3}$$

where $|\cdot|$ is the Euclidean norm on $\boldsymbol{R}^d$, $H(z/|z|) = H(-z/|z|)$ is a positive continuous function on $\boldsymbol{S}^{d-1} = \{z \in \boldsymbol{R}^d : |z| = 1\}$, and $\alpha \in (0,2)$. This implies that the variance of the jumps (1) becomes infinite.

We also assume that branching process is possible at N special points x_1, x_2,..., x_N called *branching sources*, where each particle can die or give a random number of descendants. The reproduction law at the source $x_1, i = 1,2,\dots,N$ is defined by the continuous-time Bienaymé-Galton-Watson branching process by the following infinitesimal generation function

$$f(u,x_i) = \sum_{n=0}^{\infty} b_n(x_i)u^n, \quad 0 \le u \le 1,$$

where $b_n(x_i) \ge 0$ for $n \ne 1$, $b_1(x_i) < 0$ and $\sum_n b_n(x_i) = 0$. We assume $f^{(r)}(1,x_i) < \infty$ for every $r \in N$. For the future investigation an important role will play the values

$$\beta_i = f'(1,x_i) = \sum_n n b_n(x_i) = (-b_1(x_i))\left(\sum_{n\ne 1} n\frac{b_n(x_i)}{(-b_1(x_i))} - 1\right),$$

for $i = 1,2,\dots,N$, called intensity of the branching source x_i, where the last sum is the mean number of descendants born at the point x_i.

The main objects of interest are the behavior of the local particle numbers $\mu_t(y)$ at an arbitrary point $y \in \mathbb{Z}^d$ and the total population size μ_t. We consider their conditional expectation $m(t,x) \coloneqq E_x\mu_t(y)$ under condition that at the initial time there was only one particle in the system, located at the point x. In Yarovaya (2013b), it was shown that the following equations hold

$$\frac{dm(t,x,y)}{dt} = \sum_{x'} a(x,x')m(t,x',y) + \sum_{i=1}^{N} \beta_i\delta(x-x_i)m(t,x,y), \tag{4}$$

$$\frac{dm(t,x)}{dt} = \sum_{x'} a(x,x')m(t,x') + \sum_{i=1}^{N} \beta_i\delta(x-x_i)m(t,x,), \tag{5}$$

where $m(0,x,y) = \delta(x-y), m(0,x) = 1.$. Then it was noted that equations (2), (4), (5) can be treated as the linear differential equations in a Banach space

$$\frac{dp(t,x,y)}{dt} = \big(Ap(t,\cdot,y)\big)(x), \qquad \frac{dm(t,x,y)}{dt} = \big(H_\beta m(t,\cdot,y)\big)(x), \frac{dm(t,x)}{dt} = \big(H_\beta m(t,\cdot)\big)(x) \tag{6}$$

with initial condition $p(0,x,y)=\delta(y-x), m(0,x,y)=\delta(y-x), m(t,x)=1$ respectively, where for $u \in l^p(\mathbb{Z}^d), 1 \le p \le \infty$, the operator $A: l^p(\mathbb{Z}^d) \to l^p(\mathbb{Z}^d)$ is as follows $A(u)(x)=\sum_{x'} a(x,x')u(x')$, and the operator H_β is specified by the equality $H_\beta = A + \sum_{i=1}^N \beta_i \Delta_{x_i}$, where $\Delta_x = \delta_x \delta_x^T$, and $\delta_x = \delta_x(\cdot)$ denotes the column-vector on the lattice taking a unit value at the point x and vanishing elsewhere. In $l^2(\mathbb{Z}^d)$the operator A is self-adjoint.

To investigate the population survival probability, we denote $Q(t,x) := P_x\{u_t > 0\}$ as the probability of the presence of at least one particle on the lattice. By work Yarovaya (2009), we have the equation for BRW with the single branching source located at the $0 \in \mathbb{Z}^d$, that is valid for BRW regardless of the finiteness variance of jumps

$$Q(t,x) = 1 - \int_0^t p(t-s,x,0) f\big(1-Q(s,0)\big) ds, \qquad (7)$$

where $Q(0,x)=1$.

It is convenient to study the BRW in terms of the so-called Green's function $G_\lambda(x,y) := \int_0^\infty e^{-\lambda t} p(t,x,y)\, dt$, which is the Laplace transform of the random walk transition probability. Following, for example, Yarovaya (2013a), a random walk will be called *recurrent* in the case when $G_0 := G_0(0,0) = \infty$ and *transient* in the case when $G_0 < \infty$. The behavior of a BRW essentially depends on the recurrence property.

We denote by β_c and called *critical (for BRW) intensity* of the branching source the lowest intensity of the source such that for $\beta > \beta c$ the spectrum of the operator H_β contains a positive eigenvalue. In work Albeverio, Bogachev, & Yarovaya (1998), for the BRW with a finite variance of jumps, it was established that the particle numbers, both at each lattice point and on the entire lattice, grow exponentially only for $\beta > \beta c$. In this sense, the BRW with $\beta > \beta c$ can be called *supercritical*, with $\beta = \beta c$ *critical* and with $\beta < \beta c$ *subcritical BRW*. In Yarovaya (2015), it was shown that for BRW with N branching sources, regardless of the assumptions on the variance of jumps, the following relations hold: if $G_0 = \infty$ then $\beta c = 0$ for $N > 1$, and if $G_0 = \infty$ then$\beta c = G_0^{-1}$ for $N = 1$ and $0 < \beta c < G_0^{-1}$ for $N \ge 2$.

For BRW model with a single branching source, the rejection of the variance of jumps finiteness leads to new effects. The properties and asymptotics of the function $p(t,x,y)$ and, as a consequence, the functions $G_\lambda(x,y)$ and $m(t,x,y)$ change qualitatively. In particular, the asymptotics as $t \to \infty$ of transition probability are

$$p(t,x,y) \sim \begin{cases} \gamma d^{t^{-d/2}} & for\ finite\ variane\ random\ walk, \\ h_{\alpha,d} t^{-d/\alpha}, \alpha \in (0,2) & for\ heavy-tailed\ random\ walk, \end{cases}$$

where $\gamma_d, h_{a,d} > 0$ are from Yarovaya (2007), Rytova & Yarovaya (2016) respectively. Therefore for heavy- tailed BRW the relation $G_0 < \infty$ is possible in the dimension $d = 1$ for $\alpha \in (0,1)$, and also in dimensions $d \ge 2$ for $\alpha \in$

$(0,2)$, whereas in the case of a finite variance of jumps, the relation $G_0 < \infty$ is satisfied only in dimensions $d \geq 3$. Then in $\mathbb{Z}$ and $\mathbb{Z}^d$ the βc can be strictly positive, i.e. BRW population growth can be subcritical even under a supercritical branching regime at the source, in contrast to the case of BRW with finite variance. On high-dimensional lattices, the intensity of the branching process at the source is not determines the criticality of the BRW (see Table 1 in Yarovaya (2010)). For example, under condition of jumps variance finiteness, the supercritical branching process at the source in combination with transient random walk on $\mathbb{Z}^d$, $d \geq 3$, can lead to either a supercritical, or a critical or subcritical BRW. If the branching process at the source is subcritical or critical, then on lattices $d \geq 3$only subcritical BRW is possible.

The analysis of such BRW models is implemented by studying the spectrum of operators A , H_β (see Yarovaya (2007), Yarovaya (2012), Khristolyubov, & Yarovaya (2019)), which determines the limiting behavior of solutions of differential equations (6), in particular, the tendency of the norm to a constant and the fact of monotonicity. To find the asymptotics of solutions, the Laplace transform of $p(t,x,y), m(t,x,y), m(t,x), 1 - Q(t,x)$, integral equations and Tauberian theorems (see Ch. XIII in Feller (1971)) are used. As a result, the asymptotics are expressed through the Green's function.

3. Results

For convenience, we introduce the classification of possible combinations of the lattice dimension d and the random walk jump parameter α:

	$d = 1$	$d = 2$	$d = 3$	$d \geq 4$
(a)	$\alpha \in (1,2)$			
(b)	$\alpha = 1$			
(c)	$\alpha \in (1/2,1)$	$\alpha \in (1,2)$	$\alpha \in (3/2,2)$	
(d)	$\alpha = 1/2$	$\alpha = 1$	$\alpha = (3/2)$	
(e)	$\alpha \in (0,1/2)$	$\alpha \in (0,1)$	$\alpha \in (0,3/2)$	$\alpha \in (0,2)$

In Khristolyubov, & Yarovaya (2019), a supercritical symmetric continuous-time BRW on $\mathbb{Z}^d$, $d \geq 1$, with a $N < \infty$ number of particle generation sources of varying positive intensities without any restrictions on the variance of jumps of the underlying random walk has been investigated. In Theorem 7, it was found that if the operator H_β have finite (counting multiplicity) number of positive eigenvalues, and λ_0 is the largest of them with the corresponding normalized vector u, then as $t \to \infty$ the following asymptotic relations hold

$$m(t,x,y) \sim C_1(x,y)e^{\lambda_0 t}, \quad m(t,x) \sim C_1(x)e^{\lambda_0 t}, \tag{8}$$

where $C_1(x,y) = u(y)u(x), C_1(x) = u(x)\lambda_0^{-1}\sum_{j=1}^{n}\beta_j u(x_j)$. As a consequence, $Q(t,x) \sim 1$.

Now describe the asymptotic behavior as $t \to \infty$ of the mean number of particle at the $y \in \mathbb{Z}^d$ point $m(t,x,y)$, of the mean number particles of

population $m(t,x)$ and survival probability $Q(t,x)$ in critical and subcritical BRWs with a single branching source located at the lattice origin.

Theorem 1

Let the branching process is performed at the single point $0 \in \mathbb{Z}^d, d \geq 1$*, and random walk is heavy-tailed under condition* (3)*. For each* $x \in \mathbb{Z}^d$*, as* $t \to \infty$*, the asymptotics of* $m(t,x,0), m(t,x)$ *and* $Q(t,x)$ *of the BRW is expressed in forms*

Random walk	*Branching*	$m(t,x,y)$	$m(t,x)$	$Q(t,x)$
(a)	$\beta = \beta c$	$C_{1,a}(x,y)t^{-1/a}$	$C_{1,a}(x)$	$\tilde{C}_{1,a}(x)t^{(1-a)/(2a)}$
	$\beta < \beta c$	$C_{1,a}(x,y)t^{1/a-2}$	$C_{1,a}(x)t^{1/a-1}$	$\tilde{K}_{1,a}(x)t^{(1-a)/a}$
(b)	$\beta = \beta c$	$C_{1,1}(x,y)t^{-1}$	$C_{1,1}(x)$	$\tilde{C}_{1,1}(x)(ln\ t)^{-1/2}$
	$\beta < \beta c$	$C_{1,1}(x,y)t^{-1}(ln\ t)^{-2}$	$C_{1,1}(x)(ln\ t)^{-1}$	$\tilde{K}_{1,1}(x)(ln\ t)^{-1}$
(c)	$\beta = \beta c$	$C_{d,a}(x,y)t^{d/a-2}$	$C_{d,a}(x)t^{d/a-2}$	$\tilde{C}_{d,a}(x)$
	$\beta < \beta c$	$C_{d,a}(x,y)t^{-d/a}$	$C_{d,a}(x)$	$\tilde{K}_{d,a}(x)$
(d)	$\beta = \beta c$	$C_{d,a}(x,y)(ln\ t)^{-1}$	$C_{d,a}(x)t(ln\ t)^{-1}$	$\tilde{K}_{d,a}(x)$
	$\beta < \beta c$	$C_{d,a}(x,y)t^{-d/a}$	$C_{d,a}(x)$	$\tilde{C}_{d,a}(x)$
(e)	$\beta = \beta c$	$C_{d,a}(x,y)$	$C_{d,a}(x)t$	$\tilde{K}_{d,a}(x)$
	$\beta < \beta c$	$C_{d,a}(x,y)t^{-d/a}$	$C_{d,a}(x)$	$\tilde{C}_{d,a}(x)$

and $C_{d,a}(x,y), C_{d,a}(x), \tilde{C}_{d,a}(x), \tilde{K}_{d,a}(x)$ *are some positive constants.*

4. Discussion and Conclusion

Compare asymptotics as $t \to \infty$ of survival probability $Q(t,x)$ for $\mathbb{Z}^d$, $d \geq 1$, between a BRW with finite variance of jumps and a heavy-tailed BRW:

$\mathbb{Z}^d$	Branching process	BRW with finite variance	Heavy-tailed BRW	
$d = 1$	$\beta = \beta c$	$C_1(x)t^{-1/4}$	$\tilde{C}_{1,a}(x)t^{(1-a)/(2a)}$	$\alpha \in (1,2)$
			$\tilde{C}_{1,1}(x)(ln\ t)^{-1/2}$	$\alpha = 1$
			$\tilde{C}_{1,a}(x)$	$\alpha \in (0,1)$
	$\beta < \beta c$	$K_1(x)t^{-1/2}$	$\tilde{K}_{1,a}(x)t^{(1-a)/a}$	$\alpha \in (1,2)$
			$\tilde{K}_{1,1}(x)(ln\ t)^{-1}$	$\alpha = 1$
			$\tilde{K}_{1,a}(x)$	$\alpha \in (0,1)$
$d = 2$	$\beta = \beta c$	$C_2(x)(\ln t)^{-1/2}$	$\tilde{C}_{2,a}(x)$	$\alpha \in (0,2)$
	$\beta < \beta c$	$K_2(x)(\ln t)^{-1}$	$\tilde{K}_{2,a}(x)$	$\alpha \in (0,2)$
$d \geq 3$	$\beta = \beta c$	$C_d(x)$	$\tilde{C}_{d,a}(x)$	$\alpha \in (0,2)$
	$\beta < \beta c$	$K_d(x)$	$\tilde{K}_{d,a}(x)$	$\alpha \in (0,2)$

where the constants $C_d(x), K_d(x) > 0$ are from Yarovaya (2010) and the constants $\tilde{C}_{d,a}(x)\tilde{K}_{d,a}(x) > 0$ are from Rytova & Yarovaya (2018). As can be seen, in critical and subcritical cases for Z, the random walk tail becomes heavier when the parameter α approaches to zero, and, as a result, the population survival probability becomes higher. For $\mathbb{Z}^2$, the survival probability

of a BRW with a finite variance tends to zero, but of a heavy-tailed BRW tends to a positive constant. For $\mathbb{Z}^d$, $d \geq 3$, the dimension of the lattice is already large enough even for a BRW with a finite variance, in this case the underlying random walk is transient, so that for both BRWs the population has a non-zero probability of survival.

Acknowledgement
This work is supported by the Russian Foundation for Basic Research (grant no. 17-01-0468).

References

1. Agbor, A., Molchanov, S., & Vainberg B. (2015). Global limit Theorems on the convergence of multidimen-sional random walks to stable processes. Stochastics and Dynamics.15(3): 1550024.
2. Albeverio, S., Bogachev, L. V., & Yarovaya E. B. (1998). Asymptotics of branching symmetric random walk on the lattice with a single source. C.R. Acad. Sci. Paris, Ser. I, Math.
3. Bolker, B. M., Pacala, S.W., & Neuhauser, C. (2003). Spatial dynamics in model plant communities: what do we really know? Am. Nat. 162, 135–148.
4. Borovkov, A., & Borovkov, K. (2008). Asymptotic Analysis of Random Walks. Heavy-Tailed Distributions. Cambridge University Press.
5. Ermakova, E., Makhmutova, P., & Yarovaya, E. (2019). Branching random walks and their applications for epidemic modeling. Stochastic Models, DOI: 10.1080/15326349.2019.1572519
6. Feller, W. (1971). An introduction to probability theory and its applications. Vol. II. Second edition. John Wiley & Sons, Inc., New York-London-Sydney.
7. Haldane, J. B. S. (1927). A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. PCPS 23(7): 838–844.
8. Khristolyubov, I., & Yarovaya, E. (2019). A Limit Theorem for Supercritical Branching Random Walks with Branching Sources of Varying Intensity. arXiv.org. URL: https://arxiv.org/pdf/1904.01468.
9. Molchanov, S., & Whitmeyer, J. (2017). Spatial Models of Population Processes. In: Panov V. (eds) Mod¬ern Problems of Stochastic Analysis and Statistics. MPSAS 2016. Springer Proceedings in Mathematics & Statistics, vol 208. Springer, Cham.
10. Rytova, A. I., & Yarovaya, E. B. (2016). Multidimensional Watson lemma and its applications. Mathemati-cal Notes, 99(3): 406–412.
11. Rytova, A., & Yarovaya, E. (2018). Survival Analysis of Particle Populations in Branching Random Walks. arXiv.org. URL: https://arxiv.org/abs/1812.09909.

12. Vatutin, V. A., Topchii,V. A., & Yarovaya, E. B. (2004). Catalytic branching random walk and queueing systems with random number of independent servers. Theor. Probability and Math. Statist. 69: 1–15.
13. Yarovaya, E. B. (1991). Use of spectral methods to study branching processes with diffusion in a noncompact phase space. Theor. Math. Phys.
14. Yarovaya, E.B. (2007) Branching Walks in Heterogeneous Medium, Center Appl. Studies at Moscow State Univ., Dep. Mech. and Math., Moscow, (in Russian).
15. Yarovaya, E. B. (2009). Critical branching random walks on low-dimensional lattices. Discrete Math. Appl, 19(2):191–214.
16. Yarovaya, E. B. (2010). Models of branching walks and their application in reliability theory. Autom. Re-mote Control, 71(7): 1308–1324.
17. Yarovaya, E.B. (2012). Spectral properties of evolutionary operators in branching random walk models. Math Notes 92: 115. https://doi.org/10.1134/S0001434612070139
18. Yarovaya, E. (2013a). Branching random walks with heavy tails. Commun. Statist. Theory Methods. 42(16): 3001–3010.
19. Yarovaya, E. B. (2013b). Branching random walks with several sources. Math. Popul. Stud., 20(1): 14–26.
20. Yarovaya, E. B. (2015). The structure of the positive discrete spectrum of the evolution operator arising in branching random walks. Dokl. Math. 92(1): 507–510. DOI: https://doi.org/10.1134/S1064562415040316.
21. Zhizhina, E., Komech, S., & Descombes, X. (2015). Modelling axon growing using CTRW. arXiv.org. URL: https://arxiv.org/pdf/1512.02603.

Civil registration and identity for all, a pathway to the Sustainable Development Goals (SDG's): Malaysia's perspective

Nazaria Baharudin, Mohamad Shukor Mat Lazim, Suzira Daud
Department of Statistics Malaysia

Abstract

Malaysia has adopted the resolutions proposed in the **Ministerial Conference in a Ministerial Declaration to 'Get Everyone in The Picture'** and **Regional Action Framework (RAF) for the Decade** in 2014. In 2015, Malaysia stated its commitment together with 193 countries to implement the 2030 Agenda during the United Nations General Assembly (UNGA). SDGs are a universal call for action to end poverty, promote inclusiveness and well-being of the people, protect the planet and ensure that all people enjoy peace and prosperity with the objective of leaving no one behind. A well-functioning Civil Registration and Vital Statistics system (CRVS) plays an importance role, directly and indirectly in monitoring and achieving SDG Goals, Targets and Indicators. In the context of Malaysia, the fundamental principles behind CRVS are in line with the SDGs, including the objective to support good governance and to promote inclusion in the country. This aspiration is integrated in the 11th Malaysia Plan and the 2030 Agenda of SDGs has been aligned with the strategies and initiatives of the mid-term review of the 11th Malaysia Plan.

Department of Statistics Malaysia (DOSM) is the focal point for the compilation and coordination of CRVS and SDGs Indicators for Malaysia. Thus, the aims of this study is to highlight the role of DOSM in realizing the vision of the Decade in terms of improvement in CRVS systems, issues and challenges in coordination and compilation of vital statistics towards monitoring the SDGs.

Recent monitoring and assessment exercises on CRVS and SDGs indicators for Malaysia revealed that 70 indicators relate to each other. There are about 56 per cent of the indicators are available, 23 per cent of indicators are partially available, 15 per cent are partially available & need further development, 9 per cent are not available and two per cent are not relevant. Active engagements with related agencies will be undertaken continuously at national, regional as well as state level in order to provide evidence-based statistics for SDGs monitoring.

Thus, the continuous improvements in vital statistics and strengthening the development of SDGs indicators require major investments in financing, capacity building, capacity gaps, data limitation and coordination to ensure better decision making and better target in SDGs measurement and achievement.

Keywords
CRVS, SDGs, DOSM, Decade.

1. Introduction and Background

1.1 Introduction

Malaysia plays a vital role in realising the vision of the Decade: that by 2024, all people in Asia and the Pacific will benefit from universal and responsive CRVS systems which has been declared during Ministerial Conference on CRVS in Asia and the Pacific from 24 to 28 November 2014 in Bangkok, Thailand.

CRVS is defined as the "continuous, permanent, compulsory and universal recording and production of vital statistics on the occurrence and characteristics of vital events in accordance with national laws, rules, regulations and policies including births, deaths, foetal deaths, marriages, divorces, adoptions, legitimating and recognitions" (UNSD 2014).

1.2 Background

Civil registration was first made mandatory by law in three (3) Straits Settlement states of Singapore, Penang and Malacca in the late 1860s. Cognizant of the importance of CRVS to the SDGs and well-being of the Nation and its people, the Government of Malaysia has embarked on continuous improvement and restructuring of CRVS.

The administration of CRVS remains a requisite and central priority for good governance, thus the relevant structure and systems are consistently reviewed and enhanced through the years when manual records were maintained and to the present use of ICT to record all significant events of individuals throughout the country.

Since then, the administration of CRVS in Malaysia has evolved into an extensive framework and network based on structured and extensive legal and functional coordination between all government agencies relevant to CRVS.

Since the enactment of the Ordinance on Registration of Births and Deaths in 1869 in the Straits Settlement, various legal structures have been put in place to enable structured and systematic civil registrations to this date. Currently, Malaysia has more than 20 legal instruments to legislate and facilitate CRVS in Malaysia, and Malaysia is in the midst of amending some of these laws to address changes and challenges as a result of globalization and mobilization.

The National Registration Department (NRD) is responsible for the registration of vital event that is birth, death, marriage and divorce of Non-Muslim. Muslim marriages are under the purview of the respective state religious department as well as two federal agencies namely Department of Islamic Development Malaysia (JAKIM) and Department of Syariah Judiciary

Malaysia (JKSM). Meanwhile, Department of Statistics Malaysia (DOSM) is responsible to process and disseminate vital statistics.

1.3 Governance of CRVS in Malaysia

Main National CRVS Stakeholders in Malaysia are Ministry of Home Affairs (MOHA), Ministry of Health (MOH), National Registration Department (NRD) and Department of Statistics Malaysia (DOSM), 14 states religious department, JAKIM and JKSM.

The Ministry of Home Affairs (MOHA) has commissioned two committees for the implementation of CRVS in Malaysia, namely the Steering Committee and the Technical Committee. The Steering Committee is responsible to approve the plans, monitors the progress of the implementation of measures, to assess performance targets (goals and targets) and confirmed reports of CRVS activities in Malaysia. The Steering Committee is chaired by the Secretary General of MOHA.

Meanwhile, the Technical Committee is responsible for submitting the plans, suggesting performance targets (goals and targets) and to submit reports on the implementation CRVS activities in Malaysia. The Technical Committee will be chaired by the Director General of NRD. As for vital statistics, DOSM is responsible to compile, analyse and disseminate the statistics to public. There are three working groups in DOSM which established to ensure the quality of vital statistics is in line with international standard. The working groups are namely Technical Working Group (TWG), Inter-agency Planning Group (IAPG) and DOSM's Publication Committee.

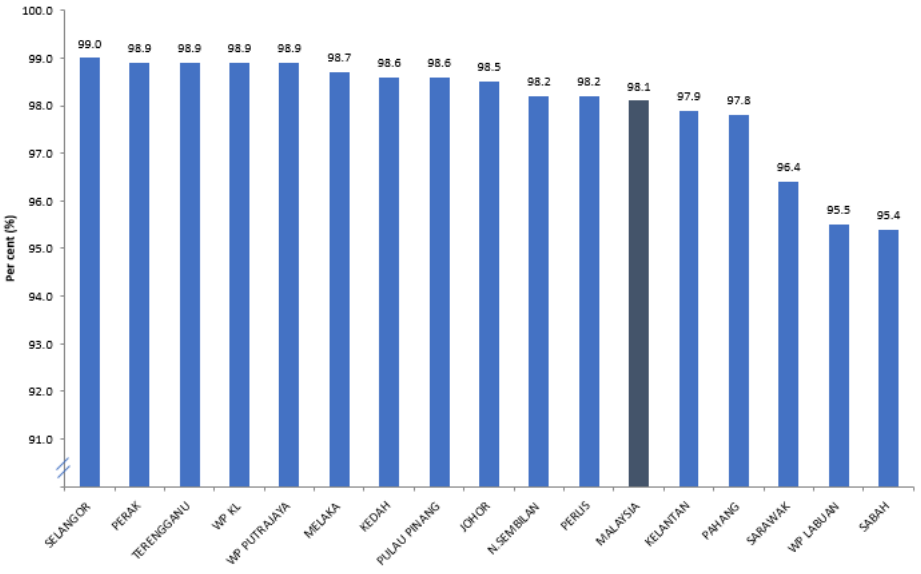
2. Recent Monitoring and Assessment Exercises

2.1 Assessment on coverage

The assessment study for coverage of birth registration has shown more than 90 per cent complete for Peninsular Malaysia from the 1960s. This comparison has been made between the number of births registered with the number of children as well as estimates of fertility from censuses have indicates that the birth registration is virtually complete. Assessment of death registration coverage was made during the preparation of life tables for Peninsular Malaysia showed no major problems in under registration of deaths except perhaps a very small amount of under registration at old ages.

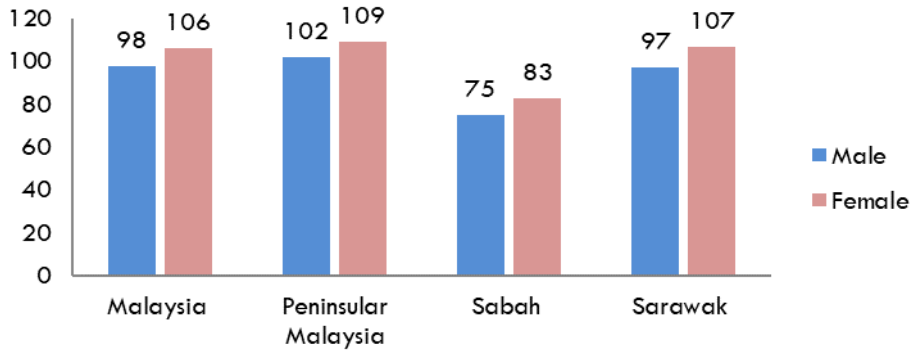
In 2016, the Department of Statistics Malaysia once again conducts an assessment review for birth registration and death coverage using data 2014. The study found that birth registration coverage rates for all states in Malaysia exceeded 95 per cent. The highest rate was recorded by Selangor of 99.0 per cent while the lowest rate was recorded by Sabah (95.4%).

Chart 1: Percentage of live births registration completeness by state, Malaysia,2014



The study also found that death registration coverage rate for Malaysia was almost 100.0 per cent except for Sabah (88.0 per cent).

Chart 2: Percentage of deaths registration completeness by region, Malaysia, 2014



2.2 Assessment on Organizational Structure

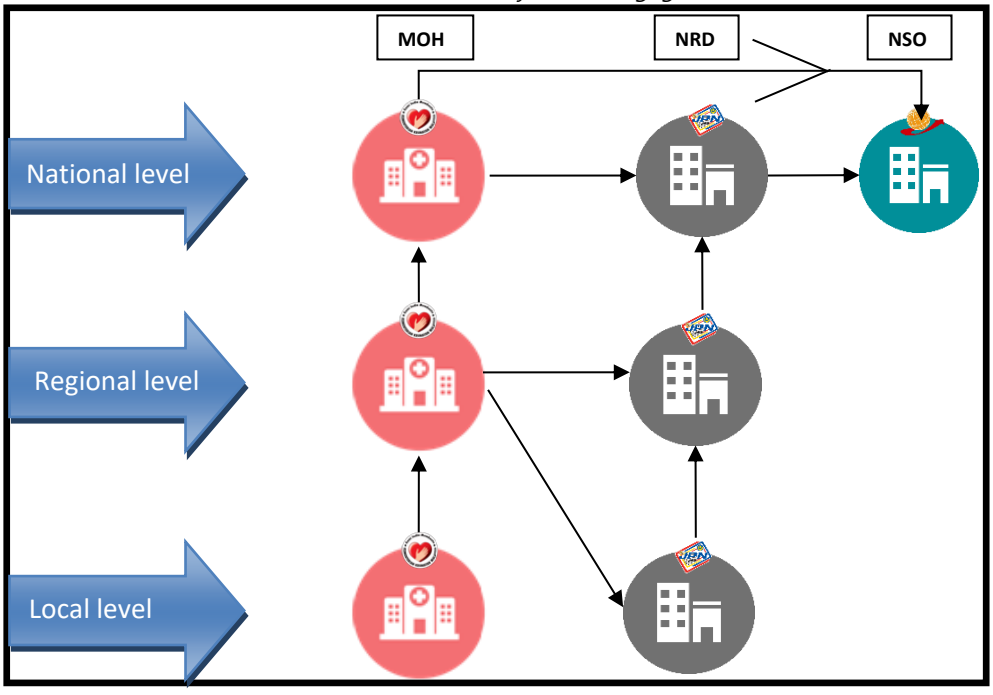
The health sector, particularly health institutions, plays a significant role in vital statistics system. First, health institutions act as informants of the occurrence of births, foetal deaths and deaths. Secondly, certifying the causes of death and this role can only be performed by physicians attached to health institutions.

In the Ministry of Health (MOH), information on health management including birth and death registration using the Medical Care Information System (SMRP). SMRP is based in the Ministry of Health and a component of Health Management Information System (HIMS). The aim of HIMS is to coordinate health information nationwide particularly in facilitate

reporting and producing statistics for inpatient services, outpatient services and support to the various levels of management in the Ministry of Health. Using this system, the whole country will adopt the information system at one time and thus, allowing performance comparison to be made between states. In addition to SMRP, there was other sub-systems such as dental, training, facilities, healthcare and others. The integration of these various information systems in HIMS enables performance to be evaluated as a whole in terms of delivering comprehensive healthcare. The information released to health personnel at various levels of administration associated with the delivery of medical care services. Administrative levels comprise national, state, district and operational levels such as hospitals and health centres.

Births and deaths occurring in health facilities are certified by Medical Officers. Medical Officer comprise those in Government Hospitals or private health facilities whom are registered with the Malaysian Medical Council as medical practitioners. Medical Officers who are on duty will fill information about births and deaths in SMRP and provide some documents and forms to parents, informants or next-of-kin of the deceased to be filled. Parents, informants or next-of-kin of the deceased will fill out birth or death registration documents and bring it to the NRD office for registration. Birth and death registrations can be done at any NRD counters as the system has allowed it to operate online.

Exhibit 1: The multi sectoral health facility-based engagements on different levels

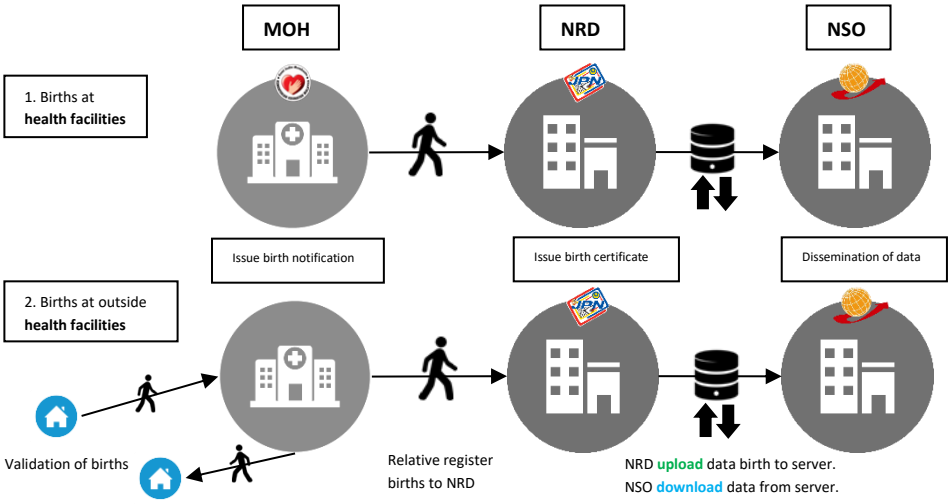


2.3 Registration process and data flows

The process of birth registrations are as follows:

i. Baby delivered in hospital / any location in Malaysia;
ii. Informant go to the nearest National Registration Department;
iii. Registrar will register the birth through online system (i-JPN); and
iv. Registrar will issue a birth certificate.

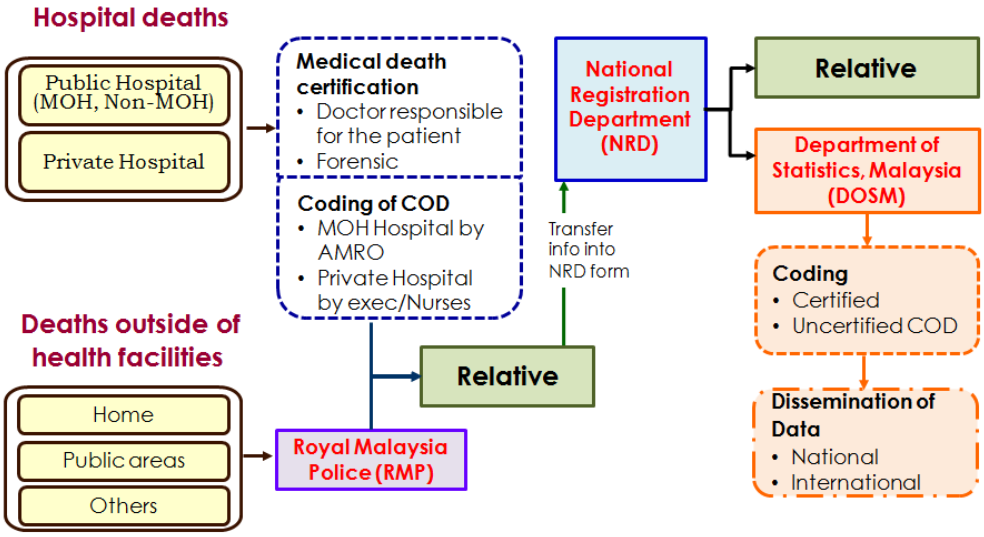
Exhibit 2: Births registration process and data dissemination flows



The process of death registrations are as follows:

i. Endorsement of Death by Medical Practitioner or Police – burial permit (Hospital/Home);
ii. Submission of application by the informer to the nearest National Registration Department;
iii. Registrar will register the death through online system (i-JPN); and
iv. Registrar will issue a death certificate.

Exhibit 3: Deaths registration process and data dissemination flows



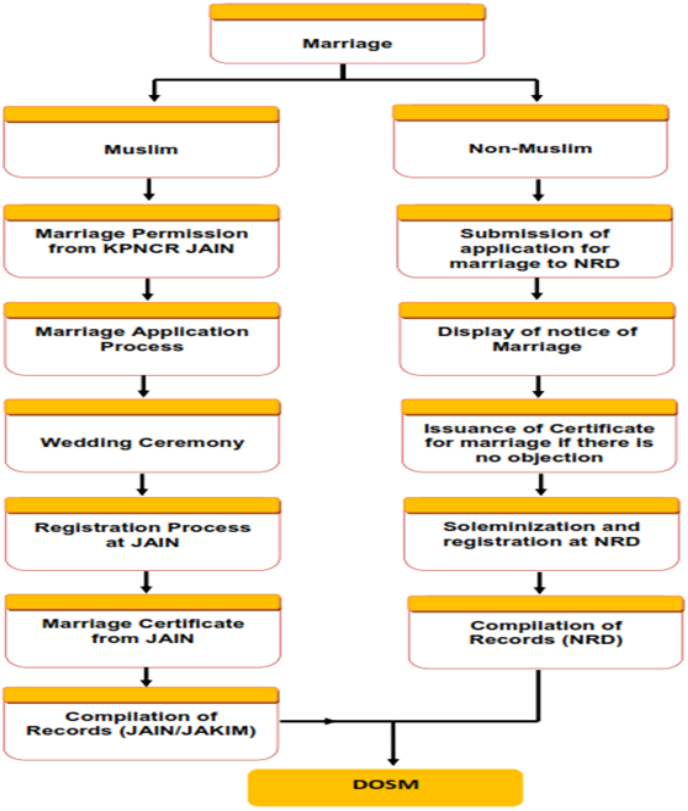
The registrations will be processed through online and stored in NRD's servers in the headquarters. All the data can be accessed by all NRD offices using i-JPN system.

The process of marriage registrations are as follows:

i. For Muslim:
   a) Confirmation of marriage application with local officials in the area where the bride/ groom resides;
   b) Marriage Application permission confirmed by the Assistant Registrar;
   c) The Registrar may only issue a consent letter of marriage after the applicant has settled the prescribed fee;
   d) Solemnization and registration at State Islamic Religion Department (JAIN); and
   e) The Registrar shall ensure that the marriage certificate is kept where the bride lives or resides.

ii. For Non-Muslim:
   a) Submission of marriage application to NRD;
   b) The applicant must reside in the district where the marriage is to take place;
   c) Submission of application for marriage registration to NRD;
   d) Issuance of Certificate for Marriage if there is no objection; and
   e) Solemnization and registration at NRD.

Exhibit 4: Marriage registrations process and data dissemination flows



The process of divorce registrations are as follows:

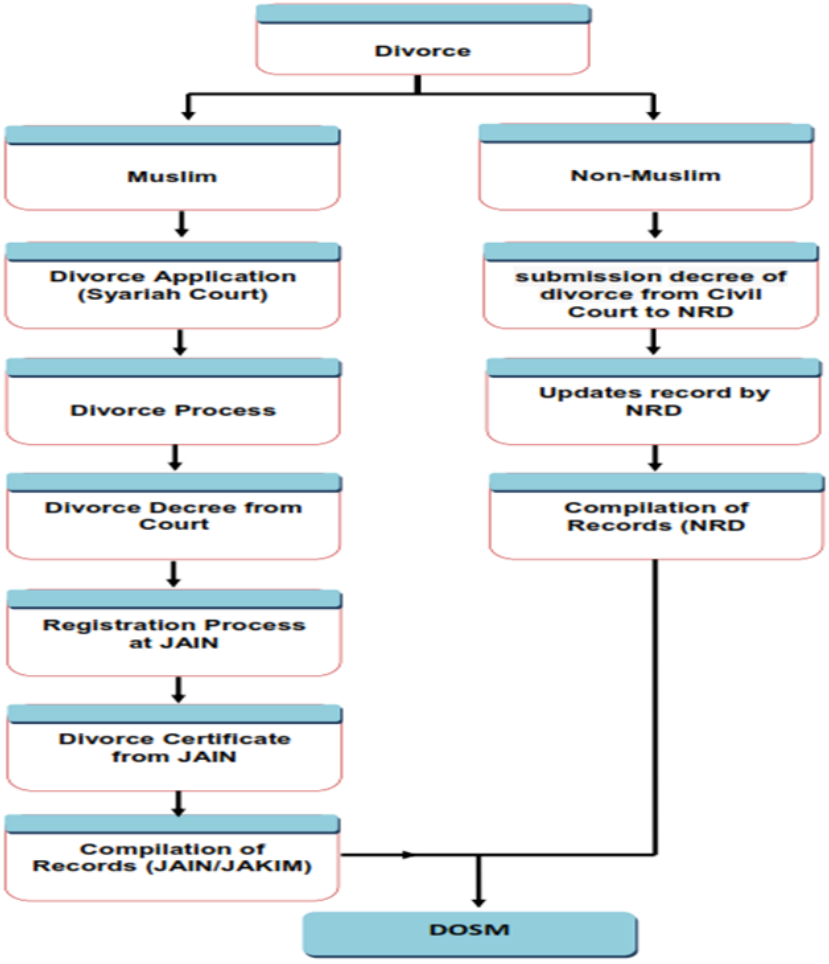
For Muslim:

i. The Registrar shall receive a copy of the results of verification of divorce issued by the Syariah Court;
ii. The Registrar shall review a copy of the results of verification of the divorce. If there is an error, the applicant needs to correct the information at the court;
iii. Once satisfied, the Registrar General shall sign a Certificate Divorce; and
iv. Registrants can only issue a Certificate of Divorce after the applicant make payments for divorce registration.

For Non-Muslim:

i. The applicant needs to submit the decree of divorce granted by high court to NRD; and
ii. NRD will update the personal marriage record regarding decree of divorce.

Exhibit 5: Divorce registrations process and data dissemination flows



2.4 Transfer of records and data flows in the civil registration system

NRD is responsible for registration of births and deaths. Registrations of birth and death at all 232 NRD offices all over the country will be keyed in on i-JPN system and the data will be stored in the server.

NRD has two (2) types of data backup which are Online Mirroring and Backup Tape. Online Mirroring is the replication of logical disk volumes onto separate physical hard disks in real time to ensure continuous availability. The replicated data is stored outside NRD's building. A mirrored volume is a complete logical representation of separate volume copies. The other type of data backup is the backup tape which is sent to DRC every day. JPN creates two (2) copies of backup tape, one (1) copy is sent to the DRC and one (1) copy is saved at Putrajaya Data Centre.

Birth, death and marriage records are obtained from its original sources. The birth information is derived from Ministry of Health (MOH) and the checking of parent's information is made through JPN's records to ensure the accuracy of data.

The original record of death is derived from MOH and Royal Malaysia Police (PDRM). The death information received from MOH is keyed in into the system by the Counter Operator and is checked by the Supervisor. Dual Layer checking is made to ensure the validity and accuracy of death data is registered (including birth and death cases).

The resource of PDRM's death record is derived through the strategic collaboration of PDRM-JPN. The integration allows quality, consistency and updated information of death accessible direct from the police station. Completeness and availability of data enables the registration of death to be verified and registered at timely manner and accurately.

On the other hand, records of marriage, divorce and ruju' (NCR) for Muslim are from JAKIM and JAIN. Meanwhile, records NCR for Non-Muslim was received from National Registration Department (NRD).

Births, deaths and NCR for Non-Muslim registered data from NRD are provided to DOSM through the Secure File Tranfers Protocol (SFTP). NRD will upload the data to server and DOSM will download from the server. Next, the data will be stored in the DOSM data warehouse that is the Statistics Data Warehouse (StatsDW). The data will also be stored at the Population and Demographic Statistics Division for backup. The data are transferred from NRD to DOSM on a monthly basis. DOSM will receive the data usually in the last week of the month.

2.5 Production and dissemination

The publication of vital statistics which include birth, death, marriage and divorce will be processed and published by DOSM. Birth and death statistics are published quarterly via Quarterly Demographic Statistics and annually via Vital Statistics Publication. Vital Statistics Publication was first published in 1963 and Quarterly Demographic Statistics was first published in 2017.

Quarterly Demographic Statistics provides statistics on live births, deaths and population on current year whereas annually Vital Statistics Publication provide more detail on statistics live births, deaths and stillbirths. The demographic indicators such as crude birth rate, crude death rate, crude rate of natural increase, total fertility rate and age-specific death rate are included in the Vital Statistics Publication. The statistics have been compiled based on concepts and guidelines from **Principles and Recommendations for a Vital Statistics System, Revision 3, United Nations Statistics Division (2014).**

Marriage and divorce statistics was first published in December 2015 for internal circulation and in 2018, the complete publication was published in December, 2018.

2.6 Incentives and disincentives for registration

In order to facilitate easy access to each citizen to register the vital events, the government, through NRD has been providing counter services at its 232

branches across the country, including 18 UTC (Urban Transformation Centre) and 11 RTC (Rural Transformation Centre). NRD offices in UTC open daily from 8.30 am to 10 pm. The UTCs are especially very popular with the urban poor as they can come to these centers after normal working hours to register or use other public services.

But, similar to other developing countries, Malaysia faces various challenges in capturing vital information of all individuals in the country largely due to geographical, resource and communication barriers. There are remote areas that pose accessibility difficulties to public officers to reach out to inhabitants of these areas or villages either to register or to educate and inform them.

To overcome this issue, NRD has intensified and strengthen the outreach program to those who have no access to the NRD office. In 2015, a total of 1,554 outreach programs were held across the country and as many as 32,151 applications were completed within the year. In 2016, as many as 1,772 outreach programs were held across the country and as many as 60,023 applications have been resolved. Target groups who benefit from this outreach program comprised of old citizens, disabled person, orphans, and those living in rural areas who do not have access to the NRD office.

3. Strategies for Improving CRVS Systems in Malaysia

Malaysia undertakes several strategies to improve the CRVS system in Malaysia. The strategies are as follows:

i. Improvements on data quality and dissemination. The activities include:
   a. Data harmonization with Ministry of Health: Maternal death, under five death, stillbirth;
   b. Data adjustment due to under reporting of deaths;
   c. Reduce time lapse to publish annual publication of Vital Statistics and Causes of Deaths from two years to one year;
   d. Production of Quarterly Vital Statistics starting 2017;
   e. Statistics by small area (by District level) starting 2018; and
   f. Producing new indicators (i.e. premature death, death rate for selected COD).

ii. Improvement on legislation
   a. Revision on Birth and Death Registration Act (Act 299) 1957 in 2016. The revision enlists the following:
      - "Special provision as to registration of birth and deaths";
      - Births and deaths can be registered at any NRD in Peninsular Malaysia (online system);
      - Extending the period of birth registration up to 60 days and abolished late birth registration; and

- The extension of the death registration/confirmation of death and post mortem from 12 hours to 7 days.

b. Combining burial permit with death registration as well as to introduce on line system for these two procedures; and
c. Undertakes National Statistical Review to strengthen DOSM to obtain administrative data from other agencies.

iii. Improvements on Cause of Deaths Statistics
   a. Improve Medically Certified Deaths by using verbal autopsy approach; and
   b. Increase certification of coders that will contribute to improve data quality on Causes of Death.

iv. Improvement in service delivery especially on birth and death registration
   a. Empowerment of NRD Outreach Program
      - Facilitate and deliver services to the people's doors, especially the rural people in remote areas that are less able in terms of transportation, physical and also income
      - Determine each citizen has access to basic services (i.e. obtain identification document)
      - Services and facilities in Rural Transformation Centre (RTC) and also Urban Transformation Centre (UTC).

v. Improve engagement and collaboration between agencies. Various engagements are undertaking particularly in address specific issues for example collaborate with Ministry of Health in providing training to DOSM's Coders for ICD 10.

4. Issues and Challenges

There are two issues with regards to births and deaths encompassing under reporting and data integration. In terms of under reporting, births and deaths statistics in health facilities only covered the registered cases to health facilities excluding those which are not reported directly to health facilities. There is a need to integrate data among agencies by maintaining the system and updating with the latest technologies require development costs. The systems of the respective agencies need to be aligned in terms of data storage, storage and analytics technology capabilities before making data integration.

For marriages and divorces statistics, there are four issues that have been identified as challenges in compiling the statistics in Malaysia. The issues are as follows:

i. The system format developed by each JAIN is different. Therefore, it will cause the compilation of NCR data to be difficult;
ii. JAKIM has developed an online system for the registration of NCR which is Management System Marriage Islam Malaysia (SPPIM) as the centralized record-keeping initiatives, but it covers only six (6) states

of Perlis, Kedah, Perak, Negeri Sembilan, Melaka and JAWI. However, approval is required from the JAIN to obtain the data. JAKIM will only supply NCR data to DOSM upon approval from JAIN;

iii. Individual records are needed by DOSM for calculating rates of marriage and divorce according to the events that will produce accurate statistics; and

iv. JAIN has supplied aggregate data by registration for NCR to DOSM but the analysis for marriage and divorce are limited.

5. Conclusion

Birth, death, marriage and divorce are an important component of the structure of the population. A complete CRVS systems will provide reliable population estimates, which are often needed in the denominator for measuring progress for examples, indicators that measure "per capita" or "per 1,000 population" or "live births" as the denominator. A number of initiatives in the CRVS improvement has enhanced the standardising and streamlining of civil registration and vital statistics processes, integration data from multiple systems, extension of registration coverage and producing accurate, complete and timely vital statistics.

CRVS systems also play an important role in the progress of achieving SDG targets and indicators. For example, birth registration is the first step in establishing legal identity for individuals and serves as the foundation for social inclusion. CRVS can also help to prevent child marriage, since marriage registration indicates the age of groom and bride. A birth certificate from birth registration can be used as proof of age and a fundamental to an individual's right to an identity and entitlements that go with it, such as to be in education, assessing health services, banking services as well as to participate in social community life. These will improve and enrich people's lives.

Therefore, the fundamental principles behind CRVS are in line with the SDGs, including the objective to support good governance and to promote inclusion among Malaysia's population. All the Investments at improving CRVS systems constitute a meaningful step towards achieving Malaysia SDGs targets.

References

1. Vibeke Oestreich, and others (2014). Status Analysis on Civil Registration and Vital Statistics (CRVS). Statistics Norway, Documents 2014/41.
2. National Registration Department and Statistics Department, Malaysia (1993). Civil Registration and Vital Statistics System in Malaysia.
3. United Nations Economics and Social Commission for Asia and the Pacific (2014). Asia Pacific Population Journal, Vol. 29, No.1/2014.

4. Principles and Recommendations for a Vital Statistics System, Revision 3, United Nations Statistics Division (2014).
5. Handbook on Civil Registration and Vitals Statistics System: Management, Operation and Maintenance (Revision 1), United Nations (2018).
6. Civil Registration and Vital Statistics (CRVS) for Monitoring the Sustainable Development Goals (SDGS), The World Bank (2017).
7. Transforming Our World: The 2030 Agenda for Sustainable Development. Resolution A/RES/70/1. New York: UN General Assembly (2015).
8. Convention on Consent to Marriage, Minimum Age for Marriage and Registration of Marriages. New York: UN General Assembly (1964).

An approach to monitoring multivariate time between events

Ross Sparks, Aditya Joshi, Cecile Paris, Sarvnaz Karimi
Data61, CSIRO Sydney, Australia

Abstract

This paper focuses on monitor plans aimed at the early detection of the increase in the frequency of events. The literature recommends either monitoring the Time Between Events (TBE) if events are rare or counting the number of events per unit non-overlapping time intervals if events are not rare. However, recently, work has suggested that monitoring counts in preference to TBE is not recommended even when counts are moderately high. Monitoring TBE is the real-time option for outbreak detection, because outbreak information is accumulated whenever an event occurs. This is preferred to waiting for the end of a period to count events. If the TBE reduces significantly, then the incidence of these events increases significantly. This paper explores multivariate TBE options (e.g., the time between flu events at all the hospitals in a state of Australia). This will be compared with the approach to monitoring counts. We consider the case when TBEs are Weibull distributed in situations where daily counts are moderately low. The paper will discuss and compare the approaches based on TBEs and counts.

Keywords

monitoring; multivariate; counts; time between events; statistical process control

1. Introduction

Dealing with multivariate time between events is challenging because events do not necessarily occur simultaneously in time for the population of interest. The application considered in the paper is symptoms that people in social media expressed that they are suffering from in terms of poor health outcomes. These symptoms are taken as diarrhoea, vomiting, headache and generally feeling unwell. Such symptoms are generally easy to self-diagnose. The only time these occur simultaneously is when the same person expresses that they are suffering from all of these, and in the dataset we considered, this never occurred. The fact that people seldom expressed that they were suffering from a combination of symptoms increases the challenge in dealing with these in a multivariate way. This was the case in our application.

The option of imputing a censored estimated value for those events that did not occur during the timing of one of these events was considered, but this option would make finding an outbreak difficult when the imputed values are based on the in-control distribution. One option was to impute these based on the out-of-control model, but we don't know the nature of this

model in real-time when engaging in prospective surveillance plans. We explored the option of aggregating the time between events within a day in this paper. This does reduce the problem to daily monitoring but still has the advantage of treating it as a multivariate problem and hence can have greater power when these are reasonably highly correlated. The next section will explore this option in more detail.

2. Multivariate TBE

The multivariate TBE considers the average daily TBE values for all events that occurred on a particular day. This means that the dimension of the multivariate monitoring plan changes with each day. If only one of the symptoms diarhoea, vomiting, headache and generally feeling unwell occurs in a day then we apply the univariate TBE monitoring as outlined in Sparks et.al. (2019a, b) for that one symptom. Otherwise the monitoring plan is multivariate with the potentially changing the dimensions of the multivariate monitoring plan each day. Treating the monitoring plan as multivariate on some days hopefully offers it greater power for such monitoring plans.

We will assume that the in-control distribution is Weibull, and this can be fitted using gamlss library in R (Stasinopoulos et al., 2006, Stasinopoulos and Rigby, 2007) for each symptom separately. This model is used to define the in-control non-homogeneous expected (parameter) values for TBE. These are predicted at the time of each event using the respective Weibull regression models. Since we don't have enough data to deliver a Phase I and Phase II, for developing our plan, we only use a retrospective surveillance approach. The model includes the following adjustments (explanatory variables): seasonal harmonics, day of the week, and within day harmonics.

Since there can be more than one event in the day, the statistic used the average time between events that occurred on that day. Since this average has a smaller standard error to those with only one event in the day, we multiply the average by the square root of the number of events within a day, so each average has the same standard error.

3. Assessment process

We assess the performance of the plans using the number of days to an alarm and train all plans to have the same false discovery rate, so they can be fairly compared. The daily counts where fitted using a negative binomial regression model as in Sparks et. al. (2010, 2011) and the EWMA is applied to the Pearson residuals of this model. If an "event" is not signalled within 7 days as unusual then it will be regarded as having been missed. The assessment process looks at univariate counting process monitoring, univariate time between events and the multivariate approach outlined above. The out-of-control false discovery rate is taken as one in 100 days for the multivariate process, whereas the univariate charts examined an out-of-control false

discovery rate is taken as one in 400 days. Note that the correlation between these symptoms considered are very low, less than 0.079.

The result are as follows. All plans monitored the Pearson residuals of the models (whether Weibull or negative binomial distributed) and these Pearson residuals for a particular symptom are smoothed over time using the EWMA statistic. The univariate TBE monitoring plans did not flag an unusual trend in the TBE in either direction. The univariate daily counts indicate lower than expected counts from 14 January 2015 to 11 February 2015. Any event that was not flagged by more than two consecutive days were ignored in the multivariate plans. The multivariate plan used the Weibull regression models for each symptom separately. Hotellings robust version of the T-squared statistic of Sparks (2015) was used to flag unusual events in terms of how large their Pearson smoothed residuals are. Note that if the TBE increases, then there is no outbreak. The TBE values need to significantly reduce to flag an outbreak. Multivariate plans flagged several events that are listed in the table below:

Dates	Reason
2015-02-24 to 2015-05-01	Larger waiting times between events for diarrhoea and headaches
2015-07-19 to 2015-07-29	Larger waiting times between events for all symptoms
2015-11-24 to 2015-12-24	Larger waiting times between events for vomiting, diarrhoea and headaches
2017-03-22 to 2017-03-31	Larger waiting times between events for vomiting and headaches
2017-08-17 to 2018-02-06	Larger waiting times between events for diarrhoea and unwell. At times there are larger waiting times between events for headaches

4. Multivariate charts are useful for diagnosing the nature of outbreaks

We use the dynamic biplot of Sparks et al (1997) to explore the nature of outbreaks using unsmoothed version of Pearson residuals from the respective Weibull regression models. Information on how to interpret this biplot can be found in Sparks et al. (2017). We are looking for periods of TBE values that are lower than expected, i.e., those in the biplot that are in the opposite quadrant to the most recent points in the observation plot. For this we only use the days with all symptoms occurring on the same day, i.e., only considering days where there is at least one of the four symptoms on the day. This was by far the most common situation – occurring in 86% of the days in the dataset.

The first 40 days are used as training data for setting up the multivariate plans. The days without an event for a symptom were excluded from the dataset being considered in this section because it resulted in missing data for the day. Thereafter we explore the average TBE event daily, when they occur

simultaneously, by investigated its nature of the trends in the Pearson residuals.

Figure 1 indicated on 14 April 2015 a non-significant increase in vomiting but flagged that there was a sustained significant decrease in people having diarrhoea since the beginning of the monitoring period (which started on 27 March 2015).

Figure 2 illustrates a period where both vomiting and unwell incidence are significantly lower than expected compared with the first 40 days of data, however diarheoa has non-significantly lower TBE values. The biplot explains 83% of the variation and therefore displays most of the variation in the four dimensional dataset. This behavior persists from 9 December 2015 to after 7 April 2016.

Figure 3 indicates a non-significant decrease in TBE values for headache and diarrhoea, but higher than expected TBE values for vomiting and unwell. The largest differences are recorded for unwell because it is in the direction of the major axis, which explains 75% of the variation, while the second axis only explains 25% of the variation. Diarrhoea and unwell have become negatively correlated with a significantly lower correlation.

Figure 4 indicates a non-significant incidence increase for diarrhoea, but non-significantly higher than expected TBE values for vomiting and unwell. The largest difference is recorded as unwell because it is in the direction close to the major axis which explains 45% of the variation while the second axis only explains 23% of the variation. Diarrhoea and unwell have become negatively correlated with significantly lower correlation.

Figure 1: Dynamic Biplot on 2015-04-14

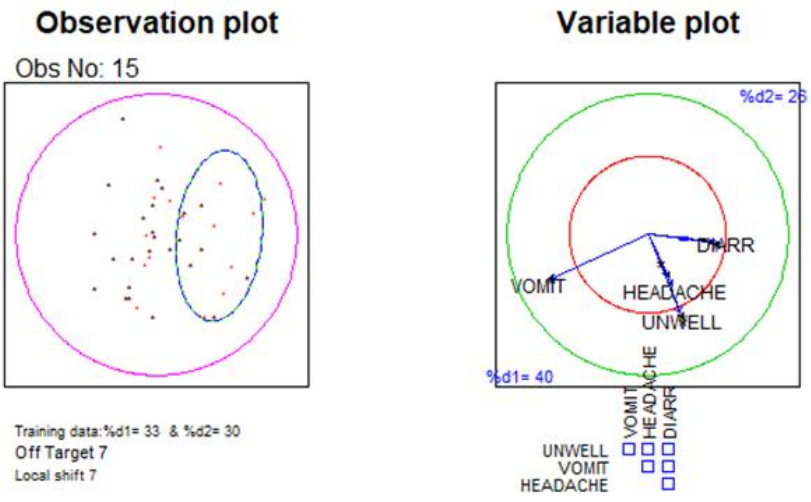


Figure 2: Dynamic Biplot on 2015-10-25

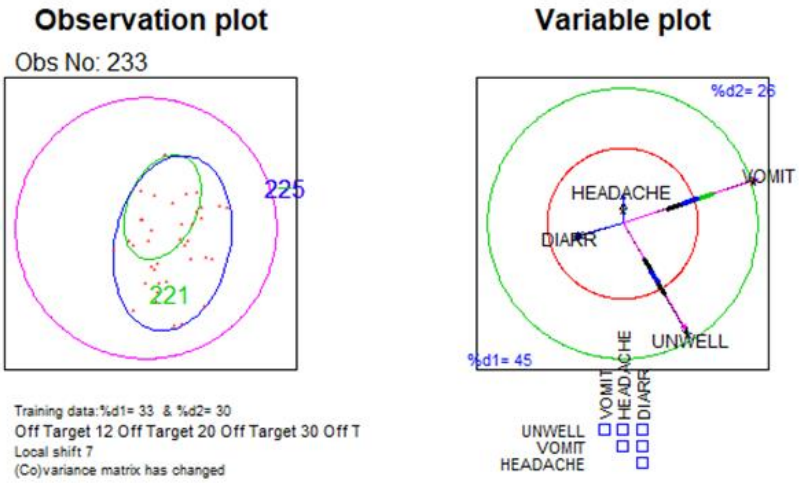


Figure 3: Dynamic Biplot on 2016-02-05

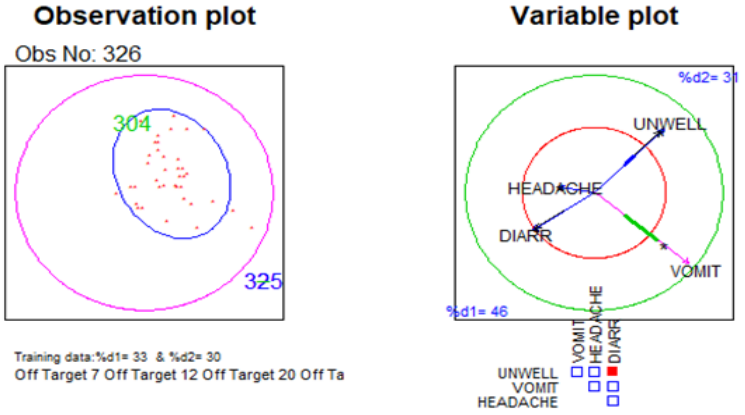


Figure 4: Dynamic Biplot on 2017-01-21

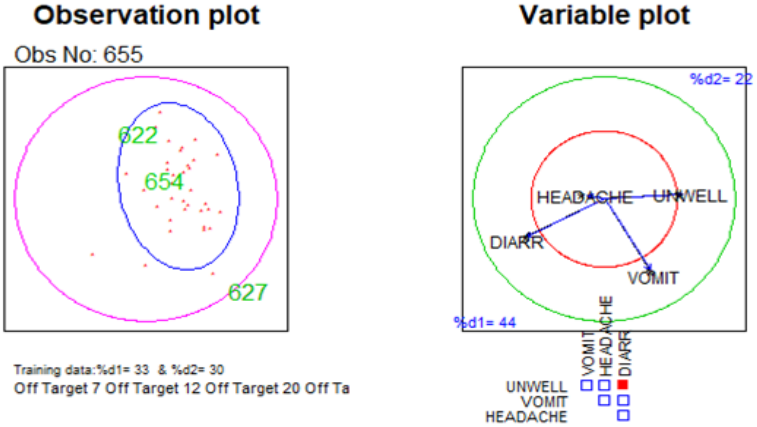
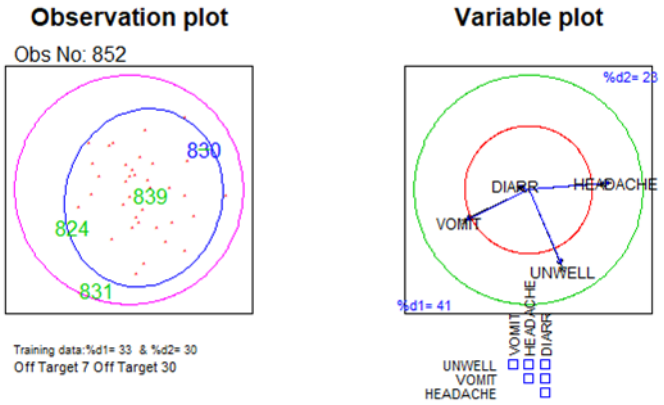


Figure 5 indicates a non-significant increase in incidence for vomiting, but higher than expected TBE values for headaches. The largest difference from baseline is recorded as headache and vomit because these are in the direction close to the major axis which explains 48% of the variation, while the second axis only explains 28% of the variation.

Figure 5: Dynamic Biplot on 2017-09



References

1. Sparks, R., Adolphson, A., & Phatak, A. (1997). Multivariate process monitoring using the dynamic biplot. International Statistical Review, 65(3), 325-349.
2. Sparks, R. (2015). Monitoring highly correlated multivariate processes using hotelling's T2 statistic: problems and possible solutions. *Quality and Reliability Engineering International*, *31*(6), 1089-1097.
3. Sparks, R., Keighley, T., & Muscatello, D. (2010). Exponentially weighted moving average plans for detecting unusual negative binomial counts. *IIE Transactions*, *42*(10), 721-733.
4. Sparks, R. S., Keighley, T., & Muscatello, D. (2011). Optimal exponentially weighted moving average (EWMA) plans for detecting seasonal epidemics when faced with non-homogeneous negative binomial counts. *Journal of Applied Statistics*, *38*(10), 2165-2181.
5. Sparks, R. S., Robinson, B., Power, R., Cameron, M., & Woolford, S. (2017). An investigation into social media syndromic monitoring. *Communications in Statistics-Simulation and Computation*, *46*(8), 5901-5923.
6. Sparks, R. S., Robinson, B., Power, R., Cameron, M., & Woolford, S. (2017). An investigation into social media syndromic monitoring. *Communications in Statistics-Simulation and Computation*, *46*(8), 5901-5923.

7. Sparks, R.S., Jin,B., Karimi, S., Paris,C. and MacIntyre, C.R. (2019). Real-time monitoring of events applied to syndromic surveillance. *Quality Engineering* with discussion. In Press.
8. Sparks, R.S., Jin,B., Karimi, S., Paris,C. and MacIntyre, C.R. (2019). Monitoring time between events using social media data. *Quality and Reliability Engineering International*. Submitted.
9. Stasinopoulos D. M., Rigby R.A. and Akantziliotou C. (2006) Instructions on how to use the GAMLSS package in R. Accompanying documentation in the current GAMLSS help files, (see also http://www.gamlss.org/).
10. Stasinopoulos D. M. Rigby R.A. (2007) Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, Vol. 23, Issue 7, Dec 2007, http://www.jstatsoft.org/v23/i07.

Healthcare fraud detection using machine learning approaches

Vassilis P. Plagianakos
Department of Computer Science and Biomedical Informatics University of Thessaly, Greece
Hellenic National Organization for the Provision of Health Services (E.O.P.Y.Y.)

Abstract

Biomedicine is undergoing a revolution driven by the explosion of biomedical data, as a result, Big Data has shifted the biomedical informatics research from case-based to data-driven-based studies. Data from hospitals and clinics form a very large data set (Big Data) since they have monthly submissions, posing several challenges under the perspective of Big Data mining and analysis. On parallel, fraud detection in the healthcare domain is an important issue, since it has considerably inflated losses for individuals, entities, and governments. Hence, there is an imperative need for new computational tools able to effectively detect fraud by exploiting the potential of Big Data. Machine Learning (ML) approaches can shed more light in healthcare fraud since they can cope with these challenges. In this study, we utilized clinical data from the National Organization for the Provision of Health Services of Greece, focusing on investigating the Clinics behavior with respect to their hospital expenditure. The core of our analysis is based on t-SNE and Density Peak, two well-established ML tools for data visualization and clustering respectively. Our results show that ML approaches can contribute to healthcare fraud detection and interpretation.

Keywords

Fraud Detection; Machine Learning; Visualization; Big Data

1. Introduction

We live in the "Big Data" era, where there is a great potential for revolutionizing the entire healthcare domain [1]. Biomedical data generation is increased constantly through the recent advancements in biomedicine field creating a large pool of heterogeneous increased with exponentially increasing rate. This data volume poses several challenges under the perspective of Big Data analysis and visualization. Given the fact that these data have ultra-high dimensionality and complexity, it is obvious that we need computational tools to cope with these challenges. Machine Learning (ML) techniques are among the best approaches to tackle these limitations [2,3]. Nevertheless, data generated in the health domain are too big, too complex and their production rate too fast for the healthcare providers to process and interpret with the existing tools. Hence, there is an imperative and urgent need

for more accurate, fast and intelligent methodological frameworks from the ML perspective. Furthermore, clinical data follows the Big Data era offering data with high diversity since these they come from different subareas [4]. Also, the majority of clinical data has high dimensionality due to a limited number of patients (small/large n, large p). However, most computational tools can handle data with large n and small p, since in high-dimensional data there exists the "curse of dimensionality" phenomenon [5].

Also, the integration of big biomedical data and advanced computational tools can contribute in healthcare fraud detection. The frauds in healthcare are classified in three main pillars related to health insurance, drug and medical. Healthcare fraud is a field with high impact since several people suffer financially with indicative examples the insurance holder who have to pay higher expenses while she/he receives reduced coverage, the business who pay increasing amounts for employer healthcare, increasing cost of doing business, clinics that charge patients for their services or charge services that should be covered by the state and so on. Indicatively, the World Health Organization (WHO) has estimated recently that every year the state is lost the 7.3% of the annual healthcare expenditure (around $470 billion) to healthcare fraud annually [6]. In this study, we utilized clinical data from National Organization for the Provision of Health Services of Greece, focusing in investigating the Clinics behavior with respect to their hospital expenditure. Our analysis is based on t-SNE and Density Peak, two well-established ML tools for data visualization and clustering respectively.

2. Machine Learning approaches in Healthcare Fraud detection

Machine learning (ML) approaches can tackle part of the complexity of fraud detection since the digitalization of health care information offers more data enabling robust Data Mining and ML frameworks [7]. These methods are classified into three categories as supervised, unsupervised learning and reinforcement learning. Briefly, the first category is the process where the algorithm constructs a function that represents given inputs (training set) at known desired outputs, with the ultimate goal of generalizing this function and for inputs with unknown output. It is used in real word problems related to classification, prediction and data interpretation. Unsupervised Learning is the process where the algorithm constructs a model for a set of inputs in the form of observations without knowing the desired outputs. We have no knowledge of the true label of data in order to compare its efficiency, as we can in previous model. It is used in real word problems related to data clustering and association analysis. The latter category concerns methods which learn a strategy of actions through direct interaction with the environment. It is used in planning problems such as robot mobility control or functions optimization in industries.

More specific, supervised learning approaches have a wide application in the domain of health care fraud detection. Indicatively, several studies have been proposed supervised learning based models for to healthcare fraud detection including classification schemes such as Neural Networks [8], decision trees [9,10] and Support Vector Machines [11, 12]. Concerning the unsupervised learning perspective, new types of fraud can be uncovered through the application of this category. usually the relative approaches are based on clustering methods. The first choice of selecting an unsupervised approach is data clustering, while fraud can be detected through samples that are nor members in a dense cluster or samples that are too far from the center of cluster. That means that this sample has less shared features among other samples and should be further evaluated. In recent literature there is a plethora of clustering approaches for fraud detection [13–17].

Also, an effective way is to search for samples outliers, which are potential fraud samples since they do not follow the behavior of the other samples [15]. Approaches using association rules is also an efficient manner for detecting healthcare fraud [18]. Furthermore, some studies have been integrated supervised and unsupervised methods proposing hybrid approaches for healthcare fraud detection. An indicative example is the study [19], where the authors examined an electronic fraud detection program that compared individual provider characteristics to their peers in identifying unusual provider behavior.

3. Analysis - Methodology

The data studied in this paper are from the National Organization for the Provision of Health Services (EOPYY), the main public purchaser of health services in Greece. EOPYY founded in early two thousand twelve, so it is still taking its first steps as i) a buyer of Health Care Services for Greek citizens and their families, ii) an assessor of Quality and Safety Services, by establishing rules in healthcare market, iii) an Health Technology Analyst of healthcare products and iv) a negotiator with healthcare stakeholders. More specific, purchasing enough and effective healthcare services for the insured citizens, the pensioners and the protected members of their families, of the insurance agencies that have been integrated with EOPYY, according to what is being foreseen in the regulatory framework of healthcare services as every time in effect. Also, the establishment of rules in designing procedures, in quality, in development, in assessing the efficiency and effectiveness of healthcare services market, the auditing of the funding process along with the rationalization in the use of public funding. The establishment of the criteria in the terms of the contracts with the providers along with the amendments of the contracts terms whenever needed. It is worth mentioning that the negotiation process with the providers regarding their remuneration, the

terms of their contracts with EOPYY, the prices of medical materials and drug reimbursement.

The clinical dataset concerns the period 01-01-2017 to 31-12-2017 and has 6379 samples that correspond to insured citizens, pensioners and the protected members of their families who have retrieved health services from the 82 regional Clinics (Branches) around Greece with respect to each corresponding Diagnosis Related Group (DRG). In this work, we are particularly interested in investigating the Clinics behavior with respect to their hospital expenditure. One would expect that Clinics should always remain in a control

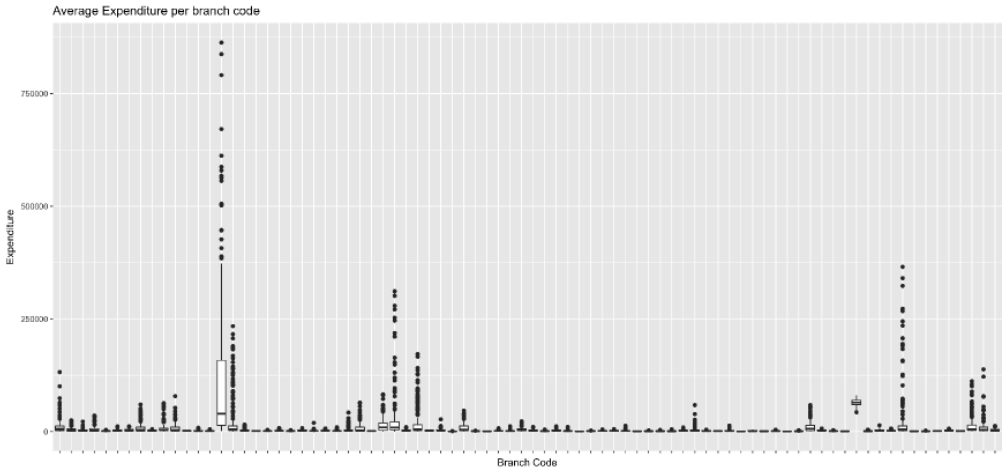


Figure 1: Average expenditure per Clinic.

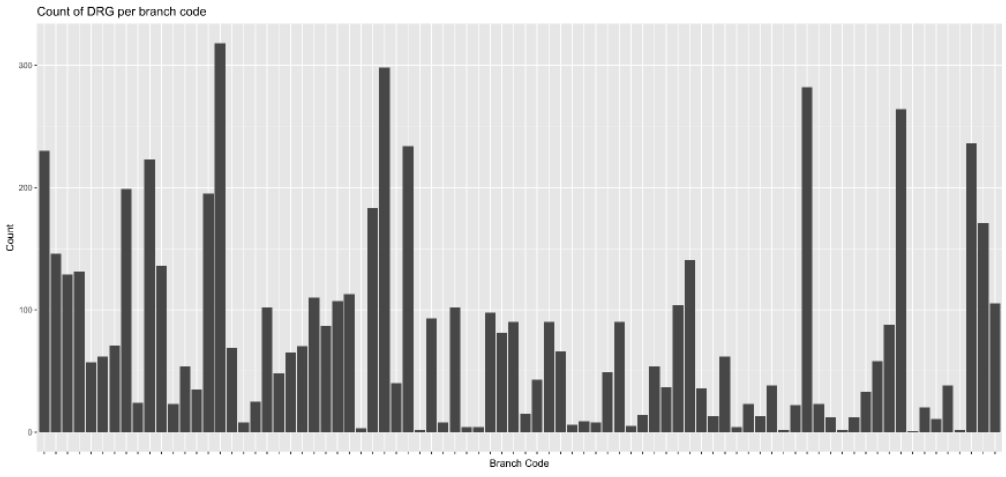


Figure 2: Count of DRGs per Clinic.

state and follow a particular pattern regarding their total expenditure taking into consideration the total number of patients (individual DRGs) they serve. A Clinic which would exceed the control limits defined by a set of parameters should be investigated further and in more detail, a process that usually

requires additional costs. To this end we may argue that most likely we will need to consider further parameters in characterizing Clinics. Instead of focusing on setting more appropriate control limits we intend to show that data driven decisions could be more appropriate for this purpose. In what follows, we present an explanatory analysis of the dataset at hand followed by the proposed scheme for analysis.

To begin with in Figure 1 we illustrate the average expenditure for each individual Clinic at the corresponding boxplot. It is evident that one can identify significant variations between the Clinics. In addition, it is interesting to examine the total count of incidents that each Clinic had to deal with during this period of time (see Figure 2). As expected, we also identify significant variation in this case. Finally, in Figure 3 we can see the total expenditure per Clinic.

To this end, we can gather information from the aforementioned statistics to create a dataset and the corresponding variables are the mean and standard

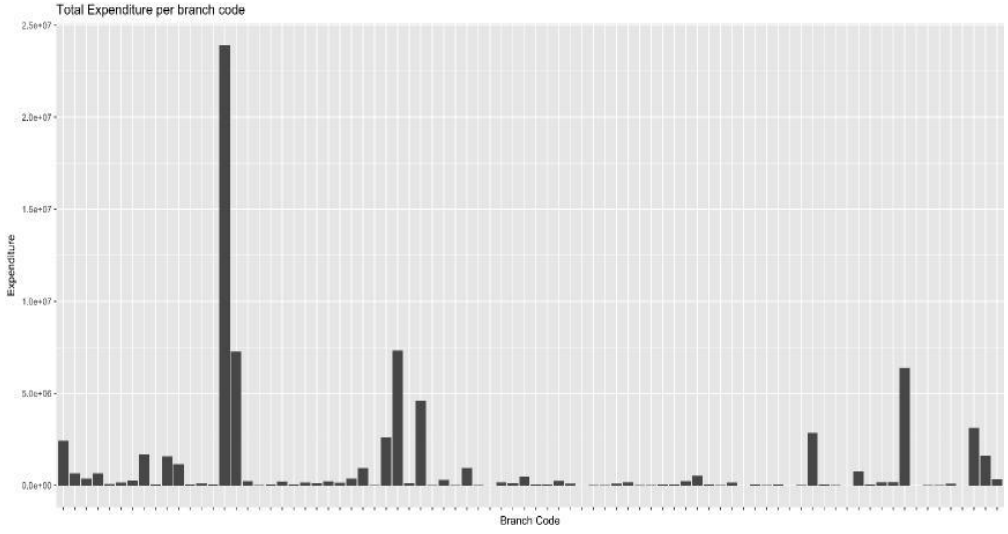


Figure 3: Total expenditure per Clinic.

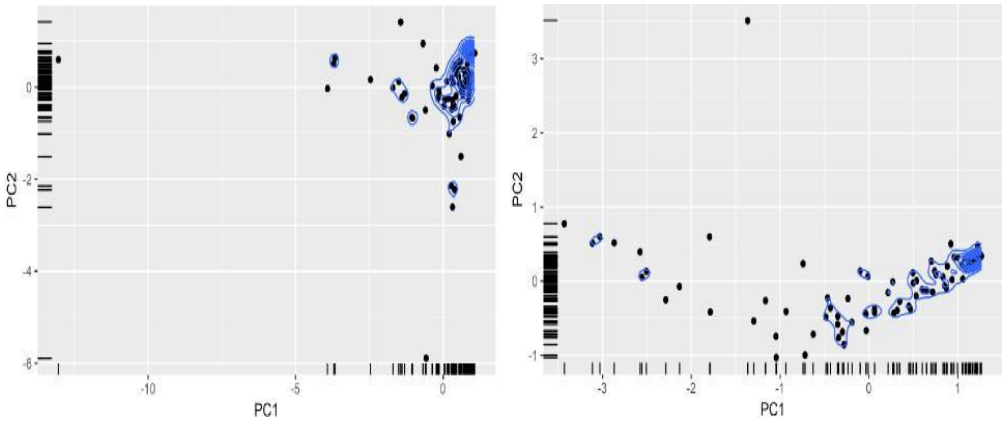


Figure 4: 2-dimensional visualization of the statistics based constructed dataset using PCA (left). 2-dimensional visualization of the complete dataset using PCA (right).

deviation of the expenditure, the count of incidents and the total expenditure (as a sum for this period). To visually investigate the structure of the resulting dataset we employ a 2 dimensional visualization using Principal Component Analysis (PCA).

As shown in Figure 4(left) although we can identify some outliers, a clear pattern is not available. Still critical information such as how are Clinics associated with different types of DRGs is missing. We already know from Figure 1 that the average expenditure varies between the Clinics, an effect that is probably caused by the variation in the types of DRGs that each Clinic deals with. To investigate this further we need to take into account whether a particular Clinic focuses on specific DRG types, for example a Clinic with high expenditure rate may deal with DRGs that are significantly more costly than others. As such we reconstruct the data incorporating information from types of DRG per Clinic. The newly generated variables are dummy variables containing the count of each DRG for each corresponding Clinic Code which is combined with the aforementioned statistics and normalized accordingly. In Figure 4(right) the 2-dimensional visualization using PCA is illustrated where we observe increased variability.

Subsequently, we may employ more advanced visualization tools for further investigation. In Figure, 5 we employ the popular t-SNE methodology for visualization. The two dimensional embedding is presented along with a cluster label denoted by a different color. This has been retrieved by applying the k-means algorithm to the original input (before the dimensionality reduction). It is shown that the clustering result fits very well the resulting visualization discovering clear patterns in the dataset for different values of k.

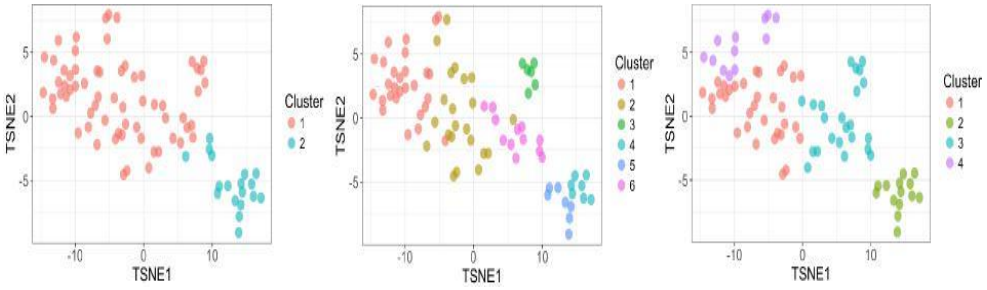


Figure 5: 2-dimensional visualization using t-SNE along with the retrieved clusters from k-means, when k is set to 2 (left) and 6 (center), respectively, and 2-dimensional visualization using t-SNE along with the retrieved clusters of Density Peak, when applied on the 2-dimensional data (right)

In the last part of our experimental analysis in an attempt to make the evaluation more accessible (assuming that we trust the visualization procedure enough), we apply a clustering methodology directly on the 2-dimensional mapping retrieved by t-SNE. This way we guaranty that the clustering result will appear in a more suitable manner. For this purpose, we also employ the Density Peak algorithm [20], which can also automatically estimate the number of existing clusters. The results are reported in Figure 5(right). It is evident that there exist clear clusters in our dataset. With this evidence one can claim that Clinics belonging to different clusters could have different control limits and bounds. The preliminary results we are reporting here can be an extremely helpful tool when designing Health Policies.

4. Conclusions

Big Data offers a great opportunity in the healthcare domain to elucidate biomedical research fields such as the healthcare fraud detection. Frauds in the health domain constitutes an important issue for both states and citizens since it holds a significant percentage of the annual healthcare expenditure globally. Nowadays, where we are in Big Data era and ML approaches have a recent advent, there is the potential for new computational tools able to handle the challenges of fraud detection in healthcare.

Our analysis based on clinical data from EOPYY, focusing in investigating the Clinics behavior with respect to their hospital expenditure. Outcomes indicates that it is obvious that there are clear patterns regarding the Clinics found in our dataset. We now have enough evidence to claim that Clinics that belong to different clusters should be examined under different circumstances. For example, control limits and bounds for Clinics could be scaled according to the cluster they belong to.

5. Acknowledgment

This project is funded by the International Research Project,"Collective wisdom driving public health poli¬cies - CrowdHEALTH", in terms of the

European Commission program Horizon 2020 from March 2017 to February 2020.

References

1. Xue-Wen Chen and Xiaotong Lin. Big data deep learning: challenges and perspectives. *IEEE access*, 2:514–525, 2014.
2. Chunhe Shi, Chengdong Wu, Xiaowei Han, Yinghong Xie, and Zhen Li. Machine learning under big data. In *6th International Conference on Electronic, Mechanical, Information and Management Society*. Atlantis Press, 2016.
3. Lidong Wang and Cheryl Ann Alexander. Machine learning in big data. *International Journal of Mathematical, Engineering and Management Sciences*, 1(2):52–61, 2016.
4. Weiqi Wang and Eswar Krishnan. Big data and clinicians: a review on the state of the science. *JMIR medical informatics*, 2(1):e1, 2014.
5. Anshu Sinha, George Hripcsak, and Marianthi Markatou. Large datasets in biomedicine: a discussion of salient analytic issues. *Journal of the American Medical Informatics Association*, 16(6):759–767, 2009.
6. Nicole F Stowell, Martina Schmidt, and Nathan Wadlinger. Healthcare fraud under the microscope: improving its prevention. *Journal of Financial Crime*, 25(4):1039–1061, 2018.
7. Guido van Capelleveen, Mannes Poel, Roland M Mueller, Dallas Thornton, and Jos van Hillegersberg. Outlier detection in healthcare fraud: A case study in the medicaid dental domain*. International journal of accounting information systems*, 21:18–31, 2016.
8. Pedro A Ortega, Cristián J Figueroa, and Gonzalo A Ruz. A medical claim fraud/abuse detection system based on data mining: A case study in chile. *DMIN*, 6:26–29, 2006.
9. Hyunjung Shin, Hayoung Park, Junwoo Lee, and Won Chul Jhee. A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications*, 39(8):7441–7450, 2012.
10. Fen-May Liou, Ying-Chan Tang, and Jean-Yi Chen. Detecting hospital fraud and claim abuse through diabetic outpatient services. *Health care management science*, 11(4):353–358, 2008.
11. Melih Kirlidog and Cuneyt Asuk. A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences*, 62:989–994, 2012.
12. Mohit Kumar, Rayid Ghani, and Zhu-Song Mei. Data mining to predict and prevent errors in health insurance claims processing. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining,* pages 65–74. ACM, 2010.
13. Qi Liu and Miklos Vasarhelyi. Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. In *29th World*

Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia, 2013.

14. Tahir Ekina, Francesca Leva, Fabrizio Ruggeri, and Refik Soyer. Application of bayesian methods in detection of healthcare fraud. *chemical engineering Transaction*, 33, 2013.
15. MingJian Tang, B Sumudu U Mendis, D Wayne Murray, Yingsong Hu, and Alison Sutinen. Unsupervised fraud detection in medicare australia. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, pages 103–110. Australian Computer Society, Inc., 2011.
16. Rasim Muzaffer Musal. Two models to investigate medicare fraud within unsupervised databases. *Expert Systems with Applications*, 37(12):8628–8633, 2010.
17. Chinho Lin, Chun-Mei Lin, Sheng-Tun Li, and Shu-Ching Kuo. Intelligent physician segmentation and management based on kdd approach. *Expert Systems with Applications*, 34(3):1963–1973, 2008.
18. Yin Shan, David Jeacocke, D Wayne Murray, and Alison Sutinen. Mining medical specialist billing patterns for health service management. In *Proceedings of the 7th Australasian Data Mining Conference-Volume 87*, pages 105–110. Australian Computer Society, Inc., 2008.
19. John A. Major and Dan R. Riedinger. Efd: A hybrid knowledge/statistical-based system for the detection of fraud. *Journal of Risk and Insurance*, 69(3):309–324.
20. Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014

Proteogenomics: statistical issues in data integration and prediction

Júlia M Pavan Soler
Statistics Department, University of São Paulo, São Paulo, SP

Abstract

Proteogenomics inaugurates a new phase of multi-omics research in Molecular Biology, seeking to integrate information of large datasets from the genome, transcriptome and proteome to clinical traits. The promise is to identify patient-specific biomarkers, which can be used on the prognostic in precision medicine. However, the expected contribution of this area depends on overcoming several interdisciplinary challenges, ranging from the design of experiments for samples preparation, storage, processing and integration of data to its analysis and interpretation. Brazil, as other countries, starts to dedicate efforts to the proteogenomic analysis of many diseases in order to identify specific and common biomarkers among world populations. Specifically, the Baependi Family Heart Study is one of the largest ongoing efforts for molecular mapping in cardiovascular diseases in our country, which includes Brazilian family information. Statistical approaches in proteogenomics are typically formulated assuming unrelated individuals, and if family structure is present and ignored, such substructures may induce to misleading results. In this talk, in the context of proteogenomics, we will consider flexible methodologies for dimensionality reduction, variable selection and structure learning taking in account sparsity, dependent observations and missing information.

Keywords

Matrix factorization; Varying coefficients; Multi-omics data; Family based design; Complex data

1. Introduction

A relevant issue that is becoming increasingly important in the big and complex data age is data integration. An early version of that trend can be seen in the multi-omics studies, as exemplified by proteogenomics studies, seeking to integrate many sources of information from large datasets measured on a common set of experimental subjects to clinical traits. In general, the integration scope in these studies try to cover the central dogma of the Molecular Biology including data from genome (such as, SNP and CNV platforms), epigenome (such as, methylation data), transcriptome (such as, Microarrays and RNA-seq data) and proteome (such as, LC-MS/MS data) to phenome (phenotype dataset). The Cancer Genome Atlas (TCGA, Weinstein et

al. 2013) project provides a powerful source of such set of data blocks. These cross-platform datasets share common information, but individually contain distinctive patterns. Disentangling between common and distinctive patterns, and also between the noise component, is critically important to perform integrative, discriminative and predictive analysis of these datasets (Smilde et al., 2017; Shu et al., 2018). Whilst single omics analyses, under an unsupervised or supervised scope, are commonly used for dimensionality reduction and selection of relevant features for specific analytical frameworks, the integration of multi-omics information is required to more fully unravel the complexities of biological systems.

The first step on the omics data analysis involves detection and adjustments for undesirable variable effects, which will tend to appear in addition to the measured variable(s) of interest among most, if not all, high-throughput technologies (Leek et al., 2010). Failing for correction of these sources of heterogeneity into the analysis can have widespread and detrimental effects on the study, not only reducing power and inducing unwanted dependence across genes, but it can also introduce sources of spurious signals. This phenomenon is true even for well-designed and randomized studies. For instance, considering whole-genome SNP platforms, Price et al. (2006) applied singular value decomposition to the genotype called data in order to account for systematic sources of variation due to population substructure. In addition, for batch effects correction and normalization in gene expression data, there are many methods based on nonparametric and parametric approaches (Wolfinger et al., 2001; Irizarry et al., 2003; Leek and Storey, 2007; Chen et al., 2011). Further, undesirable effects in mass spectrometry-based proteomics data have been treated by using smooth curves and ANOVA-Simultaneous Component Analysis (Clough et al., 2012; Mitra et al., 2016). Although all these tools are available, for database integration there is no consensus whether the normalization step should be done through uni or researchers need to work directly with the raw data, making normalization and integration in a unique step, which is both statistically and computationally challenging and a topic of current research.

Data integration can be performed through N-integration (variables integration), which consider different omics platforms evaluated on the same samples, or P-integration (sample unities integration), i.e., concatenation across studies on the same variables. Typical techniques for database N-integration use multivariate projection-based methods, as low-rank models, that embed both the sample unities and features of the data blocks into the same low dimensional vector space (Lê Cao et al., 2009; Tenenhaus et al., 2011, 2014; Ray et al., 2017). These low dimensional vectors enable effective data analytics, such as clustering, visualization and missing value imputation. In addition, these vectors are latent variables or scores possibly representing

biologically relevant molecular signatures and their analysis can suggest novel biological hypotheses.

Another class of N-integration techniques is based on a flexible regression framework. Under an unsupervised approach, probabilistic graphical models (PGMs) can be used for learning relations among multiple variables (Meinshausen et al., 2006). Tenenhaus et al. (2014) proposed a generalized canonical correlation analysis for N-integration with loads in the optimization problem defined in terms of the connections in a PGM. In addition, supervised N-integration can be performed by incorporating varying coefficients (Hastie and Tibshirani, 1993) into the regression model, with the multi-omics integration oriented for prediction of clinical outcomes. In this context, Ni et al. (2018) proposed a Bayesian hierarchical varying-sparsity regression model and apply for genomic and proteomic data integration to be prognostic for the patient's survival time.

Further, the P-integration of independent data sets measured on the same common set of variables (omics data) can be a useful opportunity to increase sample size and gain statistical power. The main challenge in this case is to prevent the analysis from systematic heterogeneities arising from the different sources of variation, as those coming from different protocols. For instance, batch and multi-center effects are unwanted variation, which often acts as strong confounders in the P-integration analysis. Such effects may lead to spurious conclusions if they are not accounted for in the statistical model.

Despite the recent progress made in the area of multi-omics integration, the methods assume independent observations (unrelated individuals), and if family structure is present and ignored in the analysis, such substructures may induce artefactual results for data integration. For instance, in the context of uni-omics data, specifically considering large pedigrees and high dimensional SNP-genotype data, de Andrade et al. (2015) obtained valid principal components estimators and showed that the latent variables taking into account the family structure are more informative than those ignoring such substructure. Ribeiro and Soler (2018), who proposed a probabilistic graphical model for learning relationships among multiple variables from family data, also consider the impact of clustered observation at the analysis. The outline of this work is as follows. First, we will review and discuss unsupervised and supervised multi-omics data integration methods, under the assumption of unrelated samples. Subsequently, we will consider family based designs, incorporate dependence among related individuals and exploit how the covariance matrix among variables is decomposed into genetic and environmental components. Finally, we will discuss the advancement of data integration methods to take into account family structure present on the data.

2. Methodology

A detailed review of multi-omics integration is presented by Huang et al. (2017). All efforts are dedicated to fully account for the uncertainties and heterogeneities in the datasets. Figure 1 shows a schematic representation of the datasets structure involved in Omic's studies. Based on matrix factorization approaches, unsupervised and supervised analysis have been used. In R package, mixOmics (Lê Cao et al., 2009; Rohart et al., 2017) is a powerful resource for integration of multi-omics datasets. In this case, multivariate projection-based methods are proposed to summarise datasets, $X_{n\times p}$, by latent components or scores ($F_{n\times m}$) and loadings ($W_{p\times m}$), such that$X \approx FW'$, m ≤ min(n, p). To properly do data reduction, different optimization problems are formulated to attain objective functions. For unsupervised uni-omics analysis, principal components or its improved version via independent components are used, and for unsupervised multi-omics, generalized canonical correlation can be a useful strategy. Considering supervised contexts, discriminant analysis combined with partial least square have been proposed. In all cases, regularised and sparse solutions are required.

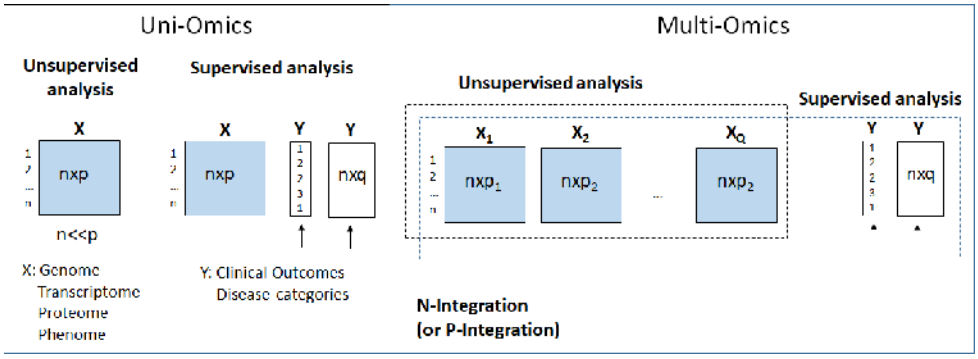


Figure 1. Schematic representation of datasets integration in Omics studies.

Regression models are powerful tools for supervised multi-omics integration. Ni et al. (2018) proposed an interesting varying coefficients regression model, which allow integration of multi-omics datasets driven for prediction of target outcomes. The model is flexible to take in account subject-specific coefficient estimation, i.e, on the patient level. Under regression formulation, regulatory axes given by proteomic (X_1) and genomic (X_2) data are connected to build clinically relevant prognostic through $Y_i \approx \sum X_{1ij}\beta_j(X_{2ij})$, where the varying coefficients $\beta_j(X_{2ij})$ define gene-protein interactions by adopting smooth functions of X_{2ij}.

All of those methods assume independent observations, and are not applied for family-based data, which are very common in genomic studies. Family data are mainly analysed using mixed model approaches that allow including familial dependences among observations. For based family

multivariate data, let X_f be a vector for all p variables and all members of the f-th family, with covariance matrix given by $\Omega = 2\phi_f \otimes \Sigma_g + I_f \otimes \Sigma_e$, where $2\phi_f$ is the kinship matrix for family f, Σ_g and Σ_e are $(p \times p)$ covariance matrix associated with polygenic and error component, respectively, and $\otimes$ is the Kronecker product. Oualkacha et al. (2012) obtained MANOVA based estimators for these covariance matrices. De Andrade et al. (2015) obtained principal components of heritability for reduction of genomic dataset in terms of ancestry scores. Different scores can be obtained from family data by operating on the covariance components, i.e., Σ_g, Σ_e, $\sum_e^{-1}\Sigma_g$ as well as $\Sigma = \Sigma_g + \Sigma_e$. Following this idea, Ribeiro and Soler (2018) proposed to learn polygenic, environmental and total graphical models from family dataset exploiting Σ_g, Σ_e and $\Sigma = \Sigma_g + \Sigma_e$. The authors also exploit to learn the multivariate relations among variables based on a univariate polygenic mixed models framework.

For multi-omics integration in family data, we are extending the multivariate projection-based methods available for independent observations to include familial dependences. Under quadratic solutions, in $\Re^{p\times p}$, it is performed considering the factorization of the polygenic and environmental components of the covariance matrix. In addition, for rectangular solutions, in $\Re^{n\times p}$, N-integration can be performed structuring data matrix through ANOVA-simultaneous component analysis (Smilde et al., 2005) and then building the reductions on the components of the data.

3. Results

Figure 2 shows two representations of n observations clustered in family structure. In (a) it is assumed independent observations, where the principal components are extracted from covariance matrix Σ. In (b) familial dependences are taking in account, where the principal components are extracted from matrix $\sum_e^{-1}\Sigma_g$. Different colors are used to discriminate members from different families. The uni-omics dataset correspond to genotype information obtained from SNP platform (Affymetrics 6.0). A detailed description of the dataset is in de Andrade et al. (2015). The figure illustrates the impact of modelling family structure on the reduction analysis. When familial dependence is used more adaptive representation of the data is obtained, allowing discriminate members between the ancestry arms found in the analysis.

Figure 3 shows probabilistic graphical models learned from family data considering multiple phenotypes extracted from the Baependi Heart Study (Oliveira et al., 2006; Egan et al., 2016). In the figure, vertices represent variables and the connections indicate partial correlations between variables. Important differences are found on the relations obtained from patterns coming from the polygenic, environmental or total covariance matrices.

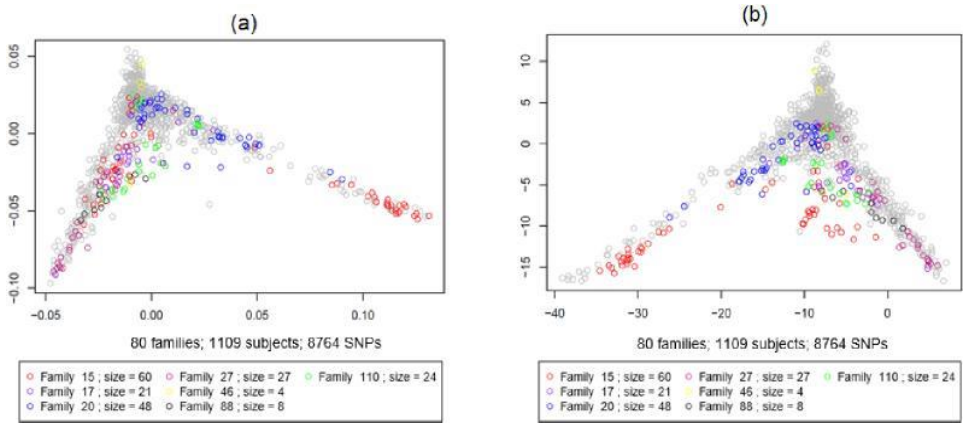


Figure 2: Representation of observations clustered in family structure. In (a), principal components were obtained under independent observations assumption. In (b) principal components are obtained by assuming familial dependences.

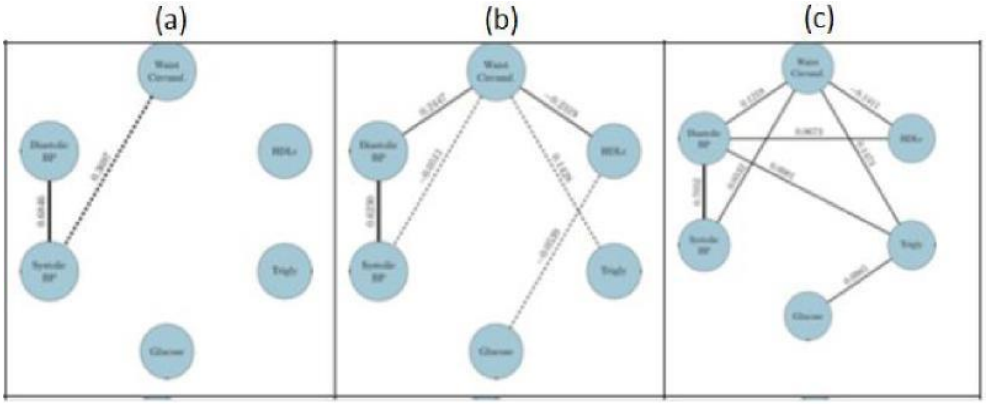


Figure 3. Probabilistic graph models to structure learning from family data. Vertices are metabolic syndrome variables: waist circumference (cm), diastolic blood pressure (mmHg), systolic blood pressure (mmHg), fasting glucose (mg/dL), triglycerides (mg/dL) and HDL-cholesterol (mg/dL). Connections indicate partial correlation between variables. In (a), polygenic covariance matrix, Σ_g, is analyzed. In (b), environmental covariance matrix, Σ_e, is used. In (c), the total covariance matrix, $\Sigma = \Sigma_g + \Sigma_e$, is used.

4. Discussion and Conclusion

It is widely recognized that integrative multi-omics analysis holds an important role for precision medicine. Despite the recent progress in the area, data integration remains a challenge, requiring combination of several software tools, mainly through bioinformatics pre-processing procedures, and extensive statistical expertise to appropriate account for the properties of heterogeneous data. To fully account for the uncertainties, data structure should be taking in account on the analysis, as integration of unsupervised or supervised datasets, N-integration or P-integration, big-n problem, independent versus dependent observations, etc. All of these topics impose challenges for conduction the analysis.

The main expected result in datasets integration is the representation of the observations under a reduced dimension, which is committed to optimizing any objective function that establishes relations among the datasets. Such relations can be based on covariance matrices or prediction functions, according to unsupervised or supervised proposals, respectively. Here we focused mainly in methods derived from matrix factorization and regression models. Most of the analyses available consider independent observations, but several multi-omics studies are based on family data that impose familial dependences among observations.

For multi-omics integration in family data we are considering strategies that decompose the problem to polygenic components integration and environmental components integration. It is a direct extension of the need to include random effect when analysing data with dependencies. Each data block is decomposed into two covariance matrices modelling different types of variation, one due the polygenic random effect, that is sharing among members from the same family and represents among-family variation, and another due the error random effect (environmental), that is the within-family variation. Then, it is performed low-rank approximation of the polygenic variation across the blocks, and low-rank approximations of the environmental variation components. The rational of our approach have been used in other contexts. Feng et al. (2018), addressing the matrices decomposition problem in datasets integration, proposed the angle-based joint and individual variation explained method that allow to compute block scores, block loadings, global loadings and global scores. We are working on the computational implementation of our methods by using the R package facilities.

References

1. Chen, C et al. (2011) Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PloS One* 6(2): e17238.
2. Clough, T et al. (2012). Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. BMC Bioinformatics 13(Suppl 16): S6.
3. de Andrade, M et al. (2015). Global Individual Ancestry Using PCs for Family Data. *Human Heredity* 80: 1-11.
4. Egan, KJ et al. (2016). Cohort profile: the Baependi Heart Study—a family-based, highly admixed cohort study in a rural Brazilian town. *BMJ Open 6*: 1:8.
5. Feng et al. (2018). Angle-based joint and individual variation explained. Journal of Multivariate Analysis 66: 241-265.
6. Hastie, T.; Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society*, Series B (Methodological): 757-796.

7. Huang, S.; Chaudhary, K; Garmire, L.X. (2017). More is better: Recent progress in Multi-Omics data integration methods. Frontiers in Genetics 8, Article 84: 1-12.
8. Irizarry, RA et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-264.
9. Lê Cao, KA; González I; Déjean, S. (2009). IntegrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* 25: 2855-2856.
10. Leek, JT; Storey, JD. (2007) Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PloS Genetics* 3 (9): e161.
11. Leek. JT et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11(10): 1-15.
12. Meinshausen, N; Bühlmann, P. et al. (2006). Highdimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34: 1436-1462.
13. Mitra, V et al. (2016). Identification of Analytical Factors Affecting Complex Proteomics Profiles Acquired in a Factorial Design Study with Analysis of Variance: Simultaneous Component Analysis. *Analytical Chemistry* 88: 4229-4238.
14. Ni, Y. et al. (2018). Bayesian Hierarchical Varying-sparsity Regression Models with Application to Cancer Proteogenomics. *Journal of the American Statistical Association* 0(0): 1-13, Applications and Case Studies.
15. Oliveira, CM et al. (2008). Heritability of cardiovascular risk factors in a Brazilian population: Baependi Heart Study. *BMC Medical Genetics* 32:1-8.
16. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38: 904-909.
17. Oualkacha, K. et al. (2012). Principal Components of Heritability for High Dimension Quantitative Traits and General Pedigrees. *Statistical Applications in Genetics and Molecular Biology* 11(2), Article 4:
18. Ray, B.; Liu, W.; Fenyö, D. (2017). Adaptive Multiview Nonnegative Matrix Factorization Algorithm for Integration of Multimodal Biomedical Data. *Cancer Informatics* 16: 1-12.
19. Ribeiro, A.H.; Soler, J.M.P. (2018). Learning Genetic and Environmental Graphical Models from Family Data. In Annals of the XXIXth International Biometric Conference, in Barcelona, Spain, July 8-13th, 2018. (article submitted to Statistics in Medicine)
20. Rohart, F. et al. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* 13(11): 1-19.
21. Smilde, A.K. et al. (2005). ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* 21(132005): 3043–48.

22. Tenenhaus A.; Tenenhaus M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika* 76(2): 257-284.
23. Tenenhaus, A. et al. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics* 15(3): 569-83.
24. Wolfinger, RD et al. (2001) Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. *Journal of Computational Biology* 8(6): 625-637.

Mendelian randomization, causal relationship, and statistical approaches

Mariza de Andrade, PhD
Mayo Clinic

Abstract

In Mendelian Randomization, one needs the outcome (Y), the instrumental variable (IV) and the mediator variable (M). In the field of biostatistics and epidemiology the genetic markers are the IVs that can vary from one to multiple genetic markers (G). Several approaches are used in MR: Structural Mean Models (SMMs) are semi-parametric models that use IVs to identify causal parameters that include multiple instrument variables for multiplicative and logistic SMMs. In this case one can use the generalized method of moments (GMM) estimator. Other approach when one have multiple mediators are to apply the two-stage least squares (2SLS) that consists of two regression stages: the first-stage regression of the exposure on the genetic IVs, i.e., (G –M) regression, the exposure is regressed on the IVs to give fitted values for the exposure (X |G) and the second-stage (X – Y) regression, the outcome is regressed on the fitted values for the exposure from the first stage regression, the outcome can be continuous and binary. One can also use likelihood-based, Bayesian, and semi-parametric methods (that includes the GMM and SMMs). We will use available data from the VTE meta-analysis as well as from the lung cancer.

Keywords

Instrumental Variables; Mediators; Two-stage Methods: Wald method, Missing data

1. Introduction

In Epidemiology studies, the researchers rely on observational data that can lead to confounding and reverse causality. In this paper will introduce concepts of making inferences in causal effects based on observational data using genetic instrumental variables as known as Mendelian randomization (MR) (1). Since children inherit their genomes from their parents at random, which means that reverse causality can be ruled out and genetic variants are not related to the environmental confounder (2). Multivariate MR (MVMR) is an approach that can be used to estimate the effect of two oe more exposures on an outcome (3). The concepts of Mendelian randomization will be introduced that include outcomes, mediators, instrumental variables among others. One of the early example of Mendelian randomization was the levels of C-reactive protein (CRP) and the risk of coronary heart disease (CHD).

However there are many risk factors that may increase the levels of CRP and the risk of CHD. First these factors need to be identified. In this particular situation one of the potential confounders was fibrinogen, a soluble blood plasma glycoprotein, that enables blood-clotting, that belongs to the inflammation pathway. This leads to the conclusion that the elevated levels of CRP are caused to changes in fibrinogen. However by using one genetic variant of Fibrinogen gene, it was concluded that fibrinogen does not play a role in CHD (4). In the era of the genome whole association studies (GWAS), the use of genetic variants as the instrumental variables to estimate the relationship between the mediators or confounders and the outcome variable, the use of Mendelian Randomization in genetic epidemiology turns up to be the way to estimate interactions (1). One can also include other omics information such as gene expression, structural variation, pathways from whole genome sequencing (WGS) (5, 6). The advantage to use Mendelian Randomization is that to make causal inferences about modifiable (non-genetic) risk factors for disease and health-related outcomes, where in these studies on can exploit the law of independt assortment, i.e., the inheritance of one trait is independent (or randomized with respect to) the inheritance of other traits as already described. In this paper I will introduce the MR models, causal estimation, and statistical approaches as well as presented real data analyses, and the list of software available for Mendelian Randomization.

2. Methodology

There are limitations of observational epidemiology for making causal inference despite the fact that conventional observational epidemiology has made relevant contributions to understanding disease etiology. For example, the identification of the link between cigarette smoking and lung cancer (3), heart disease among others as well as the limitations of randomized controlled trials (RCTs). The most common explanations are confounding by lifestyle and socioeconomic factors or by baseline health status and prescriptions, together with reverse causation and selection bias (4). Mendelian randomization is the term has been given to studies that use genetic variants in observational study of genetic epidemiology to make causal inferences about modifiable risk factors for disease and health-related outcomes (5). The issue of Mendel's law of independent assortment is not always valid due to the fact that independent assortment is generally true for genes found on non-homologous chromosomes; however it is not true for genes found in non-homologous chromosomes mainly if the genes are located close to each other, which lead to the terrn linkage disequilibrium (LD) that represents a departure from the situation that all alleles are in complete independence (LE, linkage equilibrium). Thus, many genetic association including Mendelian randomization studies exploit LD to their advantage by using genetic markers or single-nucleotide polymorphism (SNP) that are in LD with functional

variants. In summary MR study is a study that uses genetic variants that can serve as a possible proxy for an environmentally modifiable exposure to make causal inferences about the outcomes of the modifiable exposure. MR was utilized in multiple studies to determine the causal relationships between exposure and outcomes involving cancer, cardiovascular diseases, among others.

The use of genotype to determine causal inference for the effect of a modifiable (non-genetic) exposure on disease outcome is based on the general theory of Instrumental Variables (IV) analysis where IV is a variable associated with the outcome only though its robust association with an intermediary variable, which is the exposure of interest.

Assumptions of Mendelian randomization studies and IV analysis

An IV is defined as a variable that satisfies the following assumptions:

1. The ***IV*** is associated with the exposure of interest ***E***;
2. ***IV*** is independent of the confounding factors ***M*** that confound the association of ***E*** and the outcome ***Y***;
3. ***IV*** is independent of outcome ***Y*** given ***E*** and the confounding factors ***M***. These assumptions are depicted in the directed acyclic graph (DAG) shown in Figure 1.

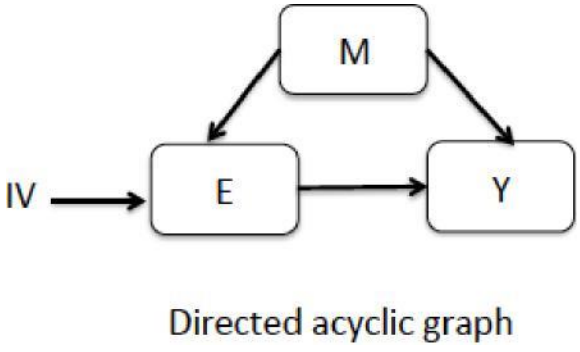


Figure 1. Directed acyclic graph (DAG) for the basic instrumental variables model. IV, instrumental variable; E, modifiable exposure of interest; Y, outcome of interest; and M, the mediators or confounders.

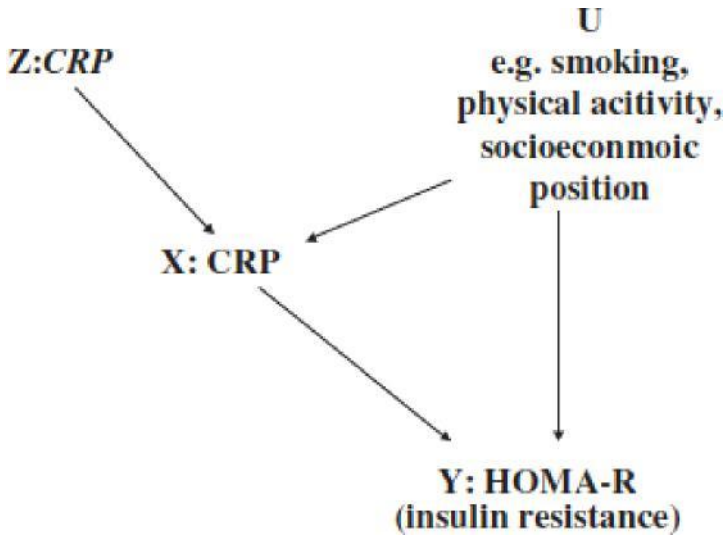


Figure 2: DAG for the effect of circulating levels of C-reactive protein on insulin resistance determined using ***CRP*** as an instrumental variable. XRP, c-Reactive protein genetic variant, the IV; E: CRP, circulating c-reactive protein levels, the modifiable exposure of interest; Y: HOMA-R, homeostasis model assessment of insulin resistance, the outcome of interest; M: (unmeasured or measured with error) confounders.

The three assumptions are sufficient for the simple case of statistical testing, i.e., using genotype as an IV to test the null hypothesis that the modifiable exposure X is not associated with outcome Y. To avoid incorrect inference due to type II errors, the usual aim of MR studies is to provide an estimate of effect with reliable confidence intervals (7). Additional assumptions are needed to estimate a causal effect with IV analysis, i.e., one should assumed the following:

4. All of the associations depicted in Figure 1 are linear and unaffected by statistical interactions. (8)

This assumption is problematic for binary outcomes since the effect estimates are represented as an odds ratio or risk ratio. In the case the one needs to estimate causal effects using IV methods using linear associations and a continuous outcome Y, the IV estimate of the regression coefficient for the effect of exposure (E) on Y is $\hat{\beta}_{IV} = \hat{\beta}_{ZY}/\hat{\beta}_{ZX}$, where $\hat{\beta}_{ZY}$ is the coefficient for the regression of outcome (Y) on the IV (Z), and $\hat{\beta}_{ZX}$ is the coefficient for the regression of exposure (X) on the IV.

The IV estimator $\hat{\beta}_{IV}$ provides an estimate of the causal effect of exposure on outcome, even in the presence of mediators of the exposure=outcome association. Several methods of IV estimation are available where more than one IV and the outcome Y is a numerical variable and associations between variables are linear. The most common used estimator is the two-stage least squares (2SLS) where can be derived by first perform the least-squares regression of the exposure X on the IV(s) Z; then

perform the least-squares regression of the outcome Y on the predicted values from the first regression. In the case that the IV is genetic information, the predicted values are the means of X for each category. There are other methods available such as limited maximum likelihood (LIML) (8) and generalized method of momments (GMM) (9). The MR was also extended to meta-analysis (10,11).

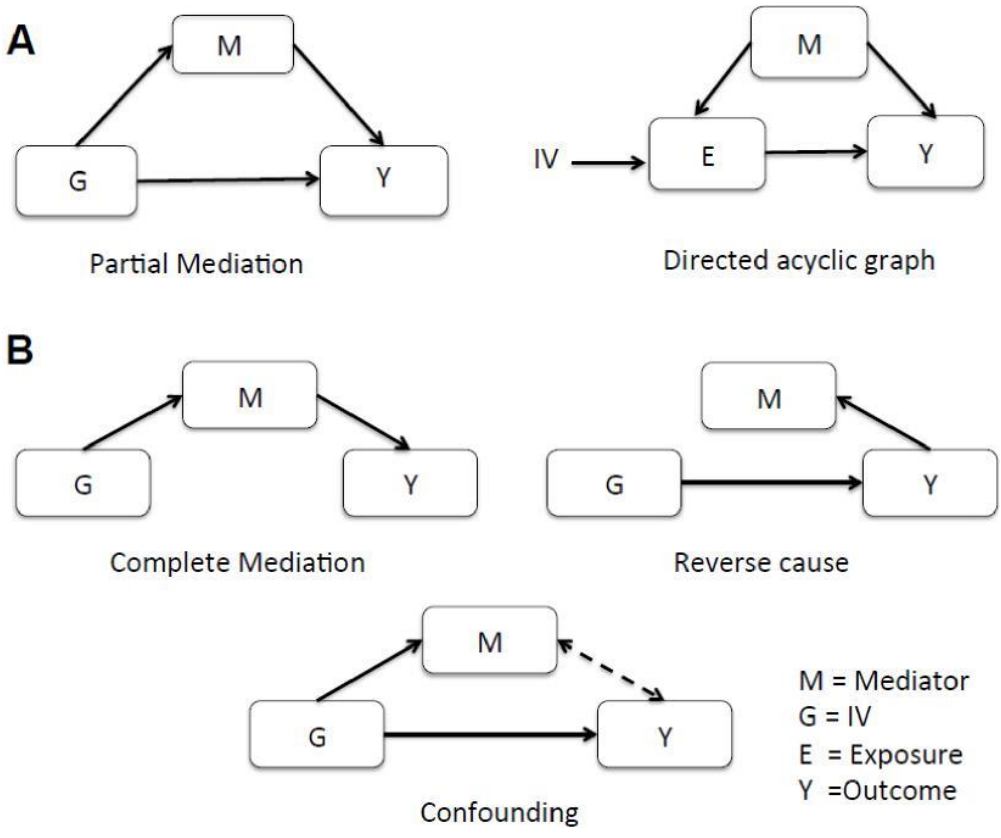


Figure 3: Mendelian Randomization Models

Approaches for Instrumental Variables (IV) analysis (1):

1. Ratio to coefficients method or the Wald method: It is the simplest way to estimate the ausal effect of the exposure on Y (continuous or binary). IV can be discrete, continuous and polytomous.

2.Two-stage least squares (2SLS) method: This consists of two regression stages: the first –stage regression of the exposure on the genetic IVs; the second-stage regression of the outcome on the fitted values of the exposure from the first stage. It is known as Two-stage least squares (2SLS). It can have multiple IVs. The outcome can be continuous and binary and can use Generalized Linear Models (GLM).

3. Likelihood-based methods: Full information maximum likelihood, limited information maximum likelihood, Bayesian methods, semi-parametric methods.
4. Generalized Method of Moments: it is a semi-parametric estimator designedas a more flexible form of 2SLS to deal with problems of heteroscedasticity. (13)

3. Results

We presented the results oftwo studies using our Venous Thromboembolism(VTE) data using MR to identify IV (height) and (BMI) for 2 studies (14,15).

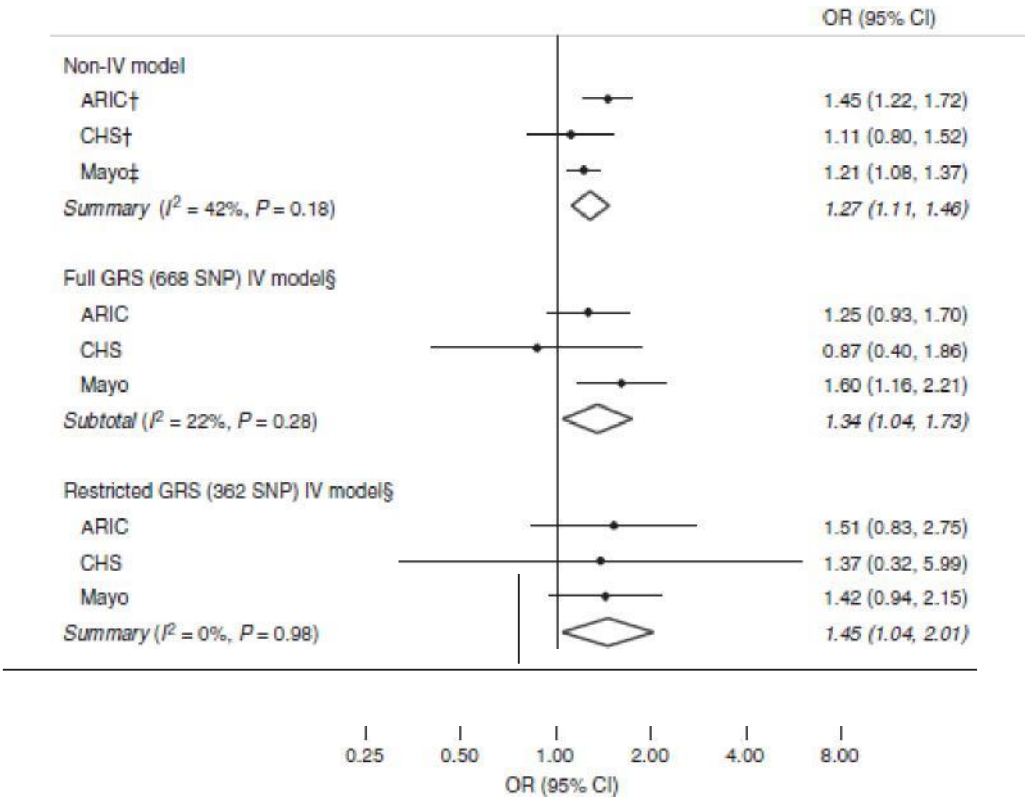
Subject characteristics and their relationship with the genetic risk score (GRS) for height in the Mayo VTE study (1994-2009)

characteristic	Mean (SD) or %		Full GRS (668 SNPs)*		Restricted GRS (362 SNPs)*	
	Cases *N*=1143	Controls *N*=1292	R^2 of GRS	*P*	R^2 of GRS	*P*
Standing height (cm)	172 (11)	171 (10)	0.075	<0.001	0.046	<0.001
Age (years)	55 (16)	56 (16)	0.002	0.09	0.003	0.04
Female sex	50	52	0.001	0.19	0.001	0.41
Body mass index (kg m^{-2})	31 (8)	29 (6)	0.000	0.80	0.001	0.30
Weight (kg)	94 (25)	84 (20)	0.018	<0.001	0.007	0.003
Prior stroke/myocardial infraction	20	11	0.000	0.78	0.000	0.76
Minnesota residency	45	55	0.001	0.23	0.003	0.05

SD, standard deviation; SNP, single-nucleotide polymorphism; VTE, venous thromboembolism.
***Among controls only**

Figure 1: Meta-analysis of non-IV and IV logistic model odds ratio (OR) and 95% CIs of VTE per 10 cm incremented in height in the ARIC, CHS, and Mayo Clinic VTE study. GRS: genetic risk score.

SNP: single-nucleotide polymorphism. † use of logistic regression adjusted for age,sex, BMI, and study site. ‡ Same as † except adjusted for age,sex, BMI, Minnesota resident,and stroke and myorcardial infarction. The modelwas fitted using IV logistic structural mean modelvia GMM estimatorwith the GRS as the IV. Mayo IV estimate are adjusted for the study matching variables;ARIC and CHS IV estimates are unadjusted.



	OR (95% CI)	***P***
BMI Study	Association between genetically predicted BMI and VTE using 95 SNPs from multi-ethnic BMI meta-analysis 1.59 (1.30-1.93)	5.8×10^{-6}
	Association between genetically predicted BMI and VTE using 75 SNPs from European ancestry BMI meta-analysis 1.58 (1.28-1.95)	2.0×10^{-5}
	Association between genetically predicted BMI (95 SNPs) and VTE based on MR Egger regression 1.90 (1.17-3.08)	0.01

4. Discussion and Conclusion

The advantages of MR is that one can add more than one mediator in the analysis to assess the IV. There are several papers recently that use the principle of MR to learn causal biological networks as well as multivariate MR. The following MR software are available:

Bioconductor 3.8: GMRP:GWAS-based Mendelian Randomization and Path Analyses.

R package: Mendelian Randomization ver:0.4.0 date: 2019/03/06- repository: CR

References

1. Burgess S, Thompson SG. Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation. 2015. CRC Press.
2. Hernan M.A., Robbins J.M. Instruments for Causal Inference: An Epidemiologist's Dream? Epidemiology. 2006;17(4):360-72.
3. North T-L., Davies N.M., Harrison S., Carter A.R., Hemani G., Sanderson E., Tilling K., Howe LD. Using Genetic instruments to estimate interactions in Mendelian Randomization studies. BioRxiv, 2019.
4. Sanderson E., Davey Smith G.,Windmeijer F., Bowden J. An examination of multivariate Mendelian randomization in the single-sample and two-sample summary data settings. International Journal of Epidemiology. 2018.
5. Keavney B., Danesh J., Parish S., et al. Fibrinogen and coronary heart disease: test of casuality by 'Mendelian randomization'. International Journal of Epidemiology 2006;35:935-943 doi:10.1093/ije/dyl114
6. Doll R, Peto R, Boreham J., Sutherland I. Mortality from cancer in relation to smoking: 50 years observations on British doctors. British Journal of Cancer 2004; 328:1519.
7. Wang J., Wang, K.,Liu,X.; Sham, P.;Zhao,Z. Next-Generation Sequencing in Human Genetics Studies: Genome Technologies and Applications to Human Genetic Studies. Hum Hered 2017/2018, 83(3):105-106. DOI: 10.1159/000494818
8. Wang, J.;Zheng,J.;Wang, Z.;Li,H.; Deng,M. Inferring Gene-Disease Association by an Integrative Analysis of eQTL Genome-Wide Association Study and Protein-Protein Interaction Data. Hum Hered 2017/2018, 83(3):117-129. DOI: 10.1159/000489761
9. Thomas DC, Lawlor DA, Thompson JR. Re: estimation of bias in nongenetic observational studies using 'Mendelian triangulation'. *Annals of Epidemiology*, 2006;16:675-680.
10. Thomas DC, Conti DV. Commentary: the concept of "Mendelian Randomization". *International Journal of Epidemiology* 2004;33: 21-25.
11. Thomas DC, Lawlor DA, Thompson JR. Re: estimation of bias in nongenetic observational studies using 'Mendelian triangulation'. *Annals of Epidemiology*, 2006;16:675-680.
12. Wooldridge JM. Introductory econometrics: A modern approach. Nelson Education, 2015, chapter 16.
13. Clarke PS, Palmer TM, Windmeijer F. Estimating structural mean models with multiple instrumental variables using the generalized method of moments. *Stat Sci* 2015; 30:96-117.
14. Roetker NS, Armasu SM, Pankrow JS et al. Taller height as a risk factor for venous
15. thromboembolism: a Mendelian randomization meta-analysis. *J Thr Haem* 2017;15:1334-1343. 15.Lindstrom S, Germain M, Crous-bou M et

al. Assessing the causal Relationship between obesity and venous thromboembolism through a Mendelian Randomization study. *Hum Genet* 2017; 136:897¬902.

Using GIS and Official Statistics to support assessments of risk to sustainable development from Environmental Degradation

Daniel Clarke
United Nations, ESCAP Statistics Division[1]

Abstract

Modern methods for risk assessment, in the insurance industry, for example, make use of a variety of sources of economic, social and environment statistics for measurement of exposure, vulnerability and coping capacity in relation to probabilities of future hazards. The Sendai Framework for Disaster Risk Reduction and the UN Sustainable Development Goals recognizes environmental hazards, associated with environmental degradation, as one of the serious risks to sustainable development. Moreover, ecosystems in good condition can boost resilience against potential hazards or other threats to sustainable development. Ecosystem degradation as well as impacts of hazards result together in losses of ecosystem services and increase countries' vulnerability both from material and financial points of view. Based on experience from conducting pilot tests of applications for geospatial data in the Southeast Asia region, this study investigates how statistical evidence can be integrated into environmental management policy-making, especially through mitigating risks from environmental hazards and by quantifying the exposure to potential hazards, in terms of population and financial exposure, which could be used as an important input for conducting cost-benefit analyses for environmental management and sustainable development planning. Using GIS and grid-based assimilation of data to integrate a broad range of datasets, exposure to potential environmental hazards are identified in order to help design policies to counteract or reduce risks to the health an sustainability of these communities. The results of this study include methods for integrating official statistics with earth observation data, which could be applied and adapted by national statistics agencies for sustainability assessments at multiple scales.

Keywords

Environmental Management; Geographic Information Systems; Risk Assessment; Official Statistics

1. Introduction

Use of geographic information system (GIS) technologies for integration of official statistics with geospatial datasets, produced from remote sensing

[1] any views expressed in this paper belong to the author only and in no way should be interpreted as a position of the United Nations.

(or earth observation) imagery, creates powerful opportunities for assessing and managing risk from environmental degradation.

The first version of the National Aeronautic and Space Administration (NASA)'s Landsat programme – which at the time was called Earth Resource Technology Satellite (ERTS-1) – launched on July 23, 1972. Six additional versions of Landsat and many other earth observation satellites have since been launched [1], providing high resolution information for deriving data and statistics for analyses of land cover, land use, qualities of ecosystems, habitats, and for analyses of change over time. The world's national and international space agencies are collaborating to produce and disseminate high quality geospatial data sets for open access and free use for non-commercial purposes, and the statistical qualities of the available data sets are constantly improving.

Case studies were developed as part of an effort to develop methodological guidance and tools to build capacities among national statistical systems in Asia and the Pacific to unlock the potential from integrating existing population and social statistics with the new sources of earth observation data for assessing sustainability of environmental management and for identifying risks from current or potential environmental degradation.

2. Methodology

Risk assessment is a methodology, developed for the insurance industry and for use by governments to reduce risk from catastrophic losses. Each main risk element (see below) is location-specific and thus are commonly mapped and integrated for the assessment using geographic information systems (GIS).

Risk = f (Hazard exposure, Vulnerability, Capacity) [2]

There are three main types of geospatial data that are integrated in order to produce statistics for risk measurement. They are: (i)point statistics (data associated with a specific point location – e.g. GPS coordinates on a map, (ii) vector (or polygon statistics), aggregations by regions (e.g. administrative regions or other defined areas like river basins) and (iii) Grid (or raster) statistics which are data associated with a selected grid system for a geographic study area.

All three types of geospatial datasets can be integrated in GIS to evaluate risks, including integration of traditional sources of population and social statistics and data from earth observation satellites.

Ecosystem condition is defined as the overall quality of the systems' key characteristics and is related to non-market benefits provided to societies via ecosystem services. [3] The main purpose for measuring condition of ecosystems is to track how they are changing over time, identify signs of

degradation, and to develop location-specific assessments of risk and the costs and benefits of intervention.

Degradation to ecosystem condition is defined in the System of Environmental-Economic Accounts Central Framework (ibid) as adverse changes to the condition (or capacity) of ecosystems to deliver ecosystem services. This capacity depends on a complex interplay of natural processes that create habitats in support of life and life-supporting conditions.

3. Geographic Data Integration

Official statistics produced from traditional sources, such as censuses and surveys are aggregated according to administrative regions, which are represented geospatially as polygons (or vector files). These statistics can be integrated with earth observation data sets data sets (typically derived as gridded, or raster data sets) via simple aggregations (in the case of grid data) and extrapolated distribution of data (for administrative-based or point-specific data). A simple example is the distribution of population statistics calculated according to censuses and surveys into a grid system utilizing correlated information from remote sensing, such as locations of residential buildings and other built-up areas.

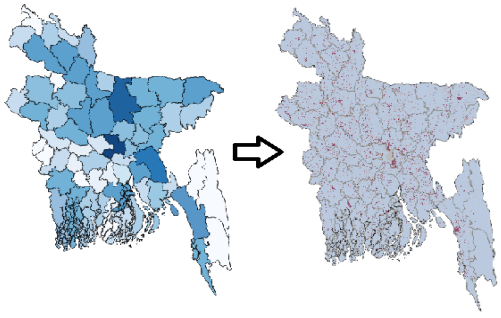


Figure 1. Sample distribution of district level population data using grid-based assimilation

Point statistics (information referenced to a specific point location) likewise are assimilated with other forms of georeferenced information using a range of spatial extrapolation methodologies commonly used by GIS specialists, such as kriging or distribution of values in space based on models of correlations with earth observation data.

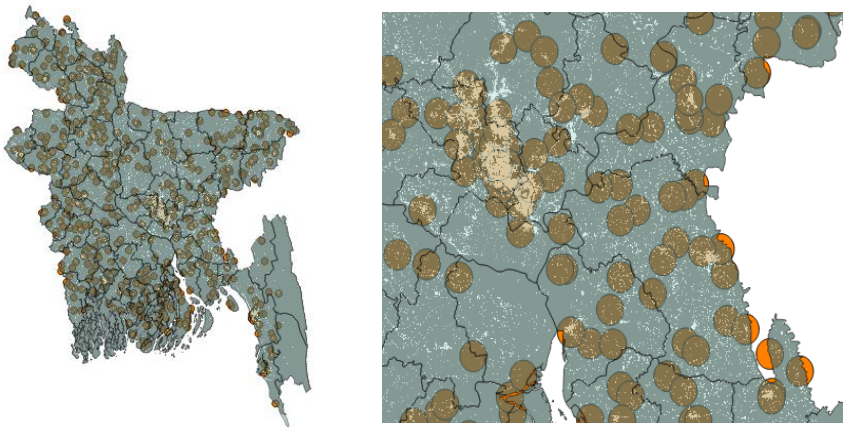


Figure 2. Sample of geo-referenced locations of sample cluster locations in relation to grid-based earth observation data

A common feature of popular methodologies for assimilation between these different types of geo-referenced data is use of a distribution function to produce a smoothed (fuzzy logic) assessment of relationships of variables in space. For these studies, the smoothed assessments represent approximations of probabilities of locations, based on fitted models that are aligned to the aggregated official statistics. The tools produce realistic calculations of proximity and probable areas or sources of risk.

For example, the image below shows a Gaussian smoothed (or blurred) distribution of values from the Global Urban Footprint (GUF) dataset. The result is a range of distributed values as an input for estimation of location (or probabilities of location) of population or other characteristics in the landscape.

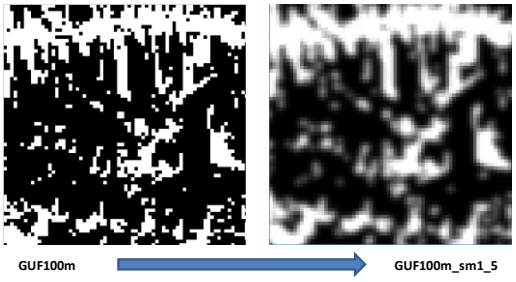


Figure 3. Sample representation of effect of distribution (fuzzy logic) of grid-based values

4. Results

An application of the basic risk assessment framework was developed for a number of cases and for analyses of risk factors at various scales across south and southeast Asia.

Included among the pilot studies is an assessment of relationships between poverty, as a metric related to vulnerability, and exposure to hydrological hazards within the large multi-county region of the Ganges-

Brahmaputra (GBM) River Basin. The assessment involved official statistics on population and poverty from four countries (Bangladesh, Bhutan, India, and Nepal). One requirement for the methodology was flexibility of scale.

A grid-based extrapolation of location of residence of poor households in the GBM was developed utilizing household survey data [4] combined with geospatial data sets, particularly the GUF [German Aerospace Centre] and visible night lights.[NOAA] The modelled extrapolation of location of poverty is overlaid with flood hazard areas, defined according to hazard maps provided by UNISDR [5], providing an isolated view of relationships between vulnerability associated with extreme poverty and exposure to a potential environmental hazard.

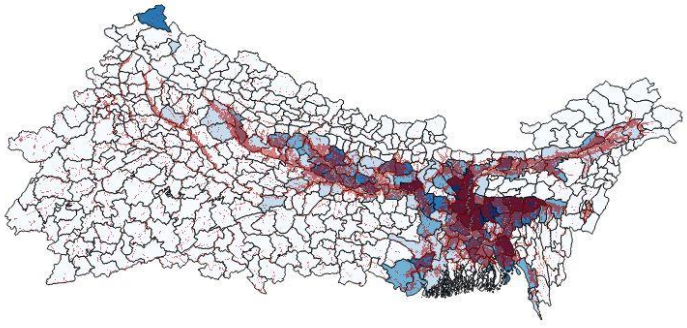


Figure 4. Estimated locations of poverty in the high flood risk exposure areas in the GBM river basin

Interpretations of the results of the integrated risk assessments are dependent on the scale of the analysis. For example, at the broadest scale, the assessment is useful for summarizing the overall extent of the challenge for environmental management in this densely populated region. The results predicted that over 500 million people in the GBM are living in areas exposed to potentially devastating impacts from a major flood, including nearly 100 million living below the international poverty line (approximately $1.90 equivalent purchasing power per day).

However, applying the same data and methodology at more detailed (higher resolution) scale of analysis can be used to reveal other important implications for environmental management. For example, by zooming in to areas near the rivers, 'hot spot' areas of particularly high exposure and high vulnerabilities to environmental hazards can be identified, especially in boundary areas near international borders.

The need for flexibility of scale for analyses of environmental degradation means that standardized tools or recommendations for integration of the input data must be built to include options for customization of scale of analysis by the users.

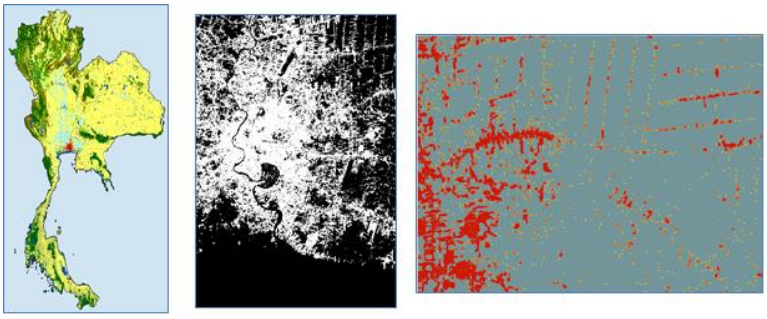
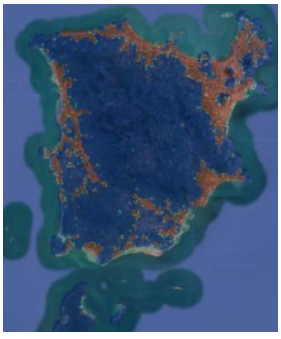


Figure 5: Analysing land cover, land use, and integration with population statistics at different scales for Thailand

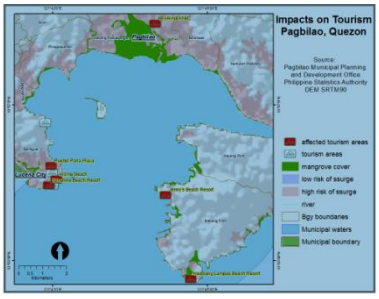
For environmental management purposes, it is useful to analyse the interactions between geospatial datasets with geographic units (shapes) of governance, i.e. to disaggregate or re-aggregate results according to environmental management zones.

Many coastal areas in Asia and the Pacific are important sources of economic wealth. Thus, these areas face a relatively high inherent exposure to environmental risks, which may be increasing due to climate changes and from impacts of increasingly intensive human activities. Through integration of social-economic earth observation data, exposure can be identified and incorporated into costs-benefit analyses and strategic environmental management policies for hot spots.



In some parts of southeast Asia, coastal areas have been temporarily closed to tourism activities as a last available option to allow coral reefs and other coastal ecosystems time to recover from effects of climate change and from over-crowding.[8] Earth observation data can be used as a validation tool to improve the accuracy, timeliness or coverage of traditional sources of data and help provide unbiased assessments for policies that could be political sensitive or have different economic effects for different groups of the population.

In a related example from the Philippines Statistics authority, a similar risk assessment methodology was used to evaluate and communicate the economic importance a specific habitats, in this case mangrove forests, as the source of protection of the local community against storm surge hazard as well as destination for ecological research and sustainable tourism. [9]



5. Discussion and Conclusion

Integration of geospatial datasets with official statistics from traditional sources creates new potential for evidence-based sustainable environmental management, and datasets with comparable definitions can be applied to analyses at flexible scales. The basic risk model can be used as an organizing framework for integration of the full range of types and formats of geospatial datasets for an integrated assessment of environmental degradation for informing sustainable environmental management.

It's crucial that population and social statistics are integrated into these environmental risk assessments because societies are a major influence on the condition of ecosystems everywhere and location of population and economic activities can be used to help policy-makers understand the potential costs and benefits from reducing risks and protecting ecosystems for future generations.

The geographic scale of analyses is crucial. Fortunately, after the datasets have been integrated into a GIS platform, the scale of analyses becomes flexible.

References

1. Kalkhan, Mohammed A. (2011) *Spatial Statistics*. CRC Press Taylor & Francis Group, Boca Raton, Florida, USA
2. ESCAP (2018). *Disaster-related Statistics Framework – Final Draft*, Expert Group on disaster-related statistics in Asia and the Pacific,
3. United Nations (2012) *System of Environmental-Economic Accounting – Central Framework*, United Nations, European Union, Food and Agriculture Organization of the United Nations, International Monetary Fund, Organisation for Economic Co-operation and Development and the World Bank. ISBNI 987-92-1-161563-0, New York, USA
4. Bangladesh DHS 2014, India DHS 2015-16, Nepal DHS 2016, Datasets, Demographic an health Surveys (DHS) Programme, accessed Jan. 2019
5. German Aerospace Centre (DLR) Global Urban Footprint (https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-9628/16557_read-40454/)
6. Night time Imagery from Earth Observation Group, NOAA National Centers for Environmental Information, National Oceanic and Atmospheric Administration, U.S. Department of Commerce. Accessed March, 2019
7. *UNISDR GAR 2015 Risk Data Platform* https://risk.preventionweb.net/capraviewer/main.jsp?countrycode=g15
8. https://www.reuters.com/article/us-southeast-asia-tourism-environment/southeast-asia-closes-island-beaches-to-recover-from-climate-change-and-tourism-idUSKBN1H3209

9. Philippines Statistics Authority (PSA). *Mangroves Forest Inventory and Estimation of Carbon Storage and Sedimentation in Pagbilao.* World Bank Group Global Partnership for Wealth Accounting and the Valuation of Ecosystem Services (WAVES). April 2017

Addressing the issue of missing or non-ideal sampling frames in household surveys in developing countries through remote sensing data

Michael Wild; Brian Blankespoor; Siobhan Murray; Talip Kilic
The World Bank

Abstract

Household surveys are the most important data source on the socio-economic conditions of the population living in Low- (LIC) and Middle-Income (MIC) countries. In most cases the surveys are design based probabilities surveys requiring a sampling frame. However, in many cases, the existing frames do not fulfill the basic requirements of an ideal frame, namely completeness, currentness and informativeness. Surveys based on an inadequate sampling frame may deliver imprecise or biased estimates. Since any information related to sampling errors is based on this frame, the errors related to inadequate sampling frames usually remain undiscovered. The problem of inadequate sampling frame is quite common. High Income Countries (HIC) have an abundance of administrative data to address these problems. However, LIC and MIC countries most likely do not have sufficient high quality administrative data and therefore exclusively rely on their decennial census. To address this considerable limitation, we compare a sample drawn from a census based sampling frame to samples using stratification from satellite data including landcover, gridded population data or share of built-up area for a single province in Malawi. We show that the deviation between the estimates and the true value are within the expected interval for all designs and frames, and that the sampling frame data derived from satellite data, performs either as good, or in some cases even outperforms the pure census-based frame results. Our research therefore provides evidence to support the use of satellite data in the construction of household sampling frames either in combination with census data or even as a stand-alone solution.

Keywords

Sampling Methods, Census, Remote Sensing

1. Introduction

Household surveys are the most important data source on the socio-economic conditions of the population living in Low- (LIC) and Middle-Income (MIC) countries. And in most cases these surveys are design-based probability surveys. One important prerequisite for this type of surveys is a sampling frame.

However, in many cases, these frames do not fulfill the basic requirements of an ideal sampling frame, namely completeness, currentness and

informativeness. Surveys based on an inadequate sampling frame may deliver imprecise or biased estimates. Since any information related to sampling errors is based on this frame, the errors related to inadequate sampling frames usually remain undiscovered.

The problem of inadequate sampling frame is quite common, even in High Income Countries (HIC). But where the latter commonly has an abundance of administrative data to address these problems, LIC and MIC countries most likely don't have this fallback option, since the quality of their administrative data is not sufficient so far. For this reason, the latter group of countries very often relies on the information collected once every 10 years.

To overcome this considerable drawback, we propose the use of remote sensing data as auxiliary information in the sampling frame. To further address the issue of non-available sampling frames, we will also us this type of data as a substitute for the census data.

2. Methodology

Simulation and Frame

To compare the efficiency of the different sampling frames and designs, we will apply an empirical sampling simulation. In this type of (Monte-Carlo style) simulation, either a true or synthetic population is used as the target population. By applying a specific sampling design, and repeated sampling (usually 1000 repetitions) under this design, we can compare the resulting population estimates with the known true population values for each run of the simulation.

The resulting distribution of these estimates is called the sampling distribution, and the average squared deviation from the underlying population value is the Mean Squared Error (MSE) or when taking its square root, the Root MSE (RMSE). To facilitate the comparison, we use the relative version expressed in percentage deviation.

Empirical sampling simulations can be considered as the "[…] ultimate tool for investigators who want to know if one sampling strategy will work better than another for their population." (Thompson, 2013). However, this requires the underlying simulation population to replicate as realistically as possible the target population.

The target variables chosen were collected during the last census. We have chosen variables of sufficient quality as well of different types (i.e. continuous vs. ratio) and with different proportionality to the MOS.

With the simulation set up in this way, we conducted the following experiments and compared the resulting estimates with each other:

i. Sampling from a conventional sampling frame stratified, by the available census variables. This is the baseline scenario, and the commonly applied approach for this type of survey. As mentioned at the outset, this approach

depends very much on the quality and, in particular, the timeliness of the sampling frame, which unfortunately is often not satisfactorily fulfilled.

ii. Sampling from a purely satellite based sampling frame, where the population values are derived from an index of the built-up area. In this scenario, we test the possibility of sampling only from a grid that measures the values of the built-up area within the grid cell. This approach would allow a sample to be taken entirely without the help of the census-based sampling frame.

iii. Linking of the satellite-based population data as well as the land coverage data to the available georeferenced census-based enumeration areas (hybrid approach). The primary purpose of this approach is to update the conventional sampling frame to reflect the required timeliness discussed above, as well as to improve its informativeness by adding the landcover data. This allows us to conduct a higher degree of stratification, resulting in a more balanced distribution of the target variable(s) population variance.

Population Estimates

Population means and totals are estimated from the sample population with the support of design weights. Design weights are the inverse of the selection probability of the final sampling unit.

The final estimate is already described above, however the selection probability in a 2-stage design can be decomposed in 2 components, one for each sampling stage:

p_design= p_1*p_2=m/M*n/N_M

for random selection at both stages, and:

p_design= p_1+p_2= 〖m*MOS〗_m/(∑_(m=1)^M▒〖MOS〗_m)*n/N_M

if 〖MOS〗_m=N_m equation .. becomes:

p_design= p_1+p_2= 〖m*N〗_m/(∑_(m=1)^M▒N_m)*n/N_M

for pps selection at the first stage and random selection at the second stage when household cluster size constitutes the measure of size. A design as this one is called epsm. Each unit has an equal chance of selection and resulting population estimates have a lower variance as. However due to non-response this result hardly holds in practice. For this purpose design weights commonly undergo some post-survey non-response adjustment. One such approach is the calibration of weights to some known population totals as described in the next section.

3. Results

One limitation of the results is related to different time periods in which the data was collected. Whereas the census was conducted in 2011, the satellite imagery comes from 2015 (built up area) and 2016 (land cover). This however only renders the results more prudent, since the estimates derived from or with the support of satellite imagery would subsequently give a more accurate picture, since it would reflect the correct distribution of the target population.

5.1. Sampling from a conventional sampling frame stratified, by the available census variables.

To address the problem of a lack of informativeness in the sampling frame, we will in a first step add the above described landcover types contained in a raster image to the sampling frame at the level of the PSU. In this way we can demonstrate, that already by adding this widely available type of data, we can substantially improve the estimates. One important perquisite for an improvement of the estimates through stratification is a sufficiently strong relationship between the variable(s) of interest and the stratification variable. In the case of the landcover, we decided to estimate the number of housing types. The landcover aggregation at the PSU level was done by calculating the share of crop area (category 4 in Table ..).

To make the most out of the additional information we also used a newly developed allocation algorithm, which optimizes stratification by simultaneously creating and allocating strata, such that the overall variance of the estimate is minimized at a prespecified level for each domain of interest. This algorithm is implemented in R by using the package sampling strata (Barcaroli, 2014).

We will first compare a sample drawn in the conventional way from the census data frame, with a sample making use of the additional stratification. Results are presented in Table 1 below

Target Value & Design	MSE	Est. Pop. Mean	CV%	Deff	n_psu	n_ssu	Est. Pop. Total
Age PPS	1.23	21.14	1.43	1.35	80	12	949363
Age PPS (wrong size)	2.21	21.16	1.92	2.37	71	12	943549
Age Random	1.56	21.11	1.6	1.68	80	12	948448
Age STR	1.56	21.13	1.52	1.58	80	12	952152
Age STRPPS	1.34	21.11	1.37	1.27	80	12	952578
Consumption PPS	0.91	563329.11	0.91	1.16	71	12	220150
Consumption PPS (wrong size)	1.45	563365.35	1.19	2.29	71	12	219830
Consumption Random	1.06	563828.89	1.05	1.59	71	12	219857
Consumption STR	0.62	563819.86	1.01	1.48	71	12	219568
Consumption STRPPS	0.52	564101.1	0.9	1.17	71	12	219749
Employment Ratio PPS	3.57	0.45	4.26	3.21	71	12	950165
Employment Ratio PPS (wrong size)	4.88	0.45	5.16	4.93	71	12	936439
Employment Ratio Random	4.37	0.45	4.38	3.43	71	12	950630
Employment STR	4.02	0.46	4.09	3.12	71	12	953019
Employment STRPPS	3.36	0.45	3.91	2.76	71	12	951303
Population Count PPS	1.38	NA	1.5	Inf	120	12	950962
Population Count PPS (wrong size)	11.99	NA	8.93	Inf	120	12	946087
Population Count Random	5.78	NA	6.14	Inf	120	12	951510
Population Count STR	5.01	NA	5.54	Inf	120	12	951357
Population Count STRPPS	1.44	NA	1.47	Inf	120	12	950941

Table 1: Census based Approach

Incorrect size measures have also been used as MOS to demonstrate the impact on the different estimates. The bigger the relationship with the target value, the larger is the impact of size on the MSE (in percent).

Stratification has been applied by different variables as described in the appendix.

1.1. Sampling from a purely satellite based sampling frame, where the population values are derived from a raster of gridded population data.

There are situations, where there is no frame at all. In this case the option exist to use gridded population data. In our simulation we have used WorldPop, and results are shown bellow.

Target Value & Design	MSE	Est. Pop. Mean	CV%	Deff	n_psu	Est. Pop. Total	Av. Samp. Size	Av. Wpop. Pop
Age PPS	3.96	21.11	3.22	12.54	100	938156	6552	9833
Age PPS (calibrated)	3.29	21.08	0.52	0.33	100	922565	6593	9962
Consumption PPS	2.21	563961.95	1.8	8.97	100	215671	6598	9927
Employment PPS	11.39	0.46	8.96	29.47	100	933408	6626	9948
Employment PPS (calibrated)	8.68	0.46	0	0	100	965882	6630	10004

Table 2: Worldpop only sample

1.2. Linking of the satellite-based population data as well as the land coverage data to the available georeferenced census-based enumeration areas (hybrid approach).

In this section we apply an option, which allows to draw a sample from a combination of geo-referenced enumeration areas and different remotely sensed data.

This situation arises, when either the census data is expected to be severely outdated, or when only the geo-referenced units but no other census data is accessible. We have produced the same population estimates, by either a pure PPS (PPS) design or a stratified PPS (STRPPS) and with or without calibration to known population totals. Calibration has been implemented at the household level, which preserves equal weights for all household members.

The MOS was the ratio of built-up area to the total EA's area. To further improve the estimates, the sample includes an additional stratification, by built-up area into 3 strata, and sampling units are proportionally allocated. Table .. bellow presents the results of this.

Target Value & Design	MSE	Est. Pop. Mean	CV%	Deff	n_psu	n_ssu	Est. Pop. Total
Age PPS	1.87	21.13	1.91	2.48	80	12	955795
Age PPS (calibrated)	0.33	21.1	0.32	0.07	80	12	951375
Age STRPPS	1.67	21.16	1.71	1.98	80	12	929429
Age STRPPS (calibrated)	0.29	21.1	0.28	0.06	80	12	951411
Consumption PPS	1.31	563728.96	1.24	2.27	71	12	221831
Consumption STRPPS 1	1.13	563774.77	1.1	1.8	71	12	215094
Employment PPS	5.37	0.45	5.74	6.24	71	12	949198
Employment PPS (calibrated)	0.26	0.46	0	0	71	12	951542
Employment STRPPS	4.99	0.45	4.87	4.41	71	12	929512
Employment STRPPS (calibrated)	0.21	0.46	0	0	71	12	951293
Population Count PPS	9.59	NA	10.06	Inf	120	12	945090
Population Count STRPPS	7.32	NA	7.53	Inf	120	12	936555

Table 3: Hybrid frame

Results are largely comparable, however for the population count the MSE is considerably larger. Other estimates compare quite well in terms of their MSE and their CV. As expected estimates can be further improved through logistic calibration, with the accompanying decrease in the CV.

Which, however, becomes particularly clear in view of the results is the improvements of the results derived by this MOS over the results derived from an incorrect MOS. In other words the application of recent satellite data should be preferred over the option of outdated population data.

It is also important to note in this regard, that the built-up area is derived from images which took place a few years after the original census, and since we only spatially allocated the units within the enumeration area, but not between them, our estimates are rather conservative, and the improvement

over any wrong MOS as well as the difference between the census and the satellite based MOS will most likely develop further in favor of the latter MOS.

As an additional comparison we have also created an MOS from the 2010 WorldPop dataset, and results are shown in .. bellow.

Target Value & Design	MSE	Est. Pop. Mean	CV%	Deff	n_psu	n_ssu	Est. Pop. Total
Age PPS (worldpop)	1.29	21.12	1.49	1.47	80	12	953002
Employment PPS (worldpop)	4.15	0.45	4.52	3.65	71	12	950801
Employment PPS (Worldpop, calibrated)	0.23	0.46	0	0	71	12	951416
Population Count PPS (WorldPop)	2.81	0	3.31	Inf	120	12	952117
Consumption PPS (WorldPop)	0.92	563706.29	0.96	1.28	71	12	219860

4. Discussion and Conclusion

Our research successfully demonstrated the use of remote sensing data to improve multi-purpose household surveys. In particular when the sampling frame is not accessible or severely outdated, the creation of a sampling frame from remotely sensed data of built-up area or gridded population data may pose a serious alternative and allows for estimates with an acceptable degree of precision. However remote sensing data may also add additional information to the sampling frame, which allows for a more efficient stratification.

With the increase in publicly available satellite date with a high degree of detail we may most likely also see an increase in applications of this approach. Processing of the data could be done with standard open-source software, however the paper is also accompanied by a cloud application which supports the herein discussed approaches through a GUI.

References

1. Barcaroli, G., 2014. Sampling Strata: An R package for the optimization of stratified sampling. Journal of Statistical Software, 61(4), pp.1-24.
2. Carfagna, Elisabetta, and F. Javier Gallego. "Using remote sensing for agricultural statistics." International Statistical Review 73, no. 3 (2005): 389-404.
3. Cochran, W.G., 1977. Sampling Techniques: 3d Ed. Wiley.
4. ESA Climate Change Initiative, 2017, Land Cover project, viewed 22 January 2018, http://2016africalandcover20m.esrin.esa.int
5. Friedl, Mark A., Damien Sulla-Menashe, Bin Tan, Annemarie Schneider, Navin Ramankutty, Adam Sibley, and Xiaoman Huang. "MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets." Remote sensing of Environment 114, no. 1 (2010): 168-182.
6. Jerven, M. (2013). Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It. Cornell Univ. Press.

7. Kilic, T., Serajuddin, U., Uematsu, H. and Yoshida, N., 2017. Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity.
8. Särndal, C.E. and Lundström, S., 2005. Estimation in surveys with nonresponse. John Wiley & Sons.
9. Singh, R., Goyal, R.C., Saha, S.K. and Chhikara, R.S., 1992. Use of satellite spectral data in crop yield estimation surveys. International Journal of Remote Sensing, 13(14), pp.2583-2592.
10. Stevens, F.R., Gaughan, A.E., Linard, C. and Tatem, A.J., 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PloS one, 10(2), p.e0107042.
11. Tiecke, T.G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kilic, T., Murray, S., Blankespoor, B., Prydz, E.B. and Dang, H.A.H., 2017. Mapping the world population one building at a time. arXiv preprint arXiv:1712.05839.
12. Thompson, S.K., 2012. Sample size. Sampling, Third Edition.
13. United Nations. Statistical Division, 2008. Designing household survey samples: practical guidelines (Vol. 98). United Nations Publications.

Implementing a geospatial data strategy in the European Statistical System

Mariana Kotzeva*, Nikolaos Roubanis, Julien Gaffuri, Hannes I. Reuter
Eurostat, European Commission, Luxembourg

Abstract

The European Statistical System is a partnership between Eurostat (the Statistical authority of the European Union) and the National Statistical Institutes and other national authorities of the European Union Member States responsible for the development, production and dissemination of European statistics. The production of statistics relies on the Generic-Statistical-Business-Process-Model, which is enhanced with a Global Statistical Geospatial Framework. This creates an information infrastructure composed of statistical and geospatial information, which is connected and conceptually integrated to spatially enable all statistics throughout the entire statistical production process. A detailed (e.g. 1km grid), comparable (e.g. across countries) and efficient data production process provides the information required for analysis to contribute to the policy decision-making processes in the EU. To facilitate the implementation of the geospatial data strategy in the European Statistical System, Eurostat is active on various levels. At the European level, the "GEOSTAT" projects were launched (1km2 population grid for Europe; standardised, point-based, geospatial, reference framework for statistics; European adaptation of the Global Statistical Geospatial Framework). At the Member State level, separate actions were initiated to align the varying levels of the geospatial data available in the statistical production process. Specific objectives were facilitated within Eurostat by, for example, providing the necessary legal instruments (e.g. Census 2021 grid regulation, Integrated Farm Structure Statistics, etc.) and with the creation of pan-European geographical datasets. The presentation will report on Eurostat's activities related to implementing a geospatial data strategy in the European Statistical System.

Keywords

UN-SDG, grid, Statistical Geospatial Framework, data, strategy

1. Introduction

European and global sustainable development programs increasingly require reliable and relevant information in terms of higher spatial and temporal resolution and increased abilities for spatial and thematic disaggregation. The UN Agenda 2030 and its Sustainable Development Goals (SDGs) is pushing for a closer integration of statistical and geospatial

information. The work on achieving and monitoring the SDGs poses substantial challenges for the statistical and geospatial communities. However, it also offers a unique opportunity to demonstrate the power of statistical-geospatial data integration across a wide range of themes. In response to the growing need to add the "where" dimension in public information and statistics, the statistical and geospatial communities have a common task to build frameworks that can support the production of relevant, accurate and timely information to allow evidence-based decision-making for all levels of society.

One of the key areas of the European Statistical System (ESS) is to harness new data sources comprising Big Data, administrative data and geospatial data. Using data from a range of sources and for multiple purposes, not only requires their integration into a common reference system of harmonised concepts, but also a common location and temporal framework. Therefore, users have not only increased their demand for location information but they also require simpler integration of data across various data sources to use in their analyses. Time and space are universal and well-defined concepts and, hence, can be used to integrate data from a wide range of topics.

The international statistical and geospatial communities recognised this challenge and responded by establishing the UN Expert Group on the Integration of Statistical and Geospatial Information (UN EG-ISGI) to develop a Global Statistical Geospatial Framework (GSGF). At the Sixth Session of the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM), held in August 2016, the five principles of the GSGF were adopted. The GSGF should act as a bridge between statistics and geospatial information, between statistical institutes and geospatial agencies, and between statistical and geospatial standards, methods, workflows and tools. Based on the ESS ambition to broaden the scope of statistical-geospatial integration, Eurostat launched a series of four GEOSTAT projects supporting financially National Statistical Institutes.

To enhance the ESS capability to integrate statistical and geospatial information, three high-level strategic objectives were set:

1. To improve the geographical granularity of statistical products. The benefit is to provide additional breakdowns for more local geographies and thus reveal finer spatial patterns at more local scales. This action increases the possibilities for spatial analysis.
2. To generalise the usage of statistical grids such as the 1km2 resolution grid. Finer and coarser resolutions should also be considered depending on the thematic domain. The main benefit of adopting statistical grids is to remove the bias introduced by statistical units with irregular sizes and shapes (e.g. Openshaw, S. & Taylor, J., 1979) and thus map statistics to the users on more reliable and stable geographies.

3. To produce new "geospatial statistics" based on geospatial data sources such as geographical databases, satellite images, etc. and the combination of these data sources together or with existing statistics. Geographical information Systems (GIS) and related methodologies are key elements to enable the production of geospatial statistics. Such geospatial statistics could address the emerging needs for new products in thematic domains with a strong spatial component (e.g. transport, agriculture, energy and population).

2. Methodology

Embedded into a long-term strategy within Eurostat and the ESS to integrate spatial information and statistics, in 2010 Eurostat launched a series of projects–at both the European and national levels.

2.1 European Level – The GEOSTAT projects

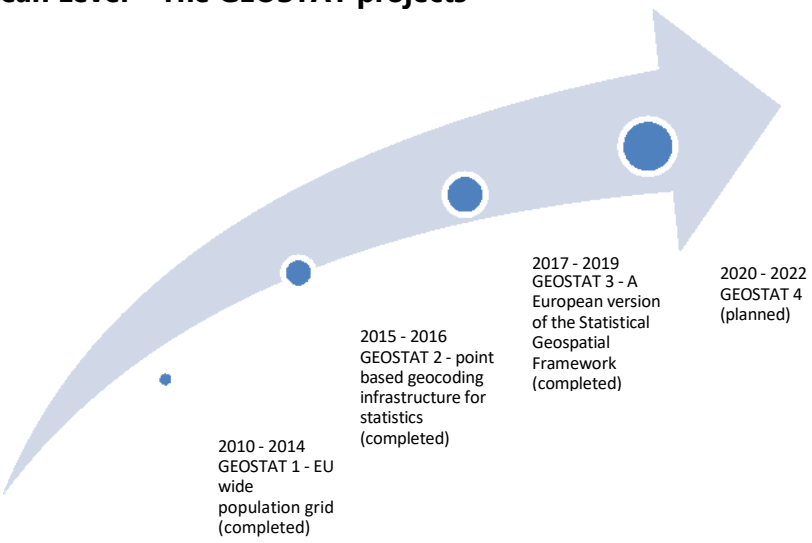


Figure 1. Development of the GEOSTAT projects over time

The first project in the series, as seen in Figure 1, was related to the 2011 Census and aimed to develop a vision and methodological foundation at geocoding various population characteristics into a 1km2 grid dataset. Grid data availability varied across countries. Not all countries in the EU were at the same level. Some countries provided data on grids, while several relied on other methods. The pre-conditions and possibilities for disaggregating data led to specific data specifications. Based on these data specifications the project made use of the Statistical production chain in the NSIs to identify relations and enable coherence by use of GIS. In addition, common processes of quality assessment, knowledge dissemination, confidentiality, business models and continuous data update were addressed or identified.

For all GEOSTAT projects, a common framework using the European Forum for Geography and Statistics (EFGS) was used for the involvement of NSIs and National Mapping and Cadastral Agencies (NMCA); for the exploitation, distribution and dissemination of the results; and for the organisation of conferences and websites (www.efgs.info).

The second project phase (GEOSTAT 2) concentrated on the development of a model for a point-based geocoding infrastructure for a more flexible, sustainable production of geospatial statistics in the ESS and was based on geocoded addresses, buildings, and dwelling registers. The aim was to enable NSIs to provide more qualified descriptions and analyses of the society, economy and environment. While also taking into account the EU context, the project was based on national practices and stakeholders (e.g. the NMCAs) and specific national challenges. GEOSTAT 2 also included an evaluation of the Generic Statistical Business Process Model the (GSBPM) and contributed to the work on the integration of statistics and geospatial information in the framework of UN-GGIM.

The third project (GEOSTAT 3) closely followed on from the previous strategic orientation and focused on the European version of the SGF adapted for the ESS, taking into account the conditions, initiatives and European and national frameworks (e.g. INSPIRE, ESS). The objectives of this third phase were, to produce recommendations on how to implement the ESS-SGF in EU member states within and outside NSIs (outlined in the implementation guide), as well as testing of the ESS-SGF with respect to the UN SDGs. In addition, the application of geospatial statistics and the integration of geospatial information into the statistical production process, within the framework of the GSBPM, were strongly promoted.

The next planned project (GEOSTAT 4 - 2019-2021) is based on the outcome of the GEOSTAT 3 project. The principles of the GSGF and the recommendations of the Implementation Guide and earlier GEOSTAT projects should be integrated into the methodological and quality framework of the ESS and prepare and support the implementation in Member States. The result of this integration work will be the GSGF-Europe, which should incorporate it into the ESS processes and national frameworks.

2.2 National Level: Eurostat support on cooperation between statistics and geospatial information

The degree of cooperation between NSIs and NMCAs varies significantly in the EU Member States. The integration of geographical and statistical information offers important opportunities to maximise the use of data collected for statistical purposes, especially for monitoring purposes. To promote the integration of statistical and geospatial information at national level, with respect to e.g. cooperation or data availability, funding was

provided to NSIs from 2012 to 2018 to work on country specific objectives. Examples of such objectives were: i) integration of geospatial and statistical information during the statistical production process; ii) processes required for continuous, (semi-) automated, regular updates of GIS under the aspect of unique national identifier systems; iii) create new information in the open data initiative and the usage of Big Data streams; iv) investigate spatial and temporal distribution of statistics; v) the usage of Earth observation (e.g. Copernicus Sentinel products) for statistics, and SDG monitoring; or; vi) access to INSPIRE data (e.g. addresses, administrative units, etc.) and services provided by the NMCAs.

3. Results

3.1 European Level – GEOSTAT Projects

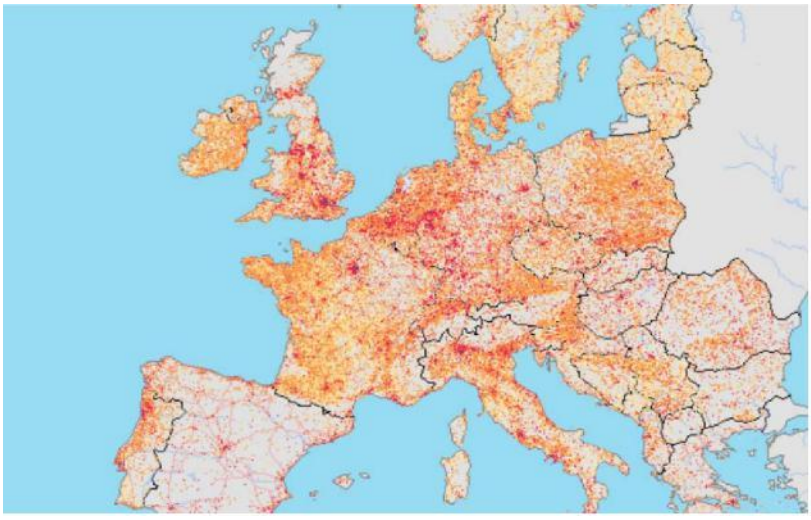


Figure 2. Population density (inhabitants per km2) based on the population grid from 2011. Source: EFGS, Eurostat

The first GEOSTAT project produced a population prototype dataset for 2006 and 2011 as seen in Figure 2. It was based on a developed methodology and data specifications for European population grid datasets (e.g. projection, resolution, confidentiality rules, etc.). Results included, for example, aggregation methods documentation for grid-based and hybrid approaches agreed within the NSIs and NMCAs, a common open data licence, a one-stop data distribution, disaggregation methods for people with unknown place of residence, testing and quality assessment and an operative case study where access to hospitals were evaluated based on the generated data.

The results from the GEOSTAT 2 project promote the application of spatial statistics and the integration of geospatial information into the statistical

production chain, within the framework of the GSBPM and proposed a generic model for national (point based) geospatial reference frameworks for statistics, building on national address, buildings and/or dwelling registers. The project identified a number of key tasks with varying importance within in each NSI such as: i) user requirements; ii) promoting geospatial statistics and the potential of geospatial information; iii) recognising geospatial data sources; iv) assessing data processing capacity; v) specifying geospatial statistics output; vi) creating a flexible production setup; vii) building the geocoded survey frame; viii) obtaining and managing geospatial data; ix) conducting geospatial data quality assessment; x) assessing identifiers to enable correct data linkage; xi) geocoding data; xii) preparing geospatial statistics products; and xiii) assessing data dissemination constraints. In essence, the project delivered a concept plus an NSI directed implementation guideline on how to set up a point-based, geocoding infrastructure for statistical production. Additionally, the outcomes were used to derive proposals for the revision of the GSBPM.

For the GEOSTAT 3 project, the focus moved further towards adaptation of the GSGF to the European specificities. The project developed and tested an implementation guide for a European version of the GSGF. The implementation guide covered the key aspects of statistical-geospatial integration as set out in the GSGF and its five principles. The focus was on comparability of statistical outputs, harmonisation of geospatial data sources and methodologies, and on interoperability of various data sources and metadata. The work was performed in close cooperation with national experts from NSIs and geospatial agencies and with UN-GGIM: Europe through its two working groups (Core Data and Data Integration). In order to assess the soundness of the requirements and recommendations proposed in the implementation guide, the project has undertaken a series of practical and technical tests. Thereby it delivered, for example, good practice cases, feasibility studies on linking SDMX and INSPIRE/OGC (Open Geospatial Consortium) web services as well as testing a selection of SDG indicators.

3.2 European Level – Supporting policy making at European level

Member states produce and maintain national geographical databases, describing with a high level of detail the geography of their territory: constructions, transport infrastructures, hydrography, topography, vegetation cover, addresses, administrative units, public services, energy production facilities, etc. The NMCAs are mandated to produce, maintain and update these databases and various services based on them. These databases have become a reference authoritative information source at national level for many activities having a strong relationship with the geographical location. They are extensively used to support various activities in land management

and civil engineering for example, but also to support national policies with spatial analyses. A key element of the ESS geospatial data strategy is also to support European policies with increased availability and utilisation of these geographical datasets at the European level.

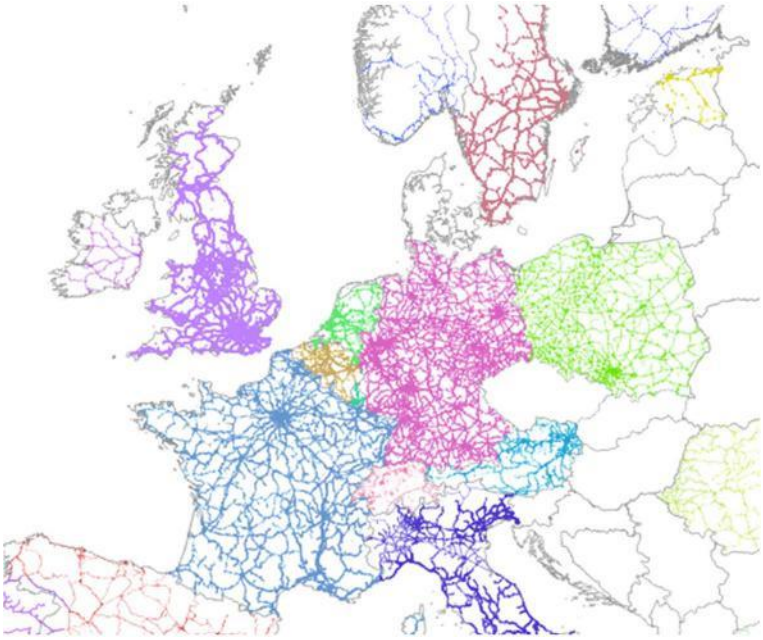


Figure 3. Example of a railway network generated from National Data sources by Eurostat

The high-level challenge to address is how to combine the mosaic of existing national geographical databases of NMCAs into pan-European geographical databases (see Figure 3). Unfortunately, this combination is not as simple as creating a union of the existing national databases: Instead, it requires harmonisation methodologies for data structures and content, specific to geospatial data, to ensure comparability and connectivity across the different borders. Several projects and initiatives have contributed to address this combination (European Spatial Data Infrastructure Network, EuroRegionalMap[1], INSPIRE) and should be completed with actions resulting in a sustainable availability of high quality, pan-European, geographical datasets easily usable for pan-European, spatial analyses. The next step toward this objective is to strengthen the collaboration between Eurostat, NSIs and the NMCAs on geographical data sharing, and to increase Eurostat's capacity to collect, harmonise and combine NMCA national databases into pan-

[1] https://eurogeographics.org/products-and-services/euroregionalmap/

European, geographical datasets following, for example, the UN-GGIM: Europe core data recommendations[2].

The main benefit for Eurostat is to increase its capacity to produce and update comparable geospatial statistics across Europe. Such geospatial analytics are regularly produced by European Commission policy DGs (Poelman & Dijkstra 2015, Poelman & Ackermans 2017). As part of its geospatial strategy, Eurostat also takes a lead in the production of such geospatial statistics and include them in its regular data catalogue.

3.3 Merging statistics and geospatial information at national level

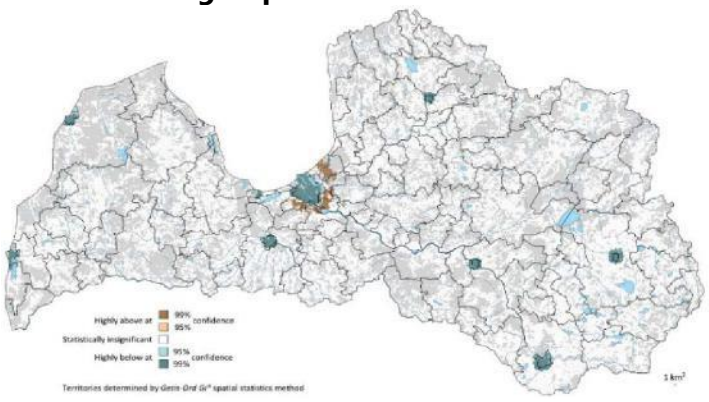


Figure 4. Areas where absolute change in usually resident population is highly above or below the average; (Courtesy of Central Statistical Bureau of Latvia

At the country level, about 50 projects provided manifold results on how Member States aligned their differences with respect to improving their integration of geospatial and statistical information (to check on all the results[3] of these grants please look to the link provided). A selection of the most important results is listed below: implementation of common registers, updating the production systems to include geospatial information (e.g. health, income, tourist, business) at various spatial levels (e.g. grid, statistical units), developing software solutions for dissemination, analysing and developing country specific implementation guidelines, experimenting and implementing linked open data approaches, analytical tasks such as OCR scanning and geocoding old census reports to analyse geospatial movements in a country in support to SDGs (see Figure 4) or commuting patterns. A forthcoming publication by Eurostat aims to document the main results of these grants.

[2] http://un-ggim-europe.org/content/wg-a-core-data

[3] https://circabc.europa.eu/w/browse/9a3fadf7-19f6-4bd4-b3cb-30179950648b

4. Discussion and Conclusion

The paper has shown how the implementation of a geospatial data strategy in the ESS was started, with the aim to improve the geographical granularity of statistical products, in support of evidence-based, policy decision-making. At different levels (e.g. European, National), a variety of actions were performed - from practical implementation over methodological development, up to strategic decisions. Their alignment to international developments was a key component. The success achieved is based on an intensive collaboration between the various actors from the NSIs, NMCAs, the EFGS and UNGGIM: Europe, etc. The development of the GSGF for Europe is one of the key outcomes and will be accompanied by Eurostat's internal activities.

5. Attribution

This paper is based on contributions from the EFGS, different GEOSTAT projects and the merging statistics and geographical information grants, which are all highly valued.

References

1. GSGF Europe - Implementation guide for the Global Statistical Geospatial Framework in Europe
2. Openshaw, S., Taylor, J., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. Statistical applications in the Spatial Sciences, 127-144.
3. Poelman, H., Dijkstra, L., 2015. Measuring access to public transport in European cities. Regional Working Paper WP 01/2015, https://ec.europa.eu/regional_policy/en/information/publications/working-papers/2015/measuring-access-to-public-transport-in-european-cities
4. Poelman, H., Ackermans, L., 2017. Passenger rail accessibility in Europe's border areas. Regional Working Paper WP 11/2017, http://ec.europa.eu/regionalpolicy/en/information/publications/working-papers/2017/passenger-rail-accessibility-in-europe-s-border-areas

Cost-effectiveness of remote sensing for Agricultural Statistics in developing and transition countries

Yakob Mudesir Seid
Statistician, Office of the Chief Statistician, FAO

Abstract

In broader terms, remote sensing enables improvements in the efficiency of agricultural statistics methodology, generates and/or validates some important agricultural related data, allows for more disaggregated data with relative low cost, and provides early information on crop production performance to engender early action. High-resolution optical and radar data are becoming more readily available from approximately 200 earth-observation satellites. However, their use in many countries is rather limited due to mainly cost, data size and technological limitations to use Geographic Information System (GIS) and image-processing software. Remote sensing use for agricultural statistics is cost effective and relates to: i) the sustained decline in image prices; ii) continued improvements in the quality of the available remote sensing data; and iii) the GIS standardisation and image analysis of open-source applications and cloud processing. This paper discusses how best to use remote sensing to improve agricultural statistics by focusing on methodological efficiency, generation and validation of data, disaggregation and early information. Moreover, the costs and benefits of using remote sensing is analysed and the cost effectiveness evaluated.

Keywords

Remote sensing, agricultural statistics, improved estimators, sensor suitability, crop monitoring and yield forecasting

1. Introduction

Since the launch of Landsat series in July 1972, agriculture has been a major beneficiary of satellite imagery. Despite some constraints posted by lack of the required expertise in statistics, image software and budget availability, remote sensing data has played a vital role in improving agricultural statistics (Hanuschank and Delince, 2004; Taylor et al., 1997).

With spatial resolution brought down to 0.5 m (Marchisio, 2014), farmers' declarations could be better-validated (Kay et al., 1997) and data precision on farming (Schumpeter, 2014) would become feasible. On other scales, with remote sensing data, generating land-cover mapping (Defourny et al., 2011; Chen, 2014) and availing data for an early warning systems (Brown and Brickley, 2012; Rembold et al., 2006 & 2013) become easier and efficient.

Footnote: Results from the research studies by the Global Strategy to Improve Agricultural and Rural Statistics

However, it is evident that these improvements come with additional costs and therefore having a clear understanding of the cost efficiency would be vital.

Remote sensing costs can be broadly divided into two categories: a) image purchase and b) data treatment (purchase and maintenance of hardware and software, recruitment of staff, and training etc.).

The cost-efficiency of using remote sensing in agricultural statistics can be evaluated by comparing the gains obtained (usually expressed as a reduction inf sampling variance) to the additional costs involved (cost of imagery, data analysis, staff training, and investment in hardware and software). Hardware and software costs have drastically decreased in recent years due to open-access software that are now widely available. With cloud-based image analysis now a standard, low-cost personal computers and disk storages allow for the analysis of very large image data sets. However, staff availability and competence needs special attention (Latham, 2017). Multi-disciplinary expertise in Geographic Information Systems (GIS), image analysis, statistics, yield modelling, agrometeorology, soil science and crop science will be required and therefore the bulk of costs will be incurred in this respect.

Thanks to initiatives undertaken by the National Oceanic and Atmospheric Administration (NOAA), the U.S. Geological Survey (USGS) and the European Space Agency (ESA), vast real-time freely accessible depositories allow for downloading or online processing thereby tackling what the United Nations Security Council considers the Big Data challenge (2015). Currently, high-, medium- and low-resolution imagery is freely available in raw format and as derived products, such as geometrically (RMS 1.5 pixels) and radiometrically (top of atmosphere) rectified imagery, vegetation indices, regional or country mosaics, and periodic cloud-free coverage. However, the VHR imagery with a ground sampling distance (GSD) lower than 5 m have to be still purchased.

2. Methodology

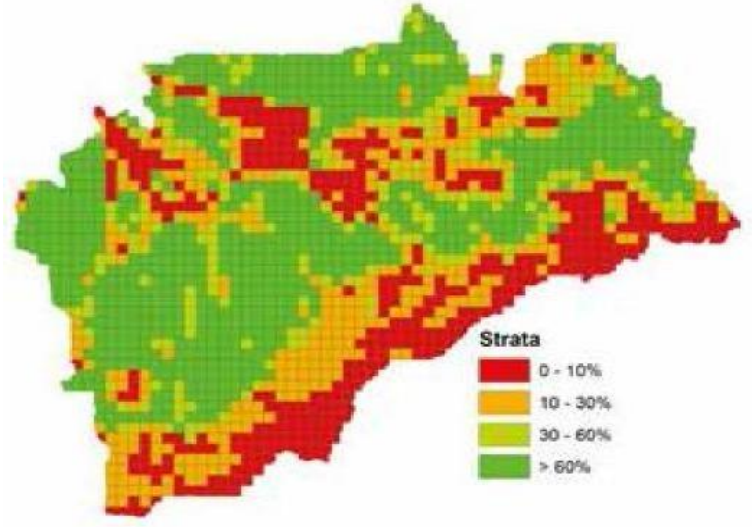
In this paper, the cost benefit for the use of remote sensing in agricultural statistics in relation to its applicability in optimizing the sampling design of agricultural surveys, improving estimators and crop monitoring and yield forecasting will be discussed.

3. Result

3.1 Optimization of sample design

During agricultural censuses or agricultural surveys, the primary activity is to have a clear delineation of the primary sampling units, usually coinciding with Enumeration Areas. An efficient method to define EAs is the use of imagery (having a resolution from 0.5 m to 2 m) in a GIS environment, seeking to subdivide the entire territory into entities with physical limits corresponding

to 50–100 holdings, such that one enumerator can collect the data relating to a subset of EAs during the census period (Geospace, 2007). Based on projects carried out in Lesotho, Namibia, the Seychelles and the United Republic of Tanzania (Loots, 2015), the significant savings that may be achieved in terms of time and fieldwork, largely repay the costs of adopting the required technology (i.e. imagery, GIS, expert consultancies and training). The general quality of undertaking the census also improves, due to better planning, transparency and traceability of the work. In addition, all of the infrastructure created can easily be reused in subsequent efforts, which is of particular interest and benefit if a Master Sampling Frame (MSF) approach is adopted (Global Strategy, 2015b).



One of the recommendations to improve sampling errors is stratification. This is a process of defining several adequate strata that are, internally, as homogeneous as possible, while being as different as possible from one another. Regardless of the master frame selected (list, area, point or multiframe), imagery is of utmost assistance to achieve the intended goals of maintaining homogeneity within the strata by providing up-to-date detailed information that is suitable for modern digital treatment. Additionally, even a rough image classification would enable obtaining a proxy for cropping intensity that can be used to optimize the sizes of the sampling unit (based on spatial correlation) and variable sampling fractions (based on the relation between sampling variance and agricultural intensity; see Benedetti et al., 2015).

For instance, an automated approach based on PSU delineation and stratification by automatic classification of the previous year's Crop Data Layers (CDLs), recently implemented by National Agriculture Statistical Service of the United States Department of Agriculture (USDA/NASS) reduced the total man-months' work per state from 30 to 12, thereby decreasing the cost by a factor of 2.5. This work includes activities related to implementing the

Area Sampling Frame consisting of dividing the territory in PSUs with physical boundaries through visual interpretation of satellite imagery, stratification of PSUs into agricultural classes intensity by photointerpretation and subdividing the sampled PSUs in to segments.

Another example of comparison can be drawn from the area frame employed by China's National Bureau of Statistics in ten provinces (covering 1 652 083 km2). It used a stratified two-stage sampling design with PPS selection in the first stage (the size of each PSU ranging between 1 km2 and 5 km2) and random selection in the second stage (each SSU having a size between 2 ha and 5 ha, with a total sampling fraction in the order of 0.2 percent). For Anhui province (139 400 km2), a sample size of 6 000 segments leads to a CV of 1.3 percent for wheat (2 200 000 ha), of 0.9 percent for middle rice (1 900 000 ha) and of 3 percent for corn (1 000 000 ha). The good level of precision obtained results from the stratification and from the PPS sampling, based on the classification of GF1 and ZY3 Chinese satellite imagery (with a resolution of 2 m). The associated costs amount to US$75 000 per province, and therefore approximately US$0.5/km2.

3.2 Improved estimators

At estimation level, merging data from the ground survey and from satellites is usually achieved through regression or calibration estimators (Global Strategy, 2015a). Gallego et al. (2014) present detailed results for a region (78 500 km2) in northern Ukraine, containing 2.45 million ha of cropland. Ninety, 4 km x 4 km, square segments were field-surveyed in 2010 (with a sampling fraction of 1.8 percent and a field size up to 250 ha). Later, the entire region was covered with MODIS, Landsat5, AWiFS, LISSIII and Rapideye imagery. Image classification was trained on data collected along the road, independently of the area frame segments. For the major crops (wheat, barley, maize and soybean), the respective mean efficiencies amounted respectively to 1.59, 1.54, 1.48, 1.50 and 1.50; therefore, the sensors' performance was approximately equal.

Comparing the cost of the field survey and the cost of imagery (today, the cost of image classification is so low that it can be set aside), the situation changes drastically, because only the two free-of-charge sensors (MODIS and Landsat TM) remain cost-effective, as the purchase price of the other three sensors make them inefficient (AWiFS, LISIIIand Rapideye).

A general context to the study include: First, field size in Ukraine tends to be large, allowing for coarse resolution sensors to compare with finer-resolution ones in terms of classification accuracy. This would not hold in most African or Asian countries, where fields tend to be rather small. Second, the study relied only on the Maximum Likelihood Classification (MLC) method - today, the USDA relies on the decision-tree classification method (See5). Finally, the availability of freely accessible imagery is increasing. MODIS (250 -

500m) or Landsat 8 (15-30m) are rectified or classified products (with resolutions in the range of 250 m to 500 m, and 15 m to 30 m, respectively) freely downloadable in near-real-time. In addition, Sentinel 1 (SAR-GRD, 9-m resolution), Sentinel 2 (10-m resolution) and Sentinel 3 (300-m resolution) are also now available from the ESA hub or Google Earth Engine.

3.3 Crop monitoring and yield forecast

Timely and reliable crop production forecasts are crucial for making informed food policy decisions and enabling rapid responses to emerging food shortfalls. In light of increasing inter-seasonal crop production variability, occasioned by the highly unpredictable climate, increasing food consumption and limited financial resources, decision-makers continue to need reliable crop monitoring system or crop production forecasts, which can provide them with adequate lead time for resource allocation and thus facilitate appropriate response and contingency planning.

Acquiring the crop condition information at early stages of crop growth is even more important than acquiring the exact production after harvest time, especially when large-scale production shortage or surplus happens. Acquiring crop condition as early as possible has great influence on the policy making on the price, circulation and storage of production (Chen Shupeng, 1990, Lin Pei, 1992, Sun Jiulin, 1996).

Regional or national crop growth estimates based on field reports are often expensive, prone to large errors, and cannot provide real-time spatially disaggregated estimates or forecasting crop condition. Moreover, obtaining data through field data collection requires quite a reasonable amount of time while real time information is needed for earlier intervention and early warning systems. In this regard, with the development of remote sensing applications and satellite along with some modelling techniques has become the uppermost approach to monitor crop condition. USDA of the U.S. and EU, as well as FAO, all have built their own crop monitoring systems using different models (Liu Haiqi, 1999, Rassmussen, 1997).

These models require different approaches, skills and data sources. The evaluation criteria for selecting the model to use should be based on the forecasting system's capacity to induce changes in the relevant agents' behaviour, resulting from their perception of risk reduction. Wilson et al. (1981) identified the ideal properties of models: reliability, objectivity, consistency with scientific knowledge, adequacy to scales, minimum cost and simplicity.

Most of the space products for yield monitoring are available free of charge, the costs mainly derive from the running costs of the monitoring system itself.

The Indian's Mahalanobis National Crop Forecast Centre has an annual budget of US$ 1.7 million to issue periodical forecasts for eight crops, to meet

the costs relating to the offices, salaries (of 31 staff members), field surveys (10 percent of the total budget), imagery (20 percent of total budget), hardware (19 workstations) and software (ERDAS, ARCGIS, GEOMATICA STAT licenses). Another example is CROPWATCH of China's RADI. Its annual budget for regional crop monitoring in China and in the major production zones worldwide (covering 31 countries and representing 80 percent of the world production of maize, wheat, rice, and soybean), amounts to US$1.5 million. Manpower costs (15 persons) represent approximately 35 percent of this total budget; those resulting from imagery data amounts only to 20 of the budget, thanks to interinstitutional data sharing.

3.4 Sensor suitability

Several examples confirm the efficiency of using remote sensing for agricultural statistics. An important factor to verify is whether the available satellites are adapted to the predominant sizes of agricultural fields in the various regions of the world.

Agricultural monitoring has various facets (cropland areas, crop type acreages, or yield monitoring at regional or field levels). This enquiry will focus on the regional acreages of major crop categories (such as cereals), of other arable land, of permanent cropland and of permanent grassland.

Table 1. Area (million ha) per field size by category and region.

Region	Cropland	Very small	Small	Medium	Large
Africa	773.0	242.6	394.1	110.4	25.9
Middle East	107.8	9.0	70.0	26.6	2.2
Asia	1411.6	472.5	673.1	179.3	86.7
Central and South	665.8	21.7	154.5	295.9	193.7
Europe	1165.9	14.0	281.2	532.8	337.9
North America	856.7	1.9	68.0	454.5	332.3
Oceania	130.5	0.0	9.8	34.3	86.4
World	**5111.3**	**761.6**	**1650.8**	**1633.9**	**1065.0**

As shown in table 1, the vast majority of agricultural land falls into the small- and medium-parcel size categories. The very-small-parcel category occupies only 15 percent of cropland at world level; Asia and Africa are less favoured from this point of view, as one third of the agricultural areas of both regions fall within the very-small-field size category.

Based on the above, it may be deduced that MODIS and Sentinel 3 sensors are adapted to monitoring the acreage of 21 percent of the world's agricultural land, with increased possibilities in Oceania and North America. However, they cannot be used in Asia and Africa. For medium-resolution satellites, such as Landsat 8, half of the world's agricultural acreages can be monitored; however,

in this case too, Asia and Africa remain disadvantaged, with less than 20 percent of their agricultural areas being able to benefit from coverage. The situation is dramatically enhanced with the arrival of Sentinel 1 (radar) and 2 (optical). As shown in Table 2, below, five of the regions reach a suitability greater than 90 percent and Africa and Asia are close to 70 percent. Total suitability is reached with VHR satellites, although they are unaffordable for all statistical systems under examination.

Table 2. Satellite resolutions and relative compatible percentage of cropland

Region	**Spot/ Rapideye**	**Sentinel 1&2**	**Landsat8 /AWiFS**	**Modis/ Sentinel 3**
Africa	100	69	18	3
Middle East	100	92	27	2
Asia	100	67	19	6
Central and South	100	97	74	29
Europe	100	99	75	29
North America	100	100	92	39
Oceania	100	100	92	66
World	100	**85**	**53**	**21**

It is noteworthy that image resolution and cost are not the only limiting factors. In tropical zones, cloud coverage can seriously hamper the percentages reported above, except for Sentinel 1 (high-resolution) and RISAT 1 (mediumresolution).

4. Discussion and Conclusion

Three main factors support the cost-effectiveness of remote sensing for agricultural statistics. The decrease in image prices, as free-of-charge long-term systems are secured by NASA and ESA at the resolutions required for crop yield monitoring (METOP, MODIS, Sentinel 3) and acreage estimation (Landsat 8, Sentinel 1 & 2). Quality is improving in terms of guaranteed long-term availability, image resolution (up to 10 m), frame size (up to 290 m x 290 m), revisiting time (up to five days) and the number of radiometric channels (above ten). Finally, open-source applications have become the standard in GIS and image analysis, as well as in access to remote cloud processing tools (hardware and software, such as Google Earth Engine).

Having an early information on yield and production estimates is essential to food security and market monitoring. To this end, remote sensing assures timely forecasts while greatly minimizing the cost of fieldwork.

Remote sensing attains cost efficiency in crop acreage estimation. Although small field sizes remain a limiting factor in 70 countries, the opportunity remains for at least 125 countries to envisage a successful use of remote sensing for crop acreage estimation in a given season. In addition,

archive imagery in sampling design optimization is used in most national statistical offices, even when a list frame approach (both for censuses and surveys) is adopted.

References

1. Gallego, F.J., Kussulb, N., Skakunb, S., Kravchenkob, O., Shelestov, A. & Kussuld, O. 2014. Efficiency assessment of using satellite data for crop area estimation in Ukraine. International Journal of Applied Earth Observation and Geoinformation, 29: 22–30.
2. Global Strategy to improve Agricultural and Rural Statistics (GSARS). 2015a. Spatial Disaggregation and Small-Area Estimation Methods for Agricultural Surveys: Solutions and Perspectives. Technical Report Series GO-07-2015. GSARS Technical Report: Rome.
   2015b. Handbook on Master Sampling Frames for Agricultural Statistics: Frame Development, Sample Design and Estimation. GSARS Handbook: Rome.
3. Global Strategy to improve Agricultural and Rural Statistics (GSARS). 2017. Handbook on Remote Sensing for Agricultural Statistics. GSARS Handbook: Rome
4. Latham, J. 2017. Organization, resources and competences. In Delincé, J. (ed.), Handbook on Remote Sensing for Agricultural Statistics (chapter 8). Global Strategy Handbook: Rome.
5. Loots, H. 2015. The use of Hexagon's Smart Client for Census software for the demarcation of census enumeration areas for the 2016 Population and housing census in Lesotho. Paper prepared for Geomatics Indaba: Conference and exhibition of surveying, geospatial information, GIS, mapping, remote sensing and location-based business, 11–13 August 2015. Gauteng, South Africa.

The accuracy and relevance of GDP measures in a digital economy

Hui Wei[1]; Yafei Wang[2,*]
[1] Tsinghua University, Beijing, China
2 Beijing Normal University, Beijing, China

Abstract

Digital economy has raised has raised questions about the conceptual basis of GDP and output, and whether current compilation methods are adequate to capture them. To this end, this paper discusses essential aspects for the measurement: (1) key features and dynamic natures of GDP to provide a basis for further account for the missing output in the digital economy; (2)a consistent and complete concept is well defined to satisfy the gap of digitalisation definition; (3) difficulties in measurement areas including quality changes ta unprecedented pace, free content and services, nonmarket production, new platforms connecting supply and demand, and globalization and impact on national accounts. In addition, the phenomenon of productivity slowdown is given major explanations: (1) measurement errors: services in particular;(2) low investment both in physical and human capital; (3) wide disparities in productivity between leaders and laggards; and (4) digital technologies are not so powerful as previous technological advances. Finally, this paper represents some research questions for measurement in the digital economy: (1) Is SNA still relevant for the changed economic environments; (2) What are the challenges for capturing changes in the economy using SNA? (3) Has mismeasurement problem worsen over time; (4) What are the promising ways moving forward.

Keywords

Digitalisation; Gross Domestic Product; Economic Measurement

1. Introduction

Digital economy refers to an economy that is based on digital computing technologies, usually people perceiving this as conducting business through markets based on the internet and getting benefits of the digitalisation in everyday life. However, the digital economy increasingly intertwining with the traditional economy, has created some new measurement challenges for macroeconomic statistics and may have exacerbated some older ones. This is known as “mismeasurement hypothesis”.

There are significant concerns that these production and benefits of digitalization may not be appropriately reflected in official statistics. Statistical agencies are typically unable to measure the output that result from the

introduction of new goods and services. This has been making responses by international statistics community, for example, OECD-IMF collaboration has made efforts to address immediate concerns about the potential scale of GDP mismeasurement in key areas where mismeasurement is often suspected.

Another bunch of studies of digital economy is focuses on challenges to mismeasurement for the productivity slowdown. The productivity slowdown has occurred in dozens of countries, and a number of commentators and researchers have suggested that this slowdown is at least in part illusory, because real output data have failed to capture the new and better products of the past decade.

New 'Solow Paradox' and GDP mismeasurement interpretation

(1) Three forces shaping the digital era (Brynjolfsson & McAfee 2014)

Sustained exponential improvement in computing; Extraordinarily large amounts of digitalized information; Recombinant innovation. These three forces enhance mental power are yielding breakthroughs that convert science fiction into everyday reality.

(2) Digital technologies transform economy (Varian 2016)

Data collection and analysis; Personalization and customization; Experimentation and continuous development; Innovations in contracting; Coordination and communication

(3) Examples of digital technologies

A small group of small drones can build a rope bridge without human control; Robots can climb ladders and walk over uneven terrain; Artificial intelligence is transforming and improving many services such as health care, education and financial services; Machine language translation.

(4) Benefits brought about by digital technologies

Digital technologies have been transforming the ways we produce, consume and distribute goods and services. More efficient and new production methods, new markets and new business models proliferate. Consumers are able to consume greater volume, variety and quality of goods and services. Consumers can enjoy more choices and leisure time

(5) GDP statistics tell different story

Real GDP growth rates have been slow across OECD countries; Real wage rates have been growing very slowly or decline. For example, U.S. labour productivity (output per worker) growth rates: 2.73% in 1947-1973; 1.54% in 1974-1994; 2.85% in 1995-2004; 1.27% in 2005-2015.

(6) Major explanations on offer

GDP is underestimated due to measurement challenges; Low investment both in physical and human capital; Wide disparities in productivity between leaders and laggards; Digital technologies are not so powerful as previous technological advances.

(7) Some credible verdicts

Official data understate the changes of real output (Feldstein 2017). According to Groshen & Moyer (2017), downward biases in the BEA official GDP are in the order of 0.39% – 0.43% per year between 2000 and 2015. Schreyer (2017) estimates downward biases for OECD GDP averaged at 0.45% since 2004. Current measurement procedures may understate the true growth of real GDP and overstate prices (Hulten 2017).

2. Methodology——improve/expanding GDP measures

(1) Taking satellite account approaches

Satellite accounts derive their name from the fact that they would orbit around the core national accounts. Satellite accounts are pursued for different purposes and hence take different forms:

Experimental in nature; Expand scope of national accounts; Provide more incremental detail of an industry or economy; Modify existing structure. Examples include digital economy, health care output, education output, R&D satellite accounts.

(2) Searching for better measurement techniques

- Measuring quality change for improved goods

Different hedonic models; Alternative data sources; Applied to more products.

- Dealing with new goods

Refining Hick's reservation price methodology; Applied to different products; Using alternative data sources.

(3) Making use of big data

- Enormous amounts of administrative data and transaction data are potential available for compiling GDP statistics

Credit card transaction data; Scanner data; Motor vehicle registration data

- Improving the accuracy of GDP statistics

Empirical evidence on better estimation methods; Checking consistencies across related measures with multi data sources

(4) Measuring intangible capital

Further research on scope and classification of intangibles; Developing output-based measures of intangibles; How to measures capital services for intangibles.

(5) Developing better output measures for services industries

- Key services industries for further research on output measurement

Health care; Education; Finance

- Major issues remaining

Health care – how to account for quality change

Education – what is the output of education

Finance – how to treat risk premium

(6) Measuring globalization

Using firm level data to study import/export components of value added across industries. Research on the impact of FDI on TFP measurement and national income account. Research on how MNEs produce knowledge-based assets globally.

3. Results

3.1 Old problems – quality change and services output measurement

'Chronical diseases' of GDP become more incurable: Quality change and new products; Services industry output measurement.

(1) Measuring quality changes

The quality change problem arises when a more desirable new model of a good does not cost much more than the old. Constructing quality adjusted price indexes has been always difficult. The challenge is much bigger today with profusion of new and improved goods in the growing digital economy. Research on measuring quality change will be on agendas of economic statistics conferences for years to come.

(2) New goods problem

The new goods problem is even more challenging, as there are no prior versions of the good on which to base price comparisons. Current procedures for incorporating new goods into existing price indexes are complicated, but may miss much of the value of these innovations. By implication, the benefits of important new information technology goods, like the Internet and the many applications it enables, may be subject to significant undervaluation.

(3) Hard- and easy-to-measure industries

Easy to measure industries: Agriculture, mining, manufacturing, Transportation, utilities; But digital services are hard to measure now. Hard to measure industries: Most are knowledge intensive services industries, professional and business services; The real challenges is can we measure the HTMI?

(4) Increasing importance of services industries

Knowledge-based services industries have been transformed by digital technologies: Health care, Finance, Education. The output measures of health care and education are still input based in official GDP statistics. The accuracy of measured economic growth critically depends on how services outputs are measured.

(5) Output concept

SNA08 (6.89): output is defined as the goods and services produced by an establishment. Output is an intuitively simple concept in a textbook production function, but there is a great diversity of products in real world, tangible and intangible, differences in quality and variety. Aggregation is necessary so that real outputs are measured as to synthetic constructs, no

longer refers to specific products. Exact units of measurement are somewhat fuzzy.

(6) Output measurement for services

Data are available on the value or cost of product transacted for service producing industries, but hard to measure real output. The key issue is how to decompose value into meaningful quantity (measurement unit) and price components.

(7) The case of health care services

The patient purchases medical treatment for improving his or her health status. Disease treatment is the output of health care services. The fundamental problem is to separate the value of medical expenses into price and quantity components in order to measure real output of health care. In what units do you measure output- not necessarily by the visit or the procedure. This is doctors' "output". It is challenging to determine the level of output and tell if improvements in technology have increased outcome-based output over time?

3.2 New problems – intangibles and impact of globalization

New diseases have developed: Increasing role of intangible capital; Globalization blurs national boundaries.

(1) Intangible capital

In the digital economy, investment in intangibles has become the dominant source of business capital formation, far outstripping the rate of investment in tangible plants and equipment. Measuring intangible capital presents a host of problems, since much of it is produced with firms on "own account" without a market transaction to fix prices and quantities. R&D, a key component of intangibles, is capitalised in national accounts in OECD countries by input approach.

(2) Economic impact of globalisation

Globalisation is the process of integrating national economic structure and production with the rest of the world ad blurs national boundaries. Firms organise their production and marketing at a global level, often spanning several countries. Household and firm spending become more international. Capital, in particular intellectual property, can be used simultaneously across the world in a multinational enterprise. Labour is mobile and income returned to home country can be an important part of national disposable income.

(3) Transactions brought about by globalization

Trade in services——Call centres, Software programming, Legal and accounting services, Medical services. Globalisation of production—— Smartphone: design & software in U.S., assembled in China. Car manufacturing: bumper, seatbelts, climate controls, etc, in different countries. Clothes: cost is 20% of retail price, made in emerging economies.

(4) Defining trade in services

Differences in defining trade in services can result in large export/import statistical discrepancies; Services are production activities that change the condition of a good or a person or that facilitate the exchange of products or financial assets; Trade in services occurs when its consumer or producer is a non-resident; Four modes of trade in services: cross-border supply; Consumption abroad; Commercial presence of foreign affiliates of MNEs; Presence of natural persons.

(5) Measurement issues in measuring trade in services

Custom administrative data is the key data sources for compiling export/import statistics; Services are delivered online, in particular business services. Such transactions are harder to be captured; Development of improved export and import price indexes is proceeding more slowly or no progress at all.

(6) Measuring Multinational Enterprises (MNES)

Measuring activities of MNEs is the most serious challenge to GDP. MNEs organise production on an international basis; Each country can only observe parts of MNE operations; Statistics based on isolated parts can be misleading Impact of internal transfer pricing. MNEs can lower their global tax burden by a number of structural arrangements: affiliates overseas to act as income recipients, holders of intellectual property rights; Transfer pricing can cause value added to be misallocated if not true reflections of the market prices.

(7) How smartphones are treated in national accounts

Apple sends design and software to China. This transaction currently is treated as an internal transfer within firm, with no separate transaction for exports of design/software. The wholesale value is counted as import and in final sales, resulting understated GDP. One option is to capture the export of intellectual property to the foreign producer on its surveys of international trade in services.

(8) Measuring international trade in intellectual property products

IPP categories: R&D, computer software & databases, artistic originals, and other IPPs.

Alternative forms of IPP: Original IPP, licences to reproduce IPP, copies of the original IPP.

International trade in IPP: No cross-border physical move; Recorded in royalties and fees rather than in trade statistics.

4. Discussion and Conclusion

Measurement challenges posed by digital economy have become more severe including Quality change, Increasing share of services, Intangibles, and Globalization.

Research opportunities in the future are:
(1) Service industry specific measurement issues
(2) Measurement issues for intangibles
(3) Measuring MNEs

References

1. Ahmad, N. and P. Schreyer (2016), "Measuring GDP in a Digitalised Economy", OECD Statistics Working Papers, 2016/07, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jlwqd81d09r-en
2. Diewert, E., et al. (2017). The Digital Economy, New Products and Consumer Welfare, Vancouver School of Economics.
3. OECD. Measuring the Digital Economy: A New Perspective[J/OL]. http://dx.doi.org/10.1787/9789264221796-en,2014.
4. BEA.Defining and Measuring the Digital Economy[J/OL]. https://www.bea.gov/digital Economy, 2018-3-15.
5. Unite Nations, System of National Accounts 2008 [M].New York,2010

Measuring the structure of digital economy - The case of China

Shujian Xiang*[1]; Yingmei Xu[2]; Wenjun Wu[2]
[1]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China.
[2]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China.

Abstract

Since the G20 officially proposed the digital economy development and cooperation initiative in 2016, the G20 summit has included the measurement of the digital economy as an important agenda to discuss in recent years. At the same time, the Intersecretariat Working Group on National Accounts (ISWGNA), supported by the European Commission, IMF, OECD, UN, World Bank and established by the United Nations Statistical Commission (UNSC), also listed the digital economy as one of the most momentous research issues at the latest 12th National Accounts Advisory Group (AEG) meeting in 2018, where it stressed the challenges confronted on measurement, entailing the satellite framework, the free products and the role of data in statistical measurement etc.. This study aims to bridge the gap between the studies of OECD and China on digital economy and tries to fill the blank between theory and practice on the measurement. In general, this paper mainly focuses on two questions: how Chinese digital economy satellite account can be designed, and how it integrates with the satellite framework proposed by OECD digital economy working group. Specifically, this paper first introduces the overall design ideas and accounting concepts, and builds up the identification mechanism of key activities. Then, the comparison of industry and products categories in digital economy was made between OECD and China. On the basis, this paper explores the feasibility of supply table, usage table and generalized investment matrix in the framework of China's digital economy satellite account. Furthermore, it provides a preliminary research paradigm for input-output structure of the digital economy, its technical relevance to the traditional economy, and the value added of China's digital economy. Finally, combined with the status quo of China's digital economy research, it suggests a systematic satellite framework for the accounting of China's digital economy, the research and development of high-frequency big data acquisition methods, and the improvement of innovation-oriented industry and product classification and other recommendations.

Keywords

Satellite account; Statistical classification; Statistical measurement; OECD

1. Introduction

At present, the global economy is gradually entering the era of digital economy. Due to the different degrees of integration between digital technology and traditional economy in the world, countries have not yet reached a consensus on the definition of digital economy. Countries at different stages of development have different definitions. On the basis of the "G20 Digital Economy Development and Cooperation Initiative", the China Academy of Information and Communications Technology (CAICT)(2018) defines the digital economy as "a digital production of knowledge and information as the key production factor, with digital technology innovation as the core driving force, with modern information networks as the important carrier, through the deep integration of digital technology and the real economy, continuously improves the digitalization and intelligence level of traditional industries, and accelerates the reconstruction of the new economic form of economic development and government governance model. The Organization for Economic Co-operation and Development (OECD) defines the connotation and extension of the digital economy from the perspective of inclusive development and digital transaction accounting. It points out that the digital economy is standardized, dynamic, and data-driven. The essential characteristics of structural dispersion, and the digital economy can be identified by whether the transaction activity is digitally ordered, platform-enabled or digitally delivered.

Regarding the overall development trend of the digital economy and relevant statistical problems, the results of CAICT studies (2018) show that the digital economy of the G20 countries is developing at a high speed. The digital economy of China in 2017 reached 4.02 trillion US dollars, second to the United States. However, the deepening of digital technology in the world is facing the Solow productivity paradox: although the digital economy in the world is developing rapidly, the productivity growth rate of all countries is generally facing a downward trend. For example, the secretariat of Asia-Pacific Economic Cooperation (APEC) (2018) estimates that from 2000 to 2017, both the industrialized and developing economies of the APEC region have a tendency to decline in labor productivity growth.

In response to this problem, Shujian Xiang and Wenjun Wu (2018) reviews that the OECD has repeatedly pointed out in recent studies that the current slowdown in productivity growth may be due to the limitations of existing statistical methods in accounting for the digital economy and conducted a series of in-depth theoretical discussions and empirical analysis on the official statistics of OECD countries. The results of existing studies indicate that the current national economic accounting framework generally lacks sufficient explanations for the digital economic phenomenon. It has limited information on the main body, the transactions and products of the digital economy. This will cause statistical mismeasurement, causing problems in the analysis of

macroeconomic data and productivity paradox. Therefore, the OECD (2018) proposes a general digital economy supply and use framework to meet the needs of macroeconomic analysis and related policy formulation.

At present, China is facing the circumstance of slowing economic growth and productivity paradox. The measurement of the structure of its economy, especially the digital economy is paid highly attention around the world. One of China's mainstream research on the structure of digital economy by CAICT carries out its statistical measurement from the two dimensions involving digital industrialization and industrial digitalization. As for the methodology, the scale of digitization is mainly based on the estimation of contribution rate of digital technology to traditional industries. However, the technical contribution rate cannot be equated with the scale of the digital economy. At the same time, it is difficult to describe the impact of different types of digital transactions on various traditional industries.

With reference to relevant research, this paper explores the feasibility of constructing a more comprehensive and in-depth input-output framework to depict the value chain between producers and consumers, as well as the supply and intermediate use of various products in the digital economy. On the basis of previous OECD research and System of National Accounts 2008 (SNA 2008), this paper develops a deeper understanding and more general characteristics of the digital economy and its satellite account framework and lays the foundation of this framework by conducting a comparison of digital industry and product classifications between OECD and China. The significance of developing the framework is to try to construct a comprehensive and feasible data accounting system and to systematically analyze the contribution and influence of the digital economy at different economic development stages. The rest of the paper is organized as follows: section2 demonstrates the theoretical foundation of related methodologies in our research; section3 presents the result and analysis; section4 draws conclusions and gives several suggestions.

2. Methodology

2.1 The basics on satellite account

The Satellite Account is an auxiliary accounting system that modifies specific accounting elements on the basis of SNA Central system with good flexibility and expandability. Broadly speaking, satellite accounts can be divided into two categories: one is called an internal satellite account, which selects and rearranges a group of key sectors from the industry and product sector classification of the central framework based on the research topic; the other category is called external satellite accounts, which is mainly designed for the extension in production, consumption and capital formation etc.. The difference between the two is that the former focuses on selective local refinement while the latter focuses on overall expansion.

China's digital economy satellite account framework combines the characteristics of the above two. It selects key departments related to the digital economy and extends the production boundary of the central framework accordingly. This will effectively link China's digital economy satellite accounts with the traditional national economic classification, and can highlight new industries and products in digital economy. In order to explain the methods and techniques of digital economy satellite account more clearly, we compare the design of digital economic satellite accounts with conventional internal satellite accounts and external satellite accounts respectively. Its main connection with internal satellite account is that both of them require to identify the key activities based on the current classification. Their difference is that digital economy satellite account also requires the identification of non-digital products or industries other than digital ones. This is because the deep integration of digital technology and the real economy enables non-digital products to be traded through digital technology to have the characteristics of digital economy. For example, although there are non-digital products of e-retailers on large e-commerce platforms such as Taobao and Jingdong in China, their use of the platform to deliver the transaction order and payment makes it an important part of the digital economy.

2.2 The accounting mechanism of digital satellite account

Based on the above analysis, the mechanism for identifying key activities in China's digital economy satellite accounts was constructed. Essentially, digital economy is the digitization of the real economy, that is, using digital technology and infrastructure to store information on production relations, economic transactions, product supply and demand etc. in the real economy in the form of binary bits. Therefore, the identification of key activities in the digital economy encompasses identification of digital economic transactions which includes three categories: the digitalization of voucher, the digitization of supply and demand information, and the digitization of content information.

First, the digitization of the voucher makes the transaction relationship between economic units have a record. At this time, the specific digital platform records the order, receipt, invoice, etc. in a digital form which broadens the credit mechanism of the transaction and makes people dare to trade with more and more strangers, thus expanding the scale and scope of related transactions; secondly, the digitization of supply and demand information reduces the friction in transactions when the digital platform collects information from both parties and obtains certain data access rights, or it may proactively provide supply and demand matching and matching services, thereby reducing the information search cost of both the supply and demand sides. Finally, the digitalization of the content information makes the real world knowledge and other information digital and its transactions are

often in non-monetary form. The OECD studies (2018) point out that digital transactions are a key factor in identifying the digital economy. They can be defined by three dimensions: digitally ordered, platform enabled and digitally delivered.

Using the methodologies mentioned above, the results below will firstly present the main problems and solutions in the digital economy, as shown in Table 1. After that, specific details on the comparison of classifications will be illustrated to bridge the gap between general conceptual framework of satellite account and statistical practices.

3. Results

3.1 Mismeasurement problems and solutions in digital satellite account

As Table 1 shows, there are mainly 10 problems faced in the research of digital economy. This paper classified the common solutions and proposes suggestions for China's Digital Economy Satellite Account in order to better measure the structure of China's Digital Economy.

Table 1: Main accounting issues and corresponding design ideas around the framework

Main problems facing digital	**Major solutions**	**Ideas of China's Digital Economy**
The definition of concept and accounting scope of digital economy	-IMF: Based on the digital department -OECD: Based on digital trading	Based on generalized digital transactions (including exchange, transfer and internal transactions)
The distinction between digital economy and other industries	OECD: Category 6 Core Digital Industry	5 core digital industries
The contribution and influence of the digital economy to the traditional economy	Splitting from existing industries makes it difficult to characterize the parts that are not coincident and deeply integrated, and it is difficult to grasp the linkage of industry.	Build a satelliteaccount containing existing industries and coordinate the input-output structure between industries and digital products

Accounting for free digital services outside the SNA production boundary	-Separate measurement of relevant economic benefits -Separate accounting for SNA production	Expand SNA production boundaries and include free services in digital products
Differentiating between different modes of platform enterprises in the digital economy	- Tradition: B2B, B2C, P2P -OECD: Digital Mediation Platform -BEA: 8 mainstream platforms	Based on the "three new" statistical Internet platform (including production, living services, technological innovation, life services and other platforms)
Differentiate between different types of digital services in the same enterprise	Further division of industrial activity units on the basis of institutional units	Further segmentation of homogenous products based on the division of industrial activity units
Identification and classification of digital products in the digital economy	-OECD:ISIC,CPA,CPC -BEA:NAICS	'Three New' statistical classification, strategic emerging industry key product classification, digital trading
Data accounting in data driven business model	- Measuring the value of a single case -cost method, market law, income law	Build a generalized investment matrix that includes data investment and capital stock for each industry
Estimation of the added value of the digital economy and the traditional industry integration (and the total output and intermediate input from the production perspective)	- Basic value-added scale plus the contribution rate based on the growth accounting model -BEA: Assumptions based on the same ratio of digital to non-digital industry inputs	The total output of the digital product obtained from the transaction, based on the input ratio assumption, or based on the productivity survey to calculate the reversed intermediate input, subtracting the added value
Accounting for digital ecological protection	There are no relevant research ideas yet.	Refer to SEEA's physical volume accounting to consider the resources and data flows of the digital environment.

3.2 Comparison of digital industry and product classifications

3.2.1 Industry classification of digital economy satellite account framework

(1) Digitally enabled industrial sector

The OECD defines the digital sector as the ICT industry in the fourth edition of the International Standard Industrial Classification (ISIC), which includes the ICT manufacturing industry, the ICT trading industry and the ICT service industry. This paper compares the ICT industry in international standards with China's National Economic Industry Classification (GB/T 4754-2017) and the 'Three New' Industry Statistics Classification (2018), and finds that in the ICT industry group of ISIC, In addition to the "2680 magnetic and optical media manufacturing", China's industry can correspond. At the same time, compared to ISIC, China's National Economic Industry Classification for "2640 Electronic Consumer Device Manufacturing", "4652 Electronics, Telecom Equipment and Parts Wholesale", "61 Telecom" and "631 Data Processing, Hosting and Related Activities; Portals The classification of the four categories of ICT industries is more detailed.

(2) Digital intermediary platform

The OECD's research defines a digital intermediary platform as a unit within the SNA production boundary that facilitates interactions between multiple individuals or business users and collects intermediary fees. The "Internet platform industry" (industry code 0502) in the "Three New" Industry Statistics Classification (2018) can meet the construction needs. There are five categories of "Internet platform industry", including: "Internet production platform" (05021), "Internet life service platform" (05022), "Internet technology innovation platform" (05023), "Internet public service platform" (05024) and "Other Internet Platforms" (05025).

(3) Corporate entities and unincorporated households that rely on the intermediary platform

In China's digital economy, the digital platform has reduced the barriers for SMEs to enter the market, making digital technology no longer limited to the way that traditional large enterprises generate income. Therefore, such industrial sectors should be listed as a type of digital industry in China's digital economy satellite accounts. At the same time, it is necessary to separately list the relevant household industry sectors. However, at present, domestic and foreign methods have not been developed to accurately identify such enterprises. This is because the proportion of enterprises that rely on the platform to create output can only rely on strong technical and economic assumptions, and this will bring many miscalculations in statistical accounting.

(4) E-sellers

China Digital Economy Satellite Account Framework can refer to OECD research as an electronic retailer (E-Talers) and an electronic distributor (E-Vendors). The difference between the two is that e-retailers sell traditional retail products online, while electronic distributors produce goods and services and transport/transport them electronically, such as subscription-based digital content.

(5) Other digital companies

In addition to the above-mentioned industries, China's digital economy also includes other types of digital industries, such as a large number of so-called "free" services. In other words, although the producers of these "free" services do not receive explicit monetary income from their users, the services they provide are essentially another way of barter transaction. Therefore, digital enterprises that do not use dominant currency transactions should also be included.

3.2.2 Product classification of digital economy satellite account framework

(1) Digital goods

For the identification of digital goods, the OECD proposes to use the ICT cargo classification in the UN standard, Central Product Classification (CPC 2.1). The 52 ICT goods in CPC can be divided into four categories: computer and peripheral equipment, communication equipment, electronic consumer equipment, other ICT components and goods.

At present, the National Bureau of Statistics of China released the "40 communication equipment, computers and other electronic equipment" products in the Catalogue of Statistical Products in 2010, which can basically cover the above four types of products, and the input and output survey in China in 2017. The corresponding intermediate consumption products can also be found in the Catalogue of Material Consumption.

(2) Paid digital services

For the identification of digital services, the OECD proposes to use the ICT service representation in CPC 2.1 and separately calculate cloud computing services and digital intermediary service products. The 46 ICT services in CPC 2.1 can be divided into six categories, including: ICT equipment manufacturing services, commercial and productivity software and its licensing services, information technology consulting and services, telecommunications services, leasing or leasing of ICT equipment, and other ICTs. service.Comparing China's product classification with UN international standards, it is found that except for China's "Catalogue Catalogue for Statistical Products", "38 Charges for Production Services and Repairs" can correspond to "ICT Equipment

Manufacturing Services" in CPC 2.1, and the other six types of ICT services have no corresponding Product Categories.

(3) Free digital services

The free digital services in the framework of China's digital economy satellite account are mainly provided by “other digital enterprises” in the digital industry, and the basic principles are not repeated here.

(4) Non-digital products

On the basis of digital products, the China Digital Economy Satellite Account can refer to the recommendations of the OECD study to select typical digital products to focus on their integration with digital technologies, including: accommodation services; food and beverage service activities; land transportation services; Travel agencies, travel accommodation services and other related activities; advertising and market research services; education services; film video and television programming services; financial and insurance services; gambling and gaming activities; retail trade.

4. Discussion and Conclusion

(1) Forming a systematic accounting system for China's digital economy satellite accounts

This paper points out the limitations of China's current digital economy accounting such as unclear concepts, narrow calibre and lack of comprehensive systems. Based on this, it proposes the design of China's digital economy satellite account framework, and proposes to continue to track international systems for accounting in the future. For the framework design of satellite accounts, the OECD's proposed digital economy satellite account framework has been widely recognized by the United Nations National Accounts Working Group (ISWGNA) Expert Advisory Group (AEG) and is expected to be based on this in the near future. In the OECD countries, the US Bureau of Economic Analysis (BEA) has clearly proposed the ultimate goal of building a digital economy satellite account, and has carried out different stages of empirical measurement and theoretical research. Among Asian countries, Malaysia is already constructing a digital economy satellite account.

These international frontier researches have not only laid an important foundation for the further improvement of the national economic accounting system, but also provided powerful guidance for the systematic accounting of China's digital economy and provided effective assistance for the development of China's digital economic accounting. In the future research on digital economic accounting methods, China should learn from others, broadly absorb the experience of digital economic practice in various countries, and closely combine the needs of China's economic development, give priority to

the study of digital industries in strategic emerging industries, and form a systematic portrayal. Data monitoring system for digital economic structure and a more scientific and accurate accounting method system.

(2) Improving methods to obtain high-frequency big data in China's digital economy

This paper analyzes the mechanism that the digital economy influences the traditional economy through the digitization of transactions, including the increase of transaction trust brought by digitalization of documents, the reduction of transaction costs brought by the digitization of supply and demand information, and the non-monetary data transactions brought by the digitization of content information. To comprehensively and accurately count high-frequency transaction data containing such information, it is necessary to research and develop stable, safe, reliable, and high-quality high-frequency data acquisition methods. Among the data sources of current accounting practice, micro-investigations have limitations such as low timeliness, cumbersome investigation costs, and high investigation costs, and commercial data is often difficult to obtain authority for the protection of trade secrets. Therefore, we can consider building a distributed statistical accounting system with anonymity and tamper-proof based on blockchain technology, and statistically record large data transactions through decentralized distributed ledgers, thus achieving high-frequency, efficient and accurate statistical accounting for the digital economy.

(3) Complementing China's digital economy industry product classification

This paper has carried out a more in-depth discussion on the industry and product classification of the digital economy, and combined with the "three new" statistical classification (2018) and the "strategic emerging industry classification (2018)" in statistical practice, etc. The advantages and limitations of statistical classification. In the future digital economic accounting work, it is recommended to further improve the statistical classification of China's digital economy satellite accounts under the support and guidance of the United Nations Statistical Commission (UNSC) Joint Accounts of the National Accounts Working Group (ISWGNA) to further increase the world. The international comparability of digital economic development in various countries, combined with the international initiatives promised by China at the G20 summit, and the inclusive development goals proposed by the G20 Digital Ministerial Conference, more accurately measure the impact of digital growth on individuals and businesses. At the same time, it should be combined with the storage and construction design of non-traditional big data such as Internet statistics to form a digital economy satellite account with profound

theoretical and practical operation, which makes the statistical system of China's digital economy more comprehensive, flexible, timely and effective.

References

1. China Academy of Information and Communications Technology (2018). G20 National Digital Economy Development Research Report.
2. APEC Policy Support Unit (2018). APEC Regional Trends Analysis: The Digital Productivity Paradox.
3. Shujian Xiang, Wenjun Wu (2018). Recent Developments in OECD Digital Economy Research and Its Enlightenment. Statistical Research.
4. Nadim Ahmad, Peter van de Ven (2018). Developing digital economy satellite accounts for macro-economic statistics. The OECD Statistics Newsletter.
5. Ahmad, N. and J.Ribarsky (2018). Towards a Framework for Measuring the Digital Economy. Paper prepared for the 35th IARIW General Conference.
6. John Mitchell, A Proposed framework for Digital Supply-Use Tables (2018). Paper prepared for the OECD Working Party on National Accounts.

The impacts of digitalisation on China's economic system

Guangwu Chen
The University of New South Wales, Sydney, Australia.

Abstract

This paper aims to quantify the impacts of digitalisation on China's economic system within the global boundary from production, consumption, and income perspectives. The key work we are going to carry out as follows: (1) We present a new industrial classification system by disaggregating the digitalised sector from related sectors and then filling the data of the new sector and also the numbers with each sector. The data are from various sources including yearbooks, digitalisation development reports and related databases. (2) We insert the new digital sector into China's economic system using a global boundary input-output table focusing on China system. We restructure the global multi-regional input-output table into two regions – China and rest of the world (RoW) and extended the new MRIO table with a new digital sector to show the digital production supply chain. (3) We use Leontief demand-driven model and Ghosh supply-driven model to quantifying production of the new goods and services, consumption from consumers and the labour and capital income. This integrated study plans to give a whole picture for assessing the size and growth of the digital economy and its possible contributions to households welfare.

Keywords

Digitalisation; economic system; multi-regional input-output table

1. Introduction

United Nations has highlighted the sustainable consumption coupled with challenges of environmental and social changes in Sustainable Development Goals (SDGs). The ever-growing digital economy poses new challenges for China to achieve a sustainable future since many consumptions through the internet-based companies has not yet to be found sustainable. Chinese government has regarded the digital economy as the future in terms of improvement to existing industry and growth of new ones. However, in order to catch up with the fast pace of development in digital markets and related companies, the regulations safeguarding the environmental and social outcomes are urgently required to update with evidence-based supports.

The quantification of footprint linking to direct, indirect and induced effects for internet-based companies has not been comprehensively investigated in previous research (Cheng *et al.* 2019). The direct effects

generated by the initial spending from consumers create additional activities and therefore environmental impacts in the local economy. In addition, the business-to-business transactions indirectly caused by the direct effects lead to further 'indirect' environmental impact embodied in supply chains. The income generated through digital platforms by households could be re-spent and induce further economic and environmental impacts, which has been largely ignored in the literature.

The digital economy also gives a boost to a free market platform such as "Airtasker" in which temporary positions are common and individuals or organizations contract with independent workers for short-term engagement (i.e. Gig Economy (Friedman 2014)). Many Chinese outsource the labor work to part-time workers or freelancers who especially for those new migration race the price to the bottom to win the job. Debate on the digital market of job outsourcing splits into two perspectives.

One argues that it increases the employment chances to get paid and supplement the low-income group people. The others claim that it leads to more individual joining the freelancing and increase the instability of the society since this form of employment has mislabeled workers as independent contractors and immunize employers from workplace injuries and harassment. Freelancing also raises the concern of responsibilities of government such as pension since the Chinese government has collected the tax from these part-time workers. However, previous research lacks of a modeling to qualify the employment of gig workers by converting it into a full-time equivalent unit (an equivalent freelancer), which will be conducted during this fellowship.

Digital platforms have made online shopping prevalent and influence subjective experiences as to make them potentially addictive activities (Chen *et al.* 2019b). Several predictive factors have been identified as drivers of compulsive online shopping including female gender, low self-esteem, low self-regulation, negative emotional state, as well as cognitive overload. Linking micro-scale consumer's socio-economic and demographical characteristics particular with mental health state to the macro-scale environmental and social footprints would help to regulate the consumer's behaviors towards a sustainable consumption future.

To challenge digital economy measurement, (1) We present a new industrial classification system by disaggregating the digitalised sector from related sectors and then filling the data of the new sector and also the numbers with each sector. The data are from various sources including yearbooks, digitalisation development reports and related databases. (2) We insert the new digital sector into China's economic system using a global boundary input-output table focusing on China system. We restructure the global multi-regional input-output table into two regions – China and rest of the world (RoW) and extended the new MRIO table with a new digital sector to show the digital production supply chain. (3) We use Leontief demand-

driven model and Ghosh supply-driven model to quantifying production of the new goods and services, consumption from consumers and the labour and capital income. This integrated study plans to give a whole picture for assessing the size and growth of the digital economy and its possible contributions to household's welfare.

2. Methodology: *Leontief and Ghosh models*

Our model start with Leontief's famous demand-side (Leontief, 1936; Leontief, 1949) and Leontief and Strout (1963). Assume that an economy can be categorized into n sectors. Let $\mathbf{T} = (t_{ij})_{n\times n}$ be a $n \times n$ intermediate transaction matrix with t_{ij} representing the input from the ith sector to the jth sector in the economy, $\mathbf{x} = (x_{ij})_{n\times 1}$ be an $n \times 1$ vector of the total output with x_i being the ith sectoral total output, $\mathbf{A} = (A_{ij})_{n\times n}$ be an $n \times n$ direct requirement coefficient matrix with a_{ij} showing the direct input from the ith sector to the jth sector to produce one unit of output, $\mathbf{L} = (\mathbf{I} - \mathbf{A})^{-1}$ be the famous Leontief Inverse Matrix representing both direct and indirect input in order to produce on unit of output; and $\mathbf{y} = (y_i)_{n\times m}$ be an $n \times m$ flow matrix including m categories of final demand and with y_i being the ith sectoral final demand. The standard Leontief's demand-driven input-output model can be shown as: $\mathbf{x} = \mathbf{Ly} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{y} = (\mathbf{I} - \mathbf{T}\hat{\mathbf{x}}^{-1})^{-1}\mathbf{y}$.

The supply-side input-output model was developed by Ghosh (Ghosh, 1958) as a supplement to the Leontief demand-driven input-output model. Ghosh's input-output model is an allocation model with the column balance equation $\mathbf{x'=i'T+v}$, **v** is the row vector of value added, **i** is a suitable unitary vector.

The direct output coefficients **B** is calculated by dividing each row of **T** by the gross output of the sector associated with that row. Its matrix $\mathbf{B} = \hat{\mathbf{x}}^{-1}\mathbf{T}$, namely allocation coefficients, represents that the distribution of outputs of the original sectors. The outputs across all sectors of the economy show inter-industrial sectors buying input from the original sectors.

Using the column balance equation, we have $\mathbf{x}' = \mathbf{i}'\mathbf{T} + \mathbf{v} = \mathbf{i}'\hat{\boldsymbol{x}}\mathbf{B} + \mathbf{v} = \mathbf{x}'\mathbf{B} + \mathbf{v} = \mathbf{v}(\boldsymbol{I} - \boldsymbol{B})^{-1} = \mathbf{vG}$. The matrix **G** is called the Ghosh inverse, relating sectoral gross production to the primary inputs — a unit of value entering the inter-industry (supply chain) system at the beginning of the process. It is termed a supply inverse and measures the production values of sectors that come in the supply chain system caused by per unit of primary input in sectors (Miller and Blair, 2009).

Attaching the satellite accounts to the supply chain system, the extended Leontief and Ghosh models can be shown respectively as $\mathbf{Q} = \mathbf{qLy}$ and $\mathbf{Q=vGq}$, where **Q** represents total digitalisation and q is the sectoral intensity vector that is calculated by $\mathbf{q} = \mathbf{Q}\hat{\mathbf{x}}^{-1}$, indicating digitalised impacts caused by producing per unit of sectoral output. Accordingly, the vector **Ly** in the

demand-side system represents the the sectoral digitalised effects on total output throughout all the supply chain system associating with one unit of final demand in sectors.

3. Results

The MR-SUT structure for the Chinese area is shown in Figure 1. It features a 42–industry and – commodity classification aggregated from the sectoral root classification for the MRIO table in the case study. The categories of the final demand blocks include rural household consumption expenditure, urban household consumption expenditure, government consumption expenditure, gross fixed capital formation, and inventory changes. The value added blocks for each region are the same as in the input-output tables published by the NBSC, i.e., compensation of employees, net taxes on production, depreciation of fixed assets, and operation surplus. The one-column rest of the world (RoW) in the right section of the MRIO table is the total exports of each region to the ROW and the RoW matrices at the bottom of the MRIO table are total imports distinguishing between intermediate input and final demand from the rest of the world to each region. The satellite account matching the MR-SUT is the industrial digitalisation input. The MR-SUT of B-T-H distinguishes 11 valuations (margins, taxes, subsidies, etc.) as mentioned above. Thus, the entire base MR-SUT for one year has the following dimensions:

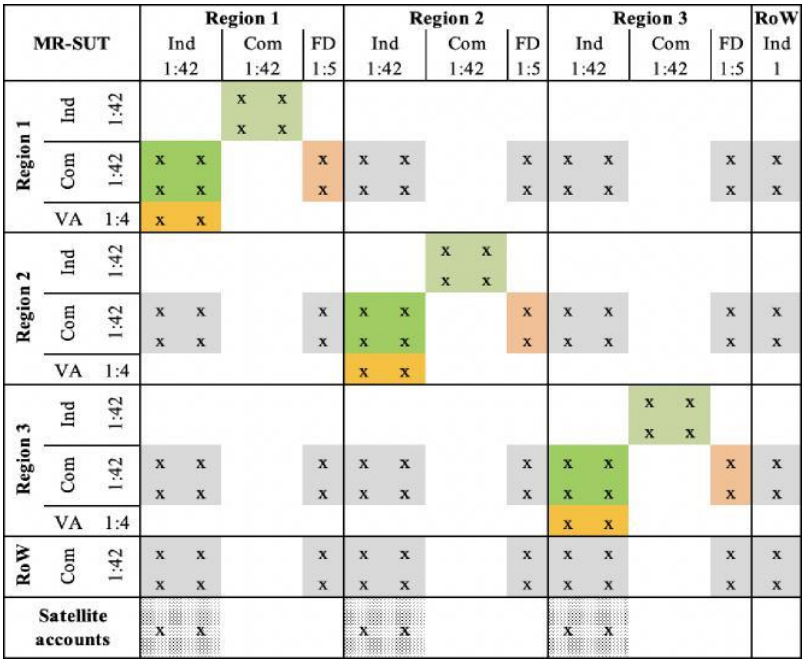


Figure 1. Structure of Chinese MRIO table used in the study. Three regions shown here is an exemplified format for illustration (Ind =industries, Com =commodities, FD =final demand, VA = value added).'

Peer-to-peer accommodation differs from the traditional accommodation sector by splitting the booking and accommodation services. The accommodation services are outsourced to hosts while the booking services are still maintained in the peer-to-peer accommodation platforms. In contrast, the traditional accommodation sector usually integrates the booking and accommodation services together. Some traditional accommodation providers may also outsource their booking services to internet-based platforms, such as Booking.com and Agoda.com, but these have not been included in this research.

In our case, CF of booking service represents 3.08 kg CO_{2e} per room per night, while CF associated with the traditional accommodation services ranges from 4.19 to 6.31 kg CO_{2e} per room per night. These figures are based on the calculation of three different scenarios as shown in the Table 6. The scenario modeling was based on an Excel-based cost accounting tool "Unlocked-Airbnb_Financial_Calculator_v2.4.6" along with a video tutorial provided by Airbnb. The fees for the Airbnb platform is about 12.3 Yuan per room per night on average. Due to the relative low carbon intensity for running the Airbnb platform (average 0.05 kgCO_{2e}/RMB), the total CF for booking services is about 3.08 kg, which seems very high compared to the CF related to the actual spending for serving guests.

In scenario 1, only the consumables were taken into account. The CF of accommodation services of Airbnb would be 4.19 kg CO_{2e} per room per night. Most of the emissions come from the electricity and other energy consumption. In scenario 2, the insurance and services e.g. gardening services and rubbish removals were further considered (scenario 2). The CF of accommodation services of Airbnb would be 4.59 kg CO_{2e} per room per night. In scenario 3, durables were also included, thus increasing the CF to 6.31 kg CO_{2e} per room per night. However, running an Airbnb is much lower than an average household's CF (27.11 kg CO_{2e} per room per night, scenario 4), which is because the other sectors including Food, Clothing and footwear, Transport, Education, Holiday etc. make up a large amount of CF.

Running an Airbnb yields lower CF than traditional accommodation and the carbon intensity of Airbnb is also lower than the traditional accommodation sector. The booking services provided by Airbnb platform and accommodation services provided by Airbnb host together generate a CF ranging 7.27-9.39 kg CO_{2e} per room per night, which is only 12%-15% of Accommodation sector (60.90 kg CO_{2e}). The carbon intensities of scenario 1-3 of Airbnb are much lower, ranging 0.02-0.09 kg/RMB, compared to 0.38 kg/RMB of accommodation sector. That is because compared to the average amount of total fee (104 Yuan per room per night) for Airbnb housing, only 2.55-9.52 dollars per room per night are spent for running Airbnb housing in Scenarios 1-3 causing a small amount of emissions, while hotels and other

accommodation industries provide a broader range of goods and services for guests thus leading to a higher cost.

Comparison of CF based on four scenarios between Airbnb and traditional accommodation sector

CF (kg CO_{2e} per room per night)	Scenario_1	Scenario_2	Scenario_3	Scenario_
Food	-	-	-	3.39
Clothing and footwear	-	-	-	0.76
Electricity	2.68	2.68	2.68	2.68
Household equipment, water and other energy	0.15	0.41	1.74	4.12
Transport	-	-	-	4.20
Communication	-	-	0.09	0.60
Education	-	-	-	0.33
Holiday	-	-	-	2.37
Miscellaneous goods and services	0.50	0.64	0.95	3.12
Direct emissions of domestic energy	0.86	0.86	0.86	0.86
Direct emissions of private transport	-	-	-	4.68
Sum of CF (kg CO_{2e} per room per night)	4.19	4.59	6.31	27.11
Sum of Spending ($AUD per room per night)	2.55	4.07	9.52	65.19
Average price ($AUD per room per night)	104	104	104	-
Carbon Intensity (kg/$)	0.02	0.04	0.09	-
CF of booking services (kg CO_{2e} per room per night)	*3.08*	*3.08*	*3.08*	-
Traditional accommodation sector CF (kg CO_{2e} per room per night)		*60.90*		
Traditional accommodation sector Intensity (kg/$)		*0.38*		

4. Discussion and Conclusion

The Chinese IELab is also the first MRIO system to include Tibet and cover the entire economic territory boundary of 31 provinces including all 2874 regions. Rapid economic growth in recent decades has meant that close regional linkages are no longer limited to the developed coastal areas but are spreading into developing and poor central and west regions. For example, with the construction of the Qinghai-Tibet railway, Tibet's economy has strongly benefited from the increase in freight in and out of Tibet and has also seen its tourism industry boom. It is therefore of crucial importance to use subnational MRIOs to guide policy-making on environmental, social and

economic issues, especially since Chinese regions exhibit a considerable degree of heterogeneity in terms of economic development.

In addition, using IELab technology to update tables annually or generate subnational Chinese MRIOs for any year becomes a less labour-intensive task and can be based on the latest official data as well as other available data sources. The Chinese IELab currently includes the longest and latest data from 1978 to 2016 and can support central and local governments in designing and testing specific policies for any region or sector in China as well as to evaluate previously implemented policies in order to guide policy adjustments and refinements. Other topical policy applications include modelling the impacts of China's abolition of the decades-long one-child since 2015 and the Chinese government's 4-trillion-yuan stimulus package in 2008 for bolster the slowing economy during global financial crisis.

References

1. Leontief, W.W., 1936. Quantitative Input and Output Relations in the Economic Systems of the United States. The Review of Economics and Statistics 18, 105-125.
2. Ghosh, A., 1958. Input-Output Approach in an Allocation System. Economica 25, 58-64.
3. Leontief, W., 1949, "Recent Developments in the Study of Interindustrial Relationships"[J], *The American Economic Review*: 211-225.
4. Lenzen, M., K. Kanemoto, D. Moran, and A. Geschke. 2012a. Mapping the Structure of the World Economy. Environmental Science & Technology 46(15): 8374-8381.
5. Lenzen, M., D. Moran, K. Kanemoto, B. Foran, L. Lobefaro, and A. Geschke. 2012b. International Trade Drives Biodiversity Threats in Developing Nations. *Nature* 486(7401): 109-112.
6. Chen, G., Wiedmann, T., Hadjikakou, M., Cheng, M., Xu, L. & Wang, Y. (2019a). Method for assessing Airbnb's direct, indirect and induced carbon footprint. *MethodX (forthcoming).*
7. Chen, G., Zhu, Y., Wiedmann, T., Yao, L., Xu, L. & Wang, Y. (2019b). Urban-rural disparities of household energy requirements and influence factors in China: Classification Tree Models *Applied Energy (forthcoming)*.
8. Cheng, M., Chen, G., Wiedmann, T., Hadjikakou, M., Xu, L. & Wang, Y. (2019). Sharing economy and sustainability: assessing Airbnb's diect, indirect and induced carbon footprint. *Annals of Tourism Research (forthcoming)*.
9. Friedman, G.J.R.o.K.E. (2014). Workers without employers: shadow corporations and the rise of the gig economy. 2, 171-188.

Using big data in monitoring indicators of sustainable development goals in Egypt

Reem Ismail Mohamed Elsybaey*
CAPMAS Cairo – Egypt

Abstract

Big Data is a collection of data sources, technologies and methodologies that have revealed from, and to, utilize the huge growth in data creation over the past decade. It is also data on a massive scale in terms of volume, intensity, and complexity that exceed the capacity of standard analytic tools. Although there is no agreed-on definition, the term is often characterized by the 3Vs—high-volume, high-velocity, and high-variety. More Vs have been added over time, such as veracity and volatility. Data sets become so large and complex that traditional data-processing applications become insufficient to capture, store, and analyze these data. Instead, a network of human skills, advanced technologies, and data access infrastructure are essential to handle big data. This is a key challenge for policymakers seeking to combine big data in monitoring their sustainable development plans. The availability of big data provides a unique opportunity to support the achievements of the SDG's like never before. As the post-2015 development agenda has now been established, strengthening data production and the use of better data in policymaking and monitoring are becoming increasingly recognized as fundamental means for development. There is potential for big data to produce new SDGs indicators, bridge time lags, and support the forecasting of existing data sets, as well as serve as a new innovative data source in the production of official statistics. To set the foundation for long-term success, an important understanding of the opportunities and challenges that come with big data is essential. There are Challenges come with big data concerned with data quality, difficulties with access, and new required skills and technologies which represents the main challenges of big data. CAPMAS the NSO (National Statistics Office) in Egypt is member of UN Global working group (GWG) on Big Data for Official Statistics. There is a communication with national, regional and international organizations in terms of open data and big data. CAPMAS also participating in official events related to accessibility to relevant sources of big and geospatial data for SDGs indicators reporting.
One of big data projects at CAPMAS is the cooperation with Hyper Markets in compiling data prices for CPI creation. This project will Increase the degree of quality of CPI and reducing overhead for respondents and errors.

Keywords

Challenges, Data quality, SDGs indicators

1. Introduction

When we work to prepare economic, social or demographic studies to develop a strategic vision, or to diagnose our reality, our first tool is the economic, social and demographic statistics and data that form the basis for decision makers and researchers.

Statistical data represent an indispensable indicators of the characteristics of the strategy or the researcher or even the analyst. Official statistics are statistics issued by government institutions and official institutions within the state. These statistics include many demographic, economic and social fields according to statistical treatments consistent with international standards "(UN).

Central Agency for Public Mobilization and Statistics (CAPMAS) is the main body of the national statistical system, which includes partners for statistical work, to be responsible for coordination among them in the field of data dissemination. CAPMAS relies on a variety of sources of data, namely censuses, field surveys, and administrative sources.

The main purpose of official statistics is to use them to identify general trends in society and decision-making, whether this decision is public or private, because they carry many benefits for the benefit of citizens. The value of official statistics is becoming more and more user-friendly because its content creates a general understanding of the state and trends in the country and the world. Therefore, official statistics are of the utmost importance as follows:

- To provide the official statistical figure based on accurate scientific bases according to the latest international recommendations and standards in the field of preparing official statistics in all fields to meet the needs of policy makers, decision makers and interested parties from one source.
- The need to raise statistical awareness through the provision of information through the media and cooperation with universities and other research institutions and to ensure the citizen's right to access information on the basis that the statistics is a public good.
- Provide official, correct and unbiased statistics on demographic, social, economic and environmental conditions and trends to serve citizens.

2. Methodology

Transformation of data to big data revealed a lot of questions Why Data Revolution? What factors led to the data revolution? What is the need for data revolution?

The report, prepared by the United Nations Secretary-General's Independent Advisory Group of Experts, speaks of an "explosion" in the volume and production of data against "the increasing demand for data from all segments of society. PARIS21 takes a complementary approach and refers to "providing the right data to the right people at the right time" (PARIS21-

2015). This definition highlights the fact that data revolution should increase the use and impact of data on results.

Types of Data:

1-Structured data: - Regular databases, tables with more reliable relationships, get accurate information (statistical tables).

2- Semi Structured data: - Partial data is structured and another is unstructured

3- Unstructured data: - We cannot deal with it by using normal databases and thus become within the big data (unstructured data represent 80% of the data in the world).

Recently, there has been an urgent need for the use of non-traditional data sources, especially big data that has not yet been used to produce official statistics or to extract information related to various statistics such as health, economics, labor, transport, migration and the environment.

The availability of big data is a unique opportunity to support the International Sustainable Development Agenda 2030, as never before. Enhanced data production and use in policy development and monitoring are increasingly recognized as an essential tool for development.

The MDG monitoring experience has clearly demonstrated that effective use of data can help promote development efforts, implementing programs, monitoring performance and improving responsiveness to the increasing demand for data. Therefore, the Sustainable development requires a data revolution to improve the availability, quality, timeliness and classification of data to support the implementation of the sustainable development plan at all levels in all regions. Also the localization of the sustainable development goals on the basis of local priorities will be key to making them concrete and relevant.

Big data sources should be effectively used to enrich the sources of official statistics so that data needs can be met in new and timely development areas, and detailed spatial data can be produced and made available to decision makers. This show that the innovative power and transformation of information technology can be utilized from data collection through mobile phones to data dissemination through advanced data visualization tools, such as visualizing data on maps.

Big data are data sources that can be –generally– described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making."

Big data characteristics:

1. Size: The number of terabytes of data obtained daily.
2. Diversity: The diversity of data between structured, unstructured and semi-structured and diverse sources of different sites and different data, whether images or videos or emails and others.

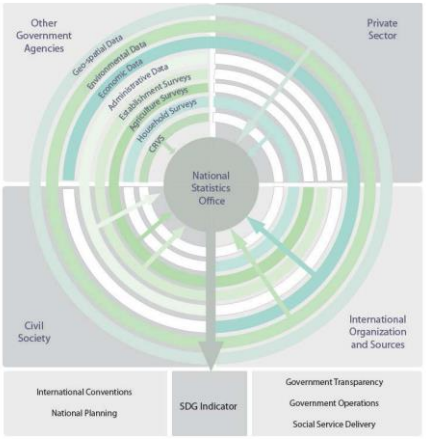
3. Speed: The speed of data access and analysis to suit business requirements.
4. Health and accuracy: the reliability of data after analysis of decision makers

Big data is characterized as data sets of increasing volume, velocity and variety; the 3 V's. Big data is often largely unstructured, meaning that it has no pre-defined data model and/or does not fit well into conventional relational databases. Apart from generating new commercial opportunities in the private sector, Big data is also potentially very interesting as an input for official statistics; either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative sources.

Figure (1): The four V's of Big Data

Sources: Mckinsey global institute, Twitter, Cisco, EMC, SAS, IBM, MEPTEC.

Figure (2): Relationship between National Statistical System and Sustainable Development goals



Source: PARIS21 et al., (2015) Data for Development - A needs assessment for SDG monitoring and statistical capacity development

3. Results

The Role of Big Data in sustainable development goals:

- It is very important for governmental and public sectors that are the most intensive users of data in both developing and developed countries.
- It is also very important for the business sector, as it provides endless opportunities to build and develop new business models; creates new products, services, and knowledge; in the framework of economies of digital data and information, internet, and the Fourth Industrial Revolution.
- The importance of capacity building, especially in developing countries, and developing global platforms for Big Data and their applications under the auspices of the UN.

Projects and initiatives related to Big Data:

- Some projects began to take the practical feature; through the initial agreement with data sources, such as the Project of Consumer Price Index.
- Some projects take the form of initiatives; which are presented by CAPMAS to some concerned parties to study the possibility of transformation into actual practical projects, such as health initiatives/ projects.

Table (1): Projects and Initiatives of Big Data

Project	Key Features
Collecting data for consumer price index	- The aim of the Project is to produce the indicators of price indexes in a more innovative way, using internet to collect prices data in a way that is faster, more accurate, higher quality, and integrated; to produce instant reports to support decision taking. - There is an actual agreement with three sources to send a list of major commodity groups. Also, there is current coordination between the Sector of Regional Branches and the Sector of Economic and Mobilization Statistics to choose ten sources to collect prices data. - CAPMAS trains and qualifies (2) employees from the source on technological courses at CAPMAS, according to the protocol signed between the source and CAPMAS.
Measuring crop production by satellite images	- The project aims at utilizing satellite images (as a source of Big Data) in addition to agricultural statistics to measure the productivity of main agricultural crops, management of field irrigation and water allocation, similar to the Australian experience in this regard. - There is current coordination with Ministry of Agriculture and a protocol for cooperation with the National Authority for Remote Sensing and Space Sciences – NARSS and the Arab League to use periodically updated satellite images in the project.
Tracking the movements of patients with epidemic diseases, such as Hepatitis	- The aim is to track the movements of patients with Hepatitis through spatial data of mobiles, to control and reduce the prevalence rate, similar to the American experience in this regard. - Coordination with Ministry of Health to choose (2) governmental hospitals and (2) private hospitals to initiate the project.
Big Data and health care	- The Project aims to collect data on each cancer patient in the country, to determine and analyze patterns or trends, to suggest ways of treatment, follow-up treatment results, and provide tips to

Project	Key Features
	avoid illness in the community. There is coordination with Ministry of Health and some hospitals of cancer treatment, such as (57357), or National Cancer Institute to detail items of cooperation with CAPMAS to apply this initiative.
Big Data and Transport statistics	- The aim of the project is to utilize data of surveillance cameras to estimate the traffic density on high ways, similar to the experience of Hungary, which helps to estimate the traffic movement across borders, in addition to using data on vehicles to estimate the traffic movement.

Source: El-Deeb, Khaled (2016) "Using Big Data in Official Statistics", Research paper presented at "Workshop of Data Ecosystem for Sustainable Development". Cairo: CAPMAS.

In 2014, the United Nations Statistical Commission established the United Nations Global Working Group (GWG) on big data for official statistics to develop and test the use of new data sources and new technologies. In this regard, CAPMAS represents Egypt as a member of the Global Working Group-on Big Data GWG, which held its first meeting in China in 2014, the second in UAE in 2015, and the third in Ireland (Duplin) in 2016. The last conference stressed on capacity building and using Big Data in an integrative way with indicators of sustainable development.

GWG aims to reduce obstacles for developing countries in the use of big data, such as satellite data, mobile phone data, scanner data, and social media data.

According to Mobile phone data, it can help in determining where tourists and migrants come from, how long they stay and where they go. This is consistent with SDG indicators that help to monitor progress on target 8.9 (promote sustainable tourism which creates jobs, promotes local culture and products) or target 10.7 (to facilitate orderly, safe, and responsible migration and mobility of people).

Big data can also be used to create passenger patterns and traffic control which related to indicator 9.1.2 (Passenger and freight volumes, by mode of transport). The accuracy of information that can be obtained through the use of mobile phone data is much higher than that obtained through traditional surveys. The time lag from data collection to analysis can also be greatly reduced.

According to Egypt, almost every person has a mobile phone and there is an increasing number of people has mobile phones and access to the internet. Also the overall prevalence rate of cell phones exceeded 110%, due to

development of their numbers to 99.50 million lines in August 2017, compared to 95.8 million lines in 2014. The number of internet users through cell phones increased to 31.8 million users in 2017, compared to 20.3 million users in 2014, with a percentage increase of 56% between the two years. This is what stated in target 9.C (significantly increase access to information and communications technology and strive to provide universal and affordable access to the Internet in least developed countries by 2020).

Table (2): Some Indicators of the ITC Sector in Egypt, August 2014 - August 2017.

Item	August 2014	August 2015	August 2016	August 2017
Subscribers of Cell phone (million line)	95.84	93.50	96.25	99.50
Prevalence rate of cell phone (%)	112.19	107.47	108.64	110.34
Users of internet through cell phones (million users)	20.28	25.24	28.77	31.78
Users of USB Modem (million user)	4.02	4.03	3.36	3.28
International Capacity for Internet (billion b/s)	406.5	649.14	961.62	1213.86
Users of ADSL (million user)	2.93	3.65	4.35	4.92
Percentage of users of internet through cell phones to total users of mobile phones (%)	21.16	26.99	29.89	31.94
Total capacity of PBX (million line)	15.42	17.50	15.88	19.21
Number of fixed/land line subscribers (million subscriber)	6.85	6.00	6.33	6.27
Prevalence rate of fixed/land line (%)	8.15	7.0	7.19	6.8
Number of centrals	1668	1580	1496	1550

Source: Summary Report on Indicators of ITC. Cairo: Ministry of ITC. Issues Dec. 2017, Sept. 2016, and Sept. 2015).

4. Discussion and Conclusion

1. It is proposed to develop a national strategy for big data and to benefit from international experiences, aiming to reflect cooperation with partners in the statistical system and partners in the big data issues.
2. It is proposed to form a national team representing the public and private sectors, civil society organizations and NGOs on the use of big data for official statistical purposes and production of official data.
3. Activate the use of big data to serve sustainable development goals from all partners in Egypt.

References

1. Data Ecosystem Report to Enhance Sustainable Development in Egypt (2018), CAPMAS.
2. Mckinsey global institute (2011), Mckinsey & company.
3. El-Deeb, Khaled (2016), "Using Big Data in Official Statistics", "Workshop of Data Ecosystem for Sustainable Development", Cairo: CAPMAS.
4. Daas.P, & der Loo. M. V. (2013)," Big Data and official statistics" Statistics Netherlands.
5. www.unglobalpulse.org.
6. https://unstats.un.org/bigdata

Statistics monopoly: No room for nostalgia.

Magued Osman
Professor of Statistics, Cairo University and Former Minister of Communications and Information Technology, Cairo, Egypt

Abstract

Technological innovations are shaping the way businesses across the world will operate in the future. Such changes will create a new world. With a strong trend toward global economic integration, changes will not be confined to developed countries and will impact developing countries as well. Disruptive innovations are on the rise and are making paradigm shifts in many sectors including manufacturing, finance, transportation, education, health, agriculture, media and entertainment. The expected growing number of robots, sensors, satellites and internet of things applications will generate a volume of data beyond what statisticians can imagine. The traditional model of bureau of statistics as the main producers of data statistics is no longer the case. Monopoly by governmental organizations in producing and disseminating data is vanishing. The large data sets, such as censuses and household surveys, that was once a pride of statistical office are a small fraction of data that are/can be generated every day by Google, Facebook, Netflix, YouTube, or Amazon. A new model is suggested to re-engineer the statistical system in developing countries to allow for an inclusive approach to all players producing and using data and big data while respecting the fundamental principles of official statistics. The model takes into consideration the challenges facing the statistics discipline in the data science era and address governance issues related to the production, analysis and dissemination of big data without compromising professional considerations, including scientific principles and professional ethics.

Keywords

Data science, big data divide, disruptive innovations, national statistical offices

1. Introduction

According to the Fundamental Principles of Official Statistics, official statistics provide an indispensable element in the information system of a democratic society, serving government, the economy and the public with data about the economic, demographic, social and environmental situation. Decision on the methods and procedures for the collection, processing, storage and presentation of statistical data is determined according to strictly professional considerations, including scientific principles and professional

ethics. Furthermore, statistical organizations should disseminate information according to scientific standards on the sources, methods and procedures derived from the theories of statistics[1].

Disruptive innovations are changing the way business is done and is creating valuable "big data" that does not conform necessarily to the principles that statistical organizations adhere to. A new model to re-engineer the eco-system of statistics is needed to avoid a complete de-association between the world of statistics and the future world of data production.

2. How the world of statistics is going to change?

Technological innovations are shaping the way businesses across the world will operate in the future. Such changes will create a new world. With a strong trend toward global economic integration, changes will not be confined to developed countries and will impact developing countries as well.

Disruptive innovation is a relatively new concept that describes businesses that create a market value from new costumers or successfully address an overlooked niche of less demanding customers[2]. Compared to sustaining innovations, disruptive innovations are a) addressing new markets rather than existing markets, b) dramatic and game changing rather than improving performance, lowering cost and incrementally changing, c) operating in unpredictable market rather than a predictable one, and, d) facing the failure of traditional business methods by creating new ones[3]. Disruptive innovation is on the rise and are making paradigm shifts in many sectors including telecommunications, manufacturing, finance, transportation, education, health, agriculture, media and entertainment.

Another game changer is artificial intelligence (AI). Estimated worldwide revenues from the AI market is expected to grow from 3 billion US$ in 2016 to 90 in 2025. AI will increase labor productivity, and hence global enterprise investment in the AI market is expected to increase from 12.4 billion US$ in 2018 to 231.9 in 2025. With one billion video cameras connected to AI by 2020, the amount of data created is growing exponentially and calls for more and better analytical solutions.

Traditional enterprise storage will soon become obsolete, as worldwide revenues from enterprise storage using cloud services is expected to increase from 17% in 2015 to 89% in 2025. The revenues from cloud services is expected to increase from 7.1 billion US$ to 50.7 billion. The expected growing number of robots, sensors, satellites and internet of things applications will generate a volume of data beyond what statisticians can imagine. The increasing revenues generated from big data will move the center of attention

[1] UN (2013)
[2] Christensen (2015)
[3] Babaian (2017)

from the traditional players of the statistical system to new players coming from the private sector and most of them are multinational companies. Medium term predictions presented in Table 1 show that the average annual growth rate in revenues resulting from digital transformation is higher than the growth expected in many other sectors of the economy. In many cases, the annual growth rate of revenues is double-digit. Such growth will result in an increasing amount of data generated as a by-product of the activities associated with digital transformation and will subsequently raise the demand for data scientists.

These complex and interacting technological changes will have its impact on the discipline of statistics whether on the academic side or on the practice side as the "world of data" will explode in volume and will diverse in nature. The volume of data sets, such as censuses and household surveys, that was once a pride of statistical offices, are only a small fraction of data that are/can be generated every day by Google, Facebook, Netflix, YouTube, Uber, twitter or Amazon. The volume of data generated outside the traditional statistical system is increasing exponentially. For example, the hours watched on Netflix per minute increases from 70K hours in 2017 to 266K hours in 2018 and to 694K in 2019. Such wealth of data is used to understand the mood of the subscribers in addition to their preferences.

The traditional model of bureau of statistics as the main producer of data will come to an end and will be challenged by other non-state actors. Shift from national to global and from public to private will dominate data ownership and will reduce the national ability to use information to support and take decisions and to formulate public policies.

The market share of governmental organizations in producing, storing, analyzing and disseminating data is going to sharply decrease. As a result, the impact of statistical governmental organizations on decision making process will shrink and their control on drafting and indorsing professional standards and ethics governing big data is at stake. Furthermore, fuzzy big data governance might jeopardize the strict use of international concepts, classifications and methods to promote international consistency and subsequently, indicators (especially economic and financial) might lack international comparability.

The World Economic Forum expects a shift in the jobs landscape. Between 2018 and 2022, it is expected that 133 million new jobs will be created and 75 million will disappear. The list of the top emerging jobs includes "data analysts and scientists" ranking number one in the list of emerging jobs and "big data specialists" ranking number six in the list. This is an extra evidence that the integration of statistics in data science is imperative to improve employability of the new generation of statisticians.

These sets of paradigm shifts will change the way statistical analysis is carried out not only in terms of computational tools and storage capacity but

also in terms of methodological issues related to violation of assumptions and representation bias. New windows of opportunities will emerge to develop better techniques to model complex data and to support decision making in an environment loaded with a mix of structured and un-structured data, a mix of global and national data, and, a mix of public and private owned data sets.

Table 1: Global impact of digital transformation in US$ billion, 2018 to 2023.

	2018	2023	Average annual growth rate[4]
Global forecast of digital payments value	3598	6687	13.2
Global eCommerce revenue forecast	1823	2854	9.4
Global Digital Media revenue forecast	142	173	4.0
Global Digital Advertising revenue forecast	291	490	11.0
Global estimated revenue for smart home products	54.5	152	22.8
Global estimated connected car[5] revenue	18.7	31.5	11.0

Source: Statista (2019)

3. Challenges facing statistical systems in the big data era

Big data revolution is attracting huge investments and is expecting to add significant revenues to all types of business. It started already creating its own identity (terminology, methods, and computing tools). One should not undermine the risk of leaving the theory and practice of statistics isolated in a bubble that is perceived as too theoretical, too expensive, and too slow to respond to urgent needs of business intelligence and decision making. This paradigm shift is creating new challenges to statisticians and require the development of new theories, methods, and tools for data integration and visualization. Statisticians should not only participate in the debate of whether big data will result in a better society and better decision-making process, but should also appreciate and interact with the volume, velocity and variability of unstructured data.

According to Secchi (2018), such an effort will result in ''a better society'', only if there are ''problems'' where the characteristics of big data can improve

[4] calculated by the author

[5] a car is considered connected as soon as it is equipped with hardware which enables internet connection.

the quality of decision making and the decision makers themselves are aware of such opportunity and willing to take advantage of it[6].

Several challenges are facing the statistical system in all countries. The challenges include:

1) Reforming academic programs on the undergraduate level without compromising the depth of theoretical statistics.
2) Designing training programs that are beyond traditional statistical programs and widen the scope to address different issues related to data science.
3) Managing fake information and opportunistic behaviors that misinform the public and negatively affect trust in statistics.
4) Balancing individual privacy and general interest.
5) Collaborating with the owners of big data and convincing them of the added value of statistics.

While some insist that Big Data may provide an opportunity to 'leapfrog' statistical systems in developing countries, others argue that Big Data is largely Big Hype, and that traditional statistical concerns and methods limit its applications for official statistics[7]. Putting this debate aside, one can expect the emerging of a set of extra concerns that need to be considered by the statistical system and the statistical community in developing countries:

1) Brain drain of talents moving from developing to developed countries. This trend might aggravate with the increasing demand of companies in nearly all sectors for data scientists. Such trend will limit the potential of innovation disruptors to contribute to the growth of developing countries.
2) Globalization and ownership of big data sets. Most of the players in the big data era will be multinational companies. With new technologies, these companies will own data sets of citizens of developing countries to reinforce a new "big data divide" between developed and developing countries. This trend is already happening with the dominance of multinational companies in the telecommunication sector, in the online commerce sector and in the transportation sector.
3) Acquisition of start-ups and successful companies. Start-ups were considered at one point of time a boost for the economy in the developing world. They might not continue to do so as a strong trend is emerging of acquisition and merging between companies that will result in reducing the added value of these start-ups in creating jobs and paying taxes in developing countries.

[6] Secchi (2018)
[7] Secchi (2018)

4. Re-engineering the statistical system

Figure 1 illustrates data production under the business as usual scenario and under expected changes in the future eco-system of the fourth industrial revolution. It is suggested to re-engineer the current statistical system to allow for an inclusive approach that integrate all producers and users of data and big data while respecting the fundamental principles of official statistics. The model should take into consideration the challenges facing the statistics discipline in the data science era and should address governance issues related to the production, analysis and dissemination of big data without compromising professional considerations, including scientific principles and professional ethics.

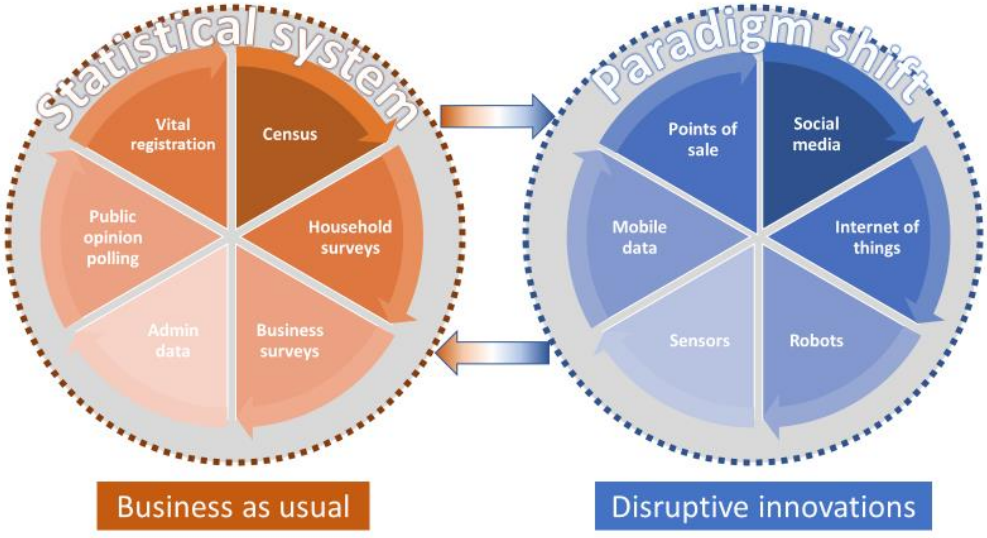


Figure 1: Framework describing the impact of disruptive innovations on data production.

To be able to conduct this re-engineering all stakeholders should embark in making necessary institutional, legislative and mind set changes. Furthermore, intellectual contributions should focus on bridging the gap between rigorous research and available big data, and, on integrating data and big data analysis in the decision-making evidence-based process.

5. Action plan

As presented above the era of big data is bringing several challenges and opportunities for the statistical community. To be able to re-engineer the eco-system, such challenges and opportunities need to be considered by all stakeholders. The following action plan is suggested:

Academia - In academia, the discipline of statistics needs to be re-shaped to take into consideration disruptive innovations that are becoming part of our daily life. Whether the term of "statistics" will remain as a brand of academic

major or university departments is becoming questionable. But more importantly, is the content of statistical education. The important question that need to be addressed is how to draw the balance while preparing students to the jobs of the future between a) what is currently taught to acquire a strong analytical and theoretical grasp of statistics and b) what should be taught to increase employability of newly graduate. More segmented recommendations should be thought of to differentiate between what should be done (if any) in statistical education for undergraduate statistical major, for undergraduate statistical minor and for graduate statistical program.

Given the pace of change associated with the fourth industrial revolution, quick fixes need to be introduced to statistical education including:

a) Introducing software for analyzing big data and for data visualization,
b) Developing joint academic programs that combine statistics and information technology,
c) Designing more state-of-the-art courses that look at data collection and data analysis in a more comprehensive way that take into consideration sources of big data,
d) Establishing protocols to foster the use of big data in research,
e) Exposing students to internet of things, artificial intelligence, and, how they will become a source of data, and,
f) Creating opportunities for students majoring in statistics for summer training or internship in information technology companies.

These fixes should not refrain the statistical community from making more strategic reform in teaching statistics.

Statistical national systems/offices – Official statistics will face the fact that Bureau of Statistics will not remain as "the" major data production organizations and should avoid ignoring other not-state actors who are mainly private sector companies. These companies might be after business opportunities in creating or analyzing big data or might own big data as a bi-product of new technologies associated with the fourth industrial revolution. In the era of data science, the following suggestions need to be taken into consideration:

a) Adopting an inclusion approach to engage owners and users of big data within the national statistical systems,
b) Changing the culture of statisticians from considering big data a threat for "good data" to an opportunity to improve the quality of big data and to reduce its limitations, as combining big data with traditional data can increase the timeliness and level of segmentation of the data.
c) Reforming the content of official statistics, such reform might include (but not limited to) transportation statistics to reflect volume of transportation network companies (Uber), expenditure to reflect national and international online shopping, foreign trade statistics to

reflect online transactions, tourism statistics to reflect online bookings, …

d) Developing methodologies to use big data in producing quick and affordable (often based on real time data) indicators to complement current economic, social and environmental indicators that might be expensive and less frequently collected to support decision making and inform policies, and,
e) Capitalize on international efforts to address the growing impact of globalization on reducing the ability of national organizations, especially in developing countries, to access big data produced by multinational companies (telecommunication, online shopping, and, transportation networking companies) and insuring the use of these data bases in addressing development goals including the sustainable development goals.

Institutional reform and governance – In a new era dominated by emerging technology breakthroughs in artificial intelligence, robotics, internet of things, cloud computing, autonomous vehicles, 3D printing, and nanotechnology, institutional reform and governance of the new statistical system should consider the following:

a) Create a consensus on national legislations and code of conducts among all stakeholders to create a win-win formula for cooperation that create value from creating and analyzing big data while maintaining trust in data as a source of evidence-based decision making.
b) Adopt an inclusive approach in the governance process to engage all stakeholders including (but not limited to) national statistical offices, governmental organizations producing data, private sector companies engaged in generating and analyzing big data, academic institutions, and, relevant professional associations.
c) Enforce that private companies producing and analyzing big data adopt and abide by statistical ethical considerations.
d) Develop a full fledge eco-system that allows the use of big data and its integration with traditional sources of data to create public knowledge and to inform policy on the national and local level.

It is obvious that the world of statistics is changing and present players in the developing countries have to step out of there comfort zone to create an eco-system that is more inclusive, more efficient and more future looking. The current producers of statistics should move away from data autocracy to data democracy and should give up on maintaining the status-co as no room for nostalgia.

References

1. Azzone, G. (2018). Big data and public policies: Opportunities and challenges. Statistics and Probability Letters. 136, 116–120. Special Issue on ''The role of Statistics in the era of Big Data''.
2. Bühlmann, P. & van de Geer, S. (2018). Statistics for big data: A perspective. *Statistics and Probability Letters,* **136**, 37–41. Special Issue on ''The role of Statistics in the era of Big Data''.
3. Secchi, P., 2018. On the role of statistics in the era of big data: A call for a debate. *Statistics and Probability Letters*, **136**, 10–14. Special Issue on ''The role of Statistics in the era of Big Data''.
4. Statista (2019). *Digital Economy Compass 2019*. https://static2.statista.com/download/pdf/DigitalEconomyCompass2019.pdf
5. UN (2013). Fundamental Principles of Official Statistics, Resolution adopted by the Economic and Social Council on 24 July 2013. https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf.
6. Christensen, C., Raynor, M. & McDonald, R. (2015). What is Disruptive innovation? https://hbr.org/2015/12/what-is-disruptive-innovation
7. Babaian, J. (2017). Disruptive innovation in healthcare. *Healthcare Leadership*. https://hcldr.wordpress.com/2017/01/10/disruptive-innovation-in-healthcare/

UN global platform as a data science collaboration environment for official statistics

Ronald W. Jansen
United Nations Statistics Division/DESA

Abstract

Traditional data collection tools such as surveys or censuses are expensive and take time to process. Some new technologies, such as the use of hand-held devices or the collection via on-line surveys, have shortened the delivery time. However, the traditional ways are not suitable to deliver on the promise of abundant and fast data for the 2030 Agenda. Therefore, the statistical community started looking into the use of new data sources to complement the traditional ones. Firstly, there is a wealth of administrative data sources, which the national statistical institutes can tap into. Then, there are the continuous streams of digital data generated by satellites, mobile networks or social media platforms. Nowadays we can access data from satellites, drones, mobile phones, social media applications and internet searches. Soon the 5G network becomes more broadly available, which makes it easier to use sensor data from cars, appliances or systems in your house or office. In 2018, the UN Statistical Commission created a Global Working Group (GWG) on Big Data for Official Statistics put into place the so-called UN Global Platform, which is a collaborative research and development environment for the global statistical community and all its stakeholder groups. Its platform organization is based on networking and marketplace principles, which facilitates the exchange, development and sharing of data, methods, tools and expertise, and accelerates data innovation. Ultimately, the purpose of the platform is to produce trusted data, trusted methods and trusted learning, which includes development of skills sets related to Data Science, since processing of Big Data requires use of advance technology and machine learning techniques. Statistical institutes have started creating data innovation centers to experiment with Data Science and Big Data sources. The reasons for wanting to have a Data Science facility in conjunction with a national statistical office are (1) to harness and exploit large digital datasets and data streams, (2) to develop and test algorithms, which lead to statistics and insights, and (3) to develop new skills in the task force of the statistical office, as well as attracts partner communities to work with the statistical office.

Keywords

Big Data; Data Science; Data innovation; Official Statistics; UN Global Platform

1. Introduction

The 2030 Agenda for Sustainable Development[1] was adopted by all United Nations Member States in 2015. It provides a shared blueprint for peace and prosperity for people and the planet, now and into the future. At its heart are the 17 Sustainable Development Goals (SDGs), which are an urgent call for action by all countries - developed and developing - in a global partnership. They recognize that ending poverty and other deprivations must go hand-in-hand with strategies that improve health and education, reduce inequality, and spur economic growth – all while tackling climate change and working to preserve our oceans and forests

In the 2030 Agenda, the global community of official statistics (through the United Nations Statistical Commission) was explicitly made responsible[2] for assuring a sufficient amount of relevant data to monitor progress on achieving the SDGs and its targets. The statisticians agreed on a global indicator framework which was adopted[3] by the UN General Assembly on 6 July 2017. This framework contains more than 230 indicators which would inform the national and international policies to achieve the SDGs. Policy makers emphasized the need for quality, accessible, timely and reliable disaggregated data to help with the measurement of progress and to ensure that no one is left behind. This new task for the community of official statistics implied also the need for the strengthening and modernizing of national statistical systems to produce statistics faster, more frequent and in more detail.

Traditional data collection tools such as surveys or censuses are expensive and take time to process. Some new technologies, such as the use of hand-held devices or the collection via on-line surveys, have shortened the delivery time. However, the traditional ways are not suitable to deliver on the promise of abundant and fast data for the 2030 Agenda. Therefore, the statistical community started looking into the use of new data sources to complement the traditional ones. Firstly, there is a wealth of administrative data sources, which the national statistical institutes can tap into. Then, there are the continuous streams of digital data generated by satellites, mobile networks or social media platforms. Nowadays we can access data from satellites, drones, mobile phones, social media applications and internet searches. Soon the 5G network becomes more broadly available, which makes it easier to use sensor data from cars, appliances or systems in your house or office. Whereas such digital footprints may raise legitimate concerns about privacy on one hand, they do provide valuable opportunities for statisticians, who need to inform policy makers and the public at large. The expectations on availability and

[1] See https://sustainabledevelopment.un.org/post2015/transformingourworld
[2] See paragraph 75 of https://sustainabledevelopment.un.org/post2015/transformingourworld
[3] See https://undocs.org/A/RES/71/313

timeliness of data are high, and surveys, which produce results that are by definition more general in nature and take more time to process, can no longer be the main collection tool of the statistical community.

This need for good and timely data is shared at the highest levels of the United Nations. In her video message[4] on 2 May 2019, Ms. Amina Mohamed, the UN Deputy Secretary-General, stated that *"Quality, relevant and timely data are essential to drive policies and programs, whether it be in our efforts to create decent work for all, monitoring environmental degradation, containing the spread of the Ebola virus or improving the living conditions in our urban areas. We need not only good data, but also real-time data."*

To make progress on data innovation, the UN Statistical Commission created a Global Working Group (GWG) on Big Data for Official Statistics in 2014, which was requested to provide strategic vision, direction and coordination for a global program on Big Data for official statistics and to promote practical use of Big Data sources. Since then, the GWG explored the benefits and challenges of the use of Big Data for official statistics and for the compilation of SDG indicators. It addressed issues pertaining to methodology, quality, technology, data access, legislation, privacy, management and finance, and provides adequate cost-benefit analyses.

More recently, in 2018, the GWG put into place the so-called UN Global Platform, which is a collaborative research and development environment for the global statistical community and all its stakeholder groups. Its platform organization is based on networking and marketplace principles, which facilitates the exchange, development and sharing of data, methods, tools and expertise, and accelerates data innovation. Ultimately, the purpose of the platform is to produce trusted data, trusted methods and trusted learning. This means that the platform will also be used as an environment for capacity building activities, which includes development of skills sets related to Data Science, since processing of Big Data requires use of advance technology and machine learning techniques.

The remainder of this paper will focus on this last point: capacity development for data innovation and data science on the UN Global Platform. The GWG has several task teams working on the various aspects of its mandate, and one of them works on training, competencies, and capacity development.

2. GWG task team on training, competencies, and capacity development

This GWG task team[5] focuses on new competencies and new skill sets, which are needed for the staff of statistical institutes to work with the new kinds of data sources, like Big Data. The main objectives of this task team are

[4] See https://unstats.un.org/unsd/bigdata/conferences/2019/default.asp

[5] See https://unstats.un.org/bigdata/taskteams/training/

to develop methods and tools for needs identification and assessment of "Big Data" competencies in National Statistical Systems (NSSs); to build a competency framework for Big Data acquisition and processing in the current data landscape; and to identify the existing supply of training of data scientists in academic and research centers. The members of the task team include national and regional statistical offices, development banks and some institutes with specific focus on Data Science, notably the Data Science Campus of the Office of National Statistics (ONS) of the United Kingdom.

The goals of ONS's Data Science Campus[6] are to investigate the use of new data sources, including administrative data and big data for public good and to help build data science capability for the benefit of the UK. A new generation of tools and technologies is being used to exploit the growth and availability of these new data sources and innovative methods to provide rich informed measurement and analyses on the economy, the global environment and wider society. The mission of the ONS Data Science Campus is to work at the frontier of data science and Artificial Intelligence – building skills and applying tools, methods and practices – to create new understanding and improve decision-making for public good. It defines data science as *"applying the tools, methods and practices of the digital and data age to create new understanding and improve decision-making".*

3. Data science for official statistics

Data science is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyse actual phenomena. Figure 1 illustrates how Data Science can be seen to sit at the intersection of mathematics, statistics, computer science, machine learning, traditional research and domain expertise. This Venn-diagram is adapted from original work of Drew Conway[7]. We will describe two examples of projects, which use Data Science to model and estimate statistics for government purposes.

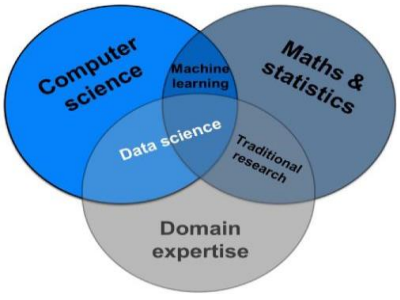


Figure 1: Data Science at the intersection of multiple disciplines

[6] See https://datasciencecampus.ons.gov.uk/about-us/

[7] See for example http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Examples of data science projects – (1) Mapping the urban forest

In September 2018, data scientists of the ONS Campus published the results of their project "Mapping the urban forest at street level". They developed an experimental method for estimating the density of trees and vegetation present at 10 meter intervals. The approach uses images sampled from Google Street View as the input to an image segmentation algorithm, which enables to derive a vegetation density map by percentage, for the road network of an entire city. The developed system is built on recent advancements in the field of deep learning for semantic image segmentation. This system makes use of an automated tree detection procedure coupled with street-level image data. The result is a consistent methodology that can be used to add value to existing tree valuation approaches, with the capability to assess urban vegetation from a remote location. Such benefits are important for policy making and urban planning.

Figure 2 depicts the segmentation labels produced by the PSPNet (Pyramid Scene Parsing Network[8]) and the Random Forest machine learning methods[9]. Challenges in the learning of these methods are that green vegetation will be less prevalent in autumn and winter months, and that not all tree species are green and not all green objects are trees. The PSPNet model is robust to all of these.

Figure 2: Comparison of vegetation segmentation methods, from left to right: ground-truth (Mapillary data), PSPNet, Random Forest, a* threshold. Images copyright Mapillary

Examples of data science projects – (2) Using mobile phone data for tourism statistics

Another example is the use of mobile phone data. Mobile phones send and receive signals from cell towers about 30 to 100 times a day with the only condition that the phone is turned on. Those signals constitute valuable information, which can be made into estimates about population density at any time of the day, can determine commuting partners, determine

[8] See https://hszhao.github.io/projects/pspnet/

[9] See https://datasciencecampus.ons.gov.uk/wp-content/uploads/sites/10/2018/09/ons-dsc-mapping-the-urban-forest.pdf

differences in weekday and weekend population, or inform about where people are going on vacation domestically and how many foreigners are visiting your country. All of those data are important to take decisions on public transportation, housing and infrastructure.

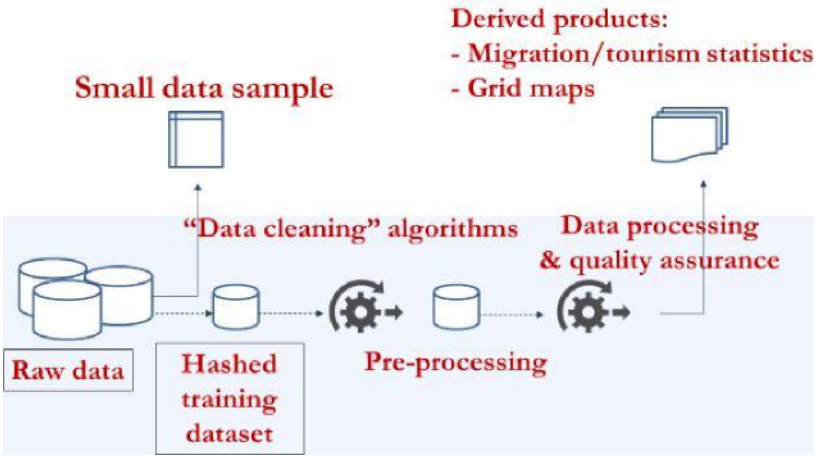


Figure 3. Flow chart of the processing of mobile phone records

The statistical office of Indonesia with support of Positium use mobile phone positioning data to estimate domestic tourism[10], internal migration, commuting patterns and population density. The positioning data (monitored over an extended period of time) will show a pattern about where a person lives, where he or she works and what some of her or his regular activities are. As illustrated in Figure 3, a lot of work is being done to pre-process data to take out "noise" and subsequently to run algorithms, which turn data records into statistics.

4. Conclusion

Tom Smith, Director of ONS Data Science Campus, presented recently[11] about "Data Science as a Team Sport" at the 5th International Conference on Big Data for official statistics in Kigali, Rwanda. He emphasized that partnerships and knowledge exchange are crucial elements for a successful application of Data Science. In the UK, a Government Data Science partnership was created to solve complex business problems using a combination of domain expertise, coding knowledge, machine learning and statistics skills on large and varied datasets.

Statistical institutes have started creating data innovation centers to experiment with Data Science and Big Data sources. Besides the Data Science

[10] See http://www.15th-tourism-stats-forum.com/pdf/Papers/S3/3_2 Indonesia's Experience of using Signaling MPD for Official Tourism Statistics.pdf

[11] See Conference agenda, Day 2, session at "Data Science And Capacity Development In Official Statistics", https://unstats.un.org/unsd/bigdata/conferences/2019/default.asp

Campus of ONS in the UK, Statistics Netherlands has created a Big Data center[12] for data innovation in official statistics. China, Korea Republic and Rwanda have concrete plans to establish such data innovation centers as well.

The reasons for wanting to have a Data Science facility in conjunction with a national statistical office are (1) to harness and exploit large digital datasets and data streams, (2) to develop and test algorithms, which lead to statistics and insights, and (3) to develop new skills in the task force of the statistical office, as well as attracts partner communities to work with the statistical office. Private sector, academia, research institutes and civil society can and are willing to support the statistical community, especially in the new domain of Data Science for official statistics. Together, significant progress can be made to fulfil the promise of timely, more frequent and more granular data to inform and achieve the sustainable development goals and targets.

References

1. United Nations (2015), *Transforming our World: The 2030 Agenda for Sustainable Development* A/RES/70/1
2. Hengshuang Zhao, Jianping Shi, et al. (2017) "*Pyramid Scene Parsing Network*", arXiv:1612.01105
3. Philip Stubbings, Joe Peskett (2018), "*Mapping the urban forest at street level*", ONS Data Science Campus, Newport, UK
4. Titi Lestari, Siim Esko, et al. (2018), "*Indonesia's Experience of using Signalling Mobile Positioning Data for Official Tourism Statistics*", paper presented at the 15th Global Forum on Tourism Statistics, 28-30 November 2018, Cusco, Peru

[12] See https://www.cbs.nl/en-gb/our-services/innovation/big-data

Measuring the rural access index in the Philippines

Candido J. Astrologo, Jr.*[1], Patricia Anne R. San Buenaventura[2], Justin Angelo O. Bantang[2]

[1] Philippine Statistics Authority, Pateros, Metro Manila, Philippines

[2] Philippine Statistics Authority, Quezon City, Philippines

Abstract

The global Sustainable Development Goals (SDG) consist of indicators classified into Tier I, Tier II, Tier III and multi-tier indicators. Among the Tier II indicators is SDG 9.1.1. – "Proportion of the rural population who live within 2 km of an all-season road". This indicator measures the share of a country's rural population that lives within 2 km of an all-season road.

In the Philippine Statistical System (PSS), efforts addressing Tier II and Tier III SDG indicators are considered as priority statistical development programs as listed in the Philippine Statistical Development Program (PSDP) 2018-2023.

This paper attempts to compute for SDG 9.1.1. taking into considerations various assumptions and limitations. The methodology is based on the initial work of the World Bank (WB) which was explored by the Asian Development Bank (ADB) relative to its work on providing support to developing countries through policy advice, research and analysis and technical assistance to reduce poverty, increase shared prosperity, and promote sustainable development.

Specifically, this paper aims to estimate the proportion of rural population living within two kilometers from an all-season road in the Philippines. Technically referred to as Rural Access Index (RAI), this indicator will help policymakers to manage investments in road sector and to formulate rural transport programs and strategies to boost agricultural growth and reduce poverty.

The WB considers RAI as a key transport headline indicator which has been established to focus on the critical role of access and mobility in the reduction of poverty in developing countries. It provides a consistent basis for estimating the proportion of the rural population which has adequate access to the transport system.

Keywords

Sustainable Development Goals; nighttime lights; road network; rural access index; Philippines

1. Introduction

In September 2015, member states of the United Nations adopted the 2030 Agenda for Sustainable Development or the Sustainable Development Goals (SDGs). The SDGs is composed of 232 unique indicators of which 93 are

classified as Tier I (indicators with established methodology , regularly collected), 72 as Tier II (with established methodology, data not regularly collected), 62 as Tier III (no established methodology, methodologies are being developed/tested) and 5 as multi-tier ((different components of the indicator, i.e., numerator and denominator, are classified into different tiers)

One of the indicators which is classified as Tier II is SDG 9.1.1. – "Proportion of the rural population who live within 2 km of an all-season road". This indicator measures the share of a country's rural population that lives within 2 kilometers of an all-season road or are within an approximate walking distance of two kilometers (around 20-25 minute walk) from an all-season road.

Considered as priority statistical development programs as stipulated in the Philippine Statistical Development Program 2018-2023, the Philippine Statistical System through the Philippine Statistics Authority (PSA) spearheads some of the methodological work in addressing Tier II and Tier III indicators. Recently, a technical assistance from the Asian Development Bank enabled the PSA to estimate SDG 9.1.1 using the initial efforts done by the World Bank.

This paper attempts to come up with measures of SDG 9.1.1, that is, to come up with an estimate of the proportion of rural population of the Philippines which live within two kilometres from an all-season road. Technically referred to as Rural Access Index, SDG 9.1.1 will enable policymakers to identify and manage investments in road sector and to formulate rural transport programs and strategies to boost agricultural growth and reduce poverty, particularly in rural areas.

2. Methodology

The methodology used in the computation of the Rural Access Index is based on the initial work of the World Bank[1] . The software QGIS[2] was used in all steps of the methodology involving maps. Source of data include the PSA (administrative maps with boundaries, urban-rural classification of levels of government, and digitized road network), and the WorldPop (high-resolution gridded population distribution map for 2015). Meanwhile, all computer operations were performed using the open-source QGIS software.

Basically, the methodology is divided into three parts:

Part 1 – Estimating the Rural Population

1. Obtain an administrative map with barangay boundaries.
2. From the map, identify the urban barangays. Barangay is the lowest level of administration in the Philippines, i.e., National, Regional, Provincial, City, Municipal and Barangay levels. Barangays are the only

[1] http://documents.worldbank.org/curated/en/3673914721178152 29/Measuring-rural-access-using-new-technologies

[2] http://www.qgis.org

administrative levels which are classified either as urban or rural. Latest classification is as of 2015.

3. Using a map with provincial boundaries, extract the rural barangays by subtracting the identified urban barangays. At this point, the result is a map of rural areas with provincial boundaries.
4. The map of rural areas was superimposed to a gridded population map (with reference year 2015) taken from WorldPop[3].
5. Apply the "Sum" operation of the "Zonal statistics" function of QGIS to estimate the total rural population (TRP). A gridded population map of rural areas is also generated at this point.

Part 2 –Identifying the 2-kilometer radius from an all-season road

For the purpose of this study, all-season roads refer to national, provincial and municipal roads[4]. The following steps are then employed:

1. From the road network map of the Philippines, identify all-season roads.
2. Using QGIS, map the 2-kilometer radius surrounding all-season roads

Part 3 – Estimating the rural population within the 2-kilomter radius from an all-season road

1. Overlay the gridded population map (from Part 1 Step 4) to the map showing the 2-kilometer radius from an all-season road (from Part 2 Step 2) to estimate the rural population living within the 2-kilometer radius from an all-season road.
2. Apply the "Sum" operation of the "Zonal statistics" function of QGIS to estimate the total rural population living within the 2-kilometer radius from an all-season road (TRP2KM). At this point, a gridded total rural population map within 2-kilometer radius from an all-season road is also generated.

Part 4 – Computing for the Rural Access Index (RAI)

To compute for the RAI, the following formula is used:

$$RAI = \frac{TRP2KM}{TRP}\ x\ 100$$

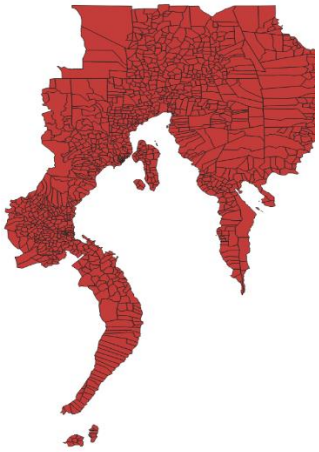
where TRP2KM = total rural population living within the 2-kilometer radius from an all-season road

TRP = total rural population

[3] https://www.worldpop.org/

[4] https://psa.gov.ph/content/road-classification

The series of images below visually describes the methodology above



Map of Davao Region with barangay boundaries

Urban Barangays in Davao Region

Rural Barangays in Davao Region

All-season road network map of Davao Region

Two-kilometer radius surrounding the all-season road network of Davao Region

Gridded rural population living within two kilometers of all-season roads in Davao Region

3. Results

In the Philippines, Davao Region has the largest contribution among regions in the southern part of the Philippines in terms of gross regional domestic product. One of its priority strategies is the Comprehensive Outcomes for Rural Empowerment Growth Triangle which aims to mobilize the region's resources to enhance its connectivity in order to achieve ease and mobility of access of people, goods and services through an integrated multi-

modal transport linkages and digital infrastructure. It is on this premise that the RAI was computed for Davao Region.

Using 2015 population data, Table 1 shows the Rural Access Index of the five provinces of Davao Region:

Table 1. Rural Access Index by Province, Davao Region: 2015

Province	TRP2KM ('000)	TRP ('000)	RAI
1. Davao del Sur	360.9	818.8	44.075
2. Davao Oriental	187.5	387.8	48.352
3. Davao Occidental	85.9	203.4	42.247
4. Compostella Valley	293.7	496.2	59.189
5. Davao Del Norte	354.6	393.7	90.057
Davao Region	1,282.7	2,300.0	55.767

TRP2KM = total rural population living within the 2-kilometer radius from an all-season road
TRP = total rural population

The RAI estimates show that three our of five provinces in Davao Region, namely Davao del sur, Davao Oriental and Davao Occidental, have values less than 50% indicating that less than half of their rural population lives within two kilometre from an all-season road in 2015.

Meanwhile, Davao del Norte has the highest RAI among the provinces in Davao Region at 90.057 indicating that 9 out of 10 persons in its rural areas live within two kilometres from an all-season road.

The whole Davao Region meanwhile has an RAI of 55.767 which implies that more than half of its total rural population are within a two-kilometer radius from an all-season road.

4. Discussion and Conclusion

In general, the RAI is a good indicator of access of rural population to roads. It provides a quantitative measure of the proportion of rural population with access to an all-season road to aid development planners to invest in road infrastructure that will facilities farm to market transport of goods.

Although the methodology is straightforward, it requires a comprehensive and extensive datasets which more often than not are not-so-dated. The availability of open source GIS software, i.e., QGIS makes the computation easy to perform and at a relatively no cost at all.

On the other hand, certain limitations and caveats are noteworthy to mention:

- There is no official definition of "all-season" road, i.e., "all-season" road is usually confused with "all-weather".
- There are different definitions of urban and rural areas per country and the classification may change over the years
- The methodology only takes into account the horizontal distance from the all-season road with no adjustments made on areas with elevation and presence of water surface.

This paper initially worked on estimating the RAI for the whole Davao Region and separately for its five provinces. The methodology can be replicated for other areas of the Philippine to come up with comparisons of RAI among its 17 regions and 81 provinces. It may be rigorous to estimate for RAI of municipalities and barangays as it will entail more computational time considering that there are 1,489 municipalities and 42,045 barangays in the Philippines as of 2018. Further, it will be difficult to identify the boundaries of road network since most road networks crosses over more than one barangay.

Future work for the PSA related to RAI include the following:

- Updating of the road network data using data collected from geo-tagging activities of the PSA.
- Use of the 2020 Census of Population and Housing (CPH) results as gridded population map.
- For Philippine Statistical System (PSS) to establish official definition of rural and urban areas.
- Adoption and approval of the methodology for the estimation of RAI in the Philippines.
- Release of statistics on RAI in the SDG Watch.

References

1. World Bank (2016). "Measuring Rural Access Using New Technologies"
2. Asian Development Bank (2019). Country Training Workshop on Data Disaggregation using Small Area Estimation and Big Data Analytics

Big data utilization for Official Statistics in Thailand

Hataichanok Puckcharern
National Statistical Office, Thailand

Abstract

As the Thai government gives importance to the use of big data, especially Government Big Data, to improve the efficiency of public administration and to improve the quality of people's life and Thai NSO also plays a big role on National Statistical System Management of the country, Ministry of Digital Economy and Society has proceeded in relation to the administration of Government Big Data in order to produce official statistics to guide and support national development, which has now begun by appointing a committee to drive policy implementation to use Big data, data centers and cloud computing, with the Deputy Prime Minister as president and 3 subcommittees have been appointed to drive the operation, consisting of

1. The Subcommittee on Enterprise Architecture Design of Government Data Integration System
2. The Subcommittee on Law and Regulation for Utilizing Big Data
3. The Subcommittee on Human Resource Development for Utilizing Big Data

Moreover, Thai NSO has conducted a study and improved the laws related to the management of statistical systems and also implemented Thailand Statistical Master Plan at the agenda base level, Function Base and Area Base in order to drive the use of big data.

Keywords

Big Data, Official Statistics, Thailand, National Statistical Office

1. Introduction

Royal Thai Government acknowledge the utilization of big data to improve efficiency in public administration and citizens' quality of life.

On 26 February 2018 The National Committee on Driving Policy Operation for Utilizing Big Data, Data Center and Cloud Computing was established. This committee objective are as follows:

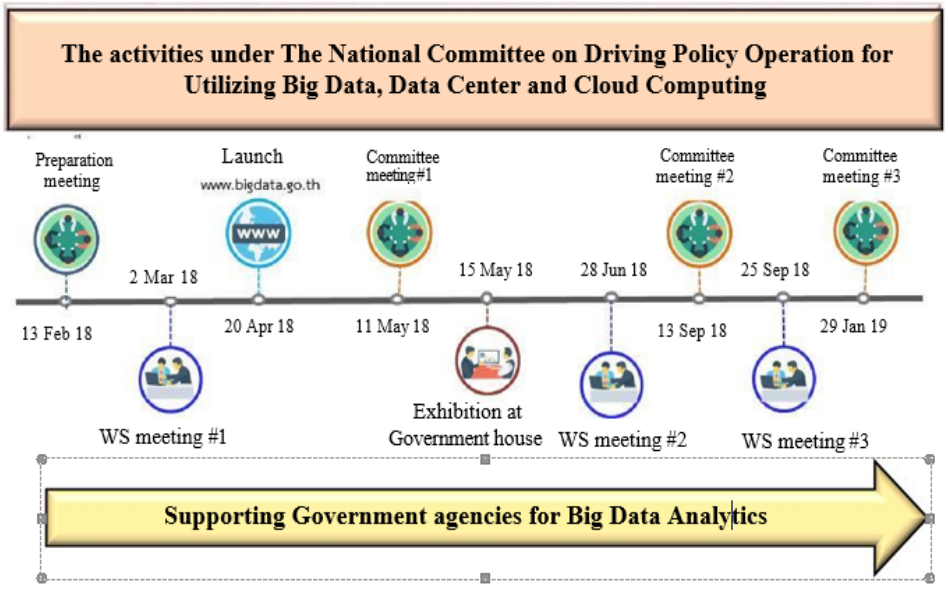
1. To plan a strategy for driving policy operation to utilize big data, data center and cloud computing
2. To address urgent and important issues to utilize big data for supporting government operation, monitoring progress of

government central database integration

3. To formulate a policy for supporting government enterprise architecture design
4. To regulate the data quality of all public sectors.

Deputy Prime Minister chairs the committee. The committee is comprised of the twentieth Permanent Secretaries of all government ministries in Thailand. The committee secretary is the Permanent Secretary of the Ministry of Digital Economy and Society and the assistant secretary of the committee is Director-General of National Statistical Office.

A week later, the first national workshop on big data utilization was launched. The objectives of this workshop are promoting the opportunity of big data utilization for public sectors to support national strategy Thailand 4.0, and integrating data across public sectors to data analysis. The workshop's participants were representatives from all ministries. In this workshop, the roadmap of driving big data utilization was presented. Besides the participants had defined three urgent and important issues which related datasets were available for analysis.



In order to perform its functions and to carry out its duties as fully and as efficiently as possible, the committee established three subcommittees. The National Committee on Driving Policy Operation for Utilizing Big Data, Data Center and Cloud Computing adopted the following resolution on May 11, 2018 in the first meeting.

Three subsidiary bodies of the committee are:

1. The Subcommittee on Enterprise Architecture

Design of Government Data Integration System
2. The Subcommittee on Law and Regulation for Utilizing Big Data
3. The Subcommittee on Human Resource Development for Utilizing Big Data

The Subcommittee on Enterprise Architecture Design of Government Data Integration System is chaired by the Permanent Secretary of the Ministry of Digital Economy and Society. The subcommittee is tasked with:

1. To design government enterprise architecture for government information and business process reformation
2. To issue an implementing guideline for driving the utilization of big data, data center, and cloud computing
3. To determine appropriate public administration process for data integration

The Subcommittee on Law and Regulation for Utilizing Big Data is chaired by Secretary-General Office of the Council of State. The subcommittee is tasked with:

1. To study rule, regulation, and law related to data access, data disclosure and data sharing across public sectors.
2. To draft an amendment of rule, regulation, and law being the limits and barriers to data access, data disclosure and data sharing across public sectors.

The Subcommittee on Human Resource Development for Utilizing Big Data is chaired by Secretary-General Office of the Civil Service Commission. The subcommittee is tasked with:

1. To study demand in human resource development on big data analytics and big data management of public sectors.
2. To design a course and to issue an implementing plan for human resource development of government on big data analytics and big data management for public sectors.

One month after the first committee meeting, the 2nd workshop was launched on June 26, 2018. This workshop aims to present the outcome of analysis the three issues from the first workshop, in addition, to build the understanding of big data analytics process and big data utilization approach, and to demonstrate how to use CAT Big Data for big data management. CAT Big Data is a platform on cloud service that assist data scientists and data analysts in handling big data analytics life cycle such as problem analysis, data preparation, modeling, data analytics, deployment, and evaluation/monitoring.

This platform was developed by CAT Telecom Public Company Limited, the state-owned company under control of Ministry of Digital Economies and Society.

Three months after the second workshop, the second committee meeting

was on 13 September 2018. The committee adopted the following resolutions:

1. The standard and guideline for information infrastructure architecture design was drafted by The Subcommittee on Enterprise Architecture Design of Government Data Integration System. This document's objective is to encourage and support data sharing across public sectors effectively.
2. The concept of roadmap for government human resource development on big data utilization was created by The Subcommittee on Human Resource Development for Utilizing Big Data.
3. The data governance framework was published by Digital Government Agency.

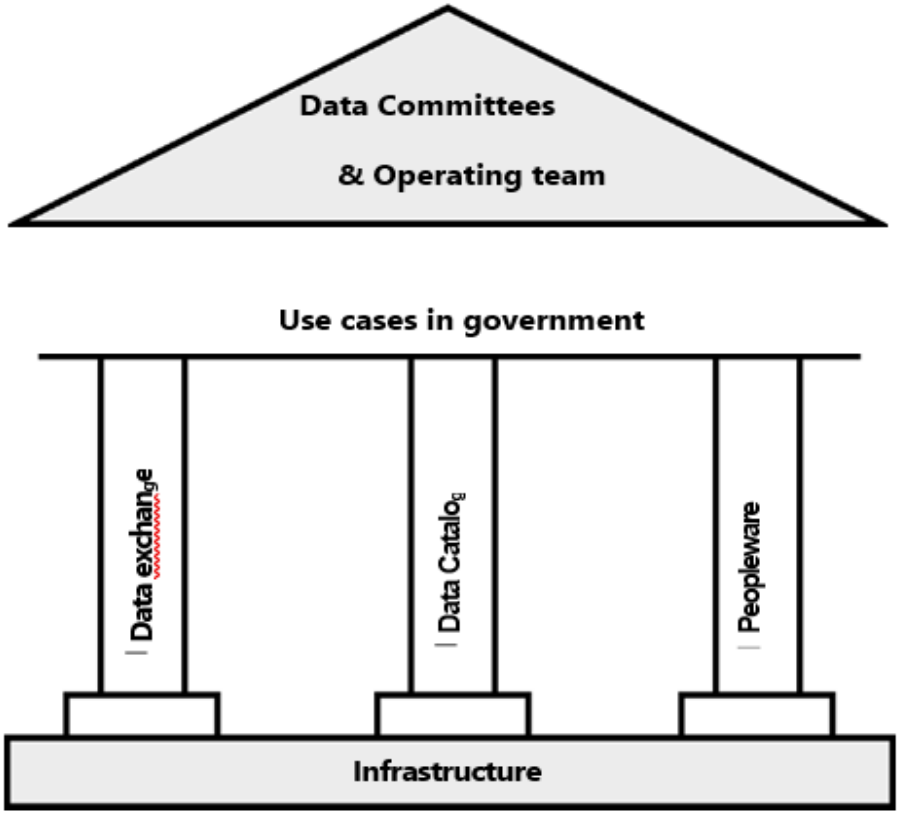
Moreover, the committee appointed Subcommittee on Enterprise Architecture Design of Government Data Integration System to assess the data center of all ministries by following the standard and guideline for information infrastructural architecture design.

The latest committee meeting, the third meeting was on January 29, 2019.

1. The Subcommittee on Enterprise Architecture Design of Government Data Integration System recommended the committee to adopt the resolution that designed Ministry of Digital Economies and Society to be the provider of Government Data Center and Cloud (GDCC). In order to service all departments from each ministries whose data center was below the standard in the document "The standard and guideline for information infrastructure architecture design". According to the results of data center assessment, the result demonstrated that many department data centers were below the standard.

In addition, this subcommittee recommended the committee to adopt the (draft) Government Big Data Analytics Framework. The Government Big Data Analytics Framework can separate into six parts as follows:

1.1 Infrastructure
1.2 Data exchange
1.3 Data catalog
1.4 Peopleware
1.5 Data committee and operating
1.6 Use cases in government



2. The Subcommittee on Law and Regulation for Utilizing Big Data recommended an amendment, which related to government data sharing across public sectors. According to the study of subcommittee about the limits and barriers related to data access, data disclosure and data sharing across public sectors, there are two limits and barriers.

   First, every ministries have their own act designed to protect government data and personal data and to control how the government uses it. It requires government agency to handle the personal data of citizen responsibly. For this barrier, the subcommittee drafted the new act called "Digital Government Act". This act will be the new mechanism to reform public administration and citizen service to One Stop Service (OSS). This leads to digital government that all government data can easily share and exchange across public sectors and the workflow of each public sectors can securely and efficiently.

   Second, in practice the government officer always reject to disclose data and to grant data access right to others officer from

other government agency. In this case, the subcommittee recommended drafting the new regulation on government data governance.

3. The Subcommittee on Human Resource Development for Utilizing Big Data drafted roadmap framework to develop government human resource for big data utilization. The objective of the framework is human resource development for utilizing big data and leads to increase economic value by targeting government officer to three groups that are users, system developer, and data analyst. The framework can be separated into three phases as follows:
    a. Short phase: to create and develop innovation called "Sandbox". The sandbox will be holistic human resource development model for utilizing big data.
    b. Middle phase: to draft proposal for human resource management.
    c. Long phase: to extend the development model to build up data science, data engineering, and data analytics skill to all public sectors. The data governance framework was published by Digital Government Agency.
    d. Moreover, the committee appointed Subcommittee on Enterprise Architecture Design

In the future, when government data can exchange across all public sectors the official statistical will be calculated automatically.

Accommodating Big Data in Nepalese Statistical System: Challenges and opportunities

Suman Raj Aryal
Cetral Bureau of Statistics

Abstract

Intellectual discourse on big data has already been introduced in the national statistical system (NSS) of Nepal. Nevertheless, the discourse in not expanded to the wide range of data producers and users of data. The big data is not yet officially used as an alternative source of data for decision making. Some private organizations working in the field of data are also found enthusiastic on big data. The recent National Strategy for the Development of Statistics (NSDS), 2018-2023 of Nepal has explicitly recognized the big data as a new initiative which sooner or later can be a complimentary source of data. As such, big data initiative is expected to stand as additional means of data supply regardless of its complexity and statistical aspects.

Irrespective of traditional sources of data, viz., census, surveys, administrative data and research outputs the big data by its very nature are difficult to handle in a country like Nepal where the statistical infrastructure is not built adequately. It needs, on the one hand, huge amount of investment to mine and process the big data and a legal acceptance from the State on the other. The Fundamental Principles of Official Statistics (FPOS) should not be violated while recognizing the big data. The question of confidentiality and the data security are equally important. Neither the national statistics offices (NSOs) nor any other agencies in the NSS have capacity to process, store and disseminate the big data at the moment. There is another important question which is about the selection of sources of big data. In the very beginning of this initiative, Nepal is not in a position to mine all sources of big data. The big data generated from social media are less likely to be appropriated since these require approval from multinational private companies. Again, the FPOS puts a dilemma, i.e., whether or not data generated by the business institutions should be regarded as official statistics. Validation of big data should pass through several statistical gates. At the moment, Nepal can use the CDR data for various purposes. The call detail records (CDR) and ATM data can be used to assess the migration, access to public services, viz, mobile banking, online shopping, unauthorized trade in the open border, employment mobility, weather, land use and so on by real time.

Big data has, as in other countries, provided us an opportunity to bring together the statisticians, the data scientists and the IT professionals. The knowledge and expertise of these three core professionals are very likely to

bring data in an advanced, objective and widely acceptable formats. Huge unused data will be supplied for the decision makings that take place at governance and business lives. It will fill gaps in socio-economic data. The new technologies and knowledge, if not replace, will gradually lessen the traditional sources of data - census and surveys. Despite the future of big data is also in 'cloud', it is already penetrating the horizons of national statistical system and minds of data users. The NSS of Nepal has curiously looked at the big data initiative and taken it as a complimentary wide source of socio-economic data. In this connection, the NSO has also introduced the digital data in its proposed new Statistics Act. It will no doubt open another door for the entry of big data in NSS of Nepal.

Keywords
Big data, data supply, NSDS, NSS, NSO, social media, statistical infrastructure, statistics act

1. Introduction

Modernization of society with the adoption of rapid change in the technology, more specifically, information and communication technology has advanced a new data sources namely Big Data. The term “Big Data” has also been coined in Nepal in recent years. The term has been famous in the sector of information, communication and technology. Big Data usually mentions to moreover large and complex data sets which are often generated with the continuous use of ICT tools or ICT based tools. The conventional software applications are not suitable to process such big data sets effectively due to the challenge of capturing, storing, transferring, and processing within a tolerable time frame. However, the big data of any sector should be transformed into usable statistics for the welfare of society. Statistical organizations should utilize the big data to produce a cost-effective statistical product that lead to better policies formulation, good decisions and strategic business moves.

Nepal has not more than 70 years of history in the production of modern statistics after the conduction of 1952/54 population census. The Nepalese statistical system exhibits the traditional characteristics in terms of data production and data exploitation for many years. However, in recent couple of years, Central Bureau of Statistics (CBS) which is a lead government statistical organization have made attempts to modernize the data production process despite the several statistical infrastructure constraints. Technology innovations and its application in day to day life cannot be ignored in global context. Big data as a by-product of day to day application of modern technology products like use of mobile phones, growing use of internet, use of digital transactions has a become a new dimension in the domain of official statistics. Nepalese Statistical System also have to be modernized in creating

new data sources and generating the new statistical products in cost effective way, timely and efficiently for smart decision making. CBS needs to make further progress in delivering timely, frequent and persistent data utilizing the big data. There needs a new dimension of approach in stepping up for the exploitation of big data in collaboration with the private sectors. This will enhance to combine and benefit from each other's strengths. This paper will discuss the issues, challenges and prospects to utilize the big data in generating the new statistical products.

2. Information, Communication and Technology (ICT) Use Perspectives in Nepal

Global technological advancements in the sector of information, communication and technology (ICT) has also not isolated Nepalese people in adopting and moving towards digital era. The number of internet users has increased rapidly in the last year in Nepal. The internet has become as the basic service of people's life. According to Nepal Telecommunication Authority (NTA) (January 15 2019) report, total number of voice service users are 39,979,561 which is 135 % of total population (29,514,745) and total number of internet subscribers (broadband service) 17,215,980 which is about 58 % of total population. This internet penetration index indicates the accessibility of the internet to the people of Nepal. Likewise, e-banking transactions has been increasing through electronic, interactive communication channels like intelligent electronic device (ATM, PDA, internet etc.) in Nepal. E-banking includes the systems that enable bank and financial institution customers, individuals or businesses, to access accounts, transact business, or obtain information on financial products and services through a public or private network, including the Internet.

As per the latest statistics from Central Bank of Nepal (Mid-March 2019), the entire country has a total of 4701 bank branches, 25.66 million bank accounts, around 6.4 million mobile banking customers and 0.86 million internet banking customers, 3049 ATM centers, 5.96 million debit card holders, and 0.1 million credit card holders. There exists an increasing online shopping's, digital billing practice in the business centers or complex. Likewise, social media (Facebook, Twitter, Instagram, WhatsApp, Skype, Viber, LinkedIn, etc.) users are also in increasing in trend in Nepal. Total number of Facebook users in Nepal were about 9.5 million in April 2018. Similarly, total number of Instagram Users in Nepal were 1.2 million, twitter users 2.3 million, YouTube users 3.8 million (Source: https://www.thesocialmediatoday.com/social-media-users-facebooktwitterinstagram-in-nepal-2018/). These facts indicate that use of ICT in daily life has been growing in Nepal. Such ICT behaviors if tracked properly and streamlined in the statistical framework, there is a high prospect of big data in producing some important aspects of statistical information, which may be useful in smart decision making. In this respect, the

role of government is necessary to be changed to adapt to quick decision driven big data technology. The sources of data from the use of internet, public private partnership programs, crowd sourcing, social media activities and data sharing activities has advanced the compulsion of big data system. The government service delivery and e-governance can be made efficient and transparent by introducing big data missions.

3. Prospects of Big Data in the Nepalese National Statistical System

Generally, NSS is composed of network of data stakeholders in terms of data respondents, data production, data users, academic in the field of statistics. According to OECD definition, the system is the collaborative of statistical organizations and units within a country that jointly collect (demand and data), compile, process and disseminate official statistics to its' users. The national statistical system of Nepal is envisaged by Statistical Act 1958. Other laws are also prevalent which has designated different agencies to produce statistics of their respective areas. In the future, NSS of Nepal need to deal with Big Data as a new dimension of data sources in the context of technological advancements. It is quite challenging to address the issues to operationalize within NSS in terms of defining and identifying data producers, data users, data suppliers and data products. However, NSDS of Nepal has made provision to integrate big data in national statistical system. A compatible statistical infrastructure has to be established while adopting the big data in the system.

Statistical literacy along with big data to users and producers are must to make the users to demand data as well as supply proper statistics with the utilization of big data. Academia and the research institutions are also one of the parts of NSS which should also involve big data in research and academic activities. The ICT oriented NSS can only fully exploit the Big Data in generating the quality statistics. Similarly, the NSS should address the issues on legal framework, in-depth policy, professionalism, standard concepts, definitions, classifications and methodologies, scientific methods, reference period, coordination, designated agencies for specific statistics, human resources capacity development, research and training, strengthening organizational units, advocacy and sustainable investment towards establishing Big Data statistical system.

4. National Strategy for Development of Statistics of Nepal as a platform to Big Data

National Strategy for Development of Statistics (NSDS) is the country's overall vision, mission, strategies and activities on the development of the national statistical system. The constitution of Nepal has ensured the statistical duties of federal, provincial and local level governments. It has made conducive environment to enhance the Statistical system. There has been

growing demand of official statistics from the users of private sectors, academia, media and public at large. In this connection, national strategy for the development of statistics of the country was necessary to be formulated. Government of Nepal (GON) recently endorsed NSDS of the country for five years (2018-2023). The strategy is the guideline to move ahead towards fulfilling data gaps in the various sectors by conducting statistical activities of the international standards. The NSDS is expected to generate nationally owned and produced data for all SDG indicators.

The strategy is made in compliance with the fundamental principle of official statistics to achieve the objectives of national statistical system. The strategy has given priority to supply the statistics as demanded by the users. NSS of the country needed to be in compliance with the international standards to support for the national development efforts. The long-term vision the NSDS of Nepal is to establish coordinated, active and strengthened national statistical system. The mission of NSDS is to develop the system that would produce, manage and supply the quality official statistics in compliance with the requirement of federal governance of the country and support the policies related to equitable development and people's welfare. The NSDS has set three strategic objectives to achieve the mission and vision namely:

1. Develop Statistical System establishing coordination among central, provincial and local governments in involved in statistical production activities
2. Arrange a regular supply of reliable and quality statistics that meet the needs of all levels of governments and the wider users for evidence-based policy formulation and development
3. Institutional Strengthening by improving legal and procedural provisions for the management of statistical operations

The NSDS includes the different strategies to fulfill the above-mentioned strategic objectives as follow:

1. Make institutional arrangement as per federal system
2. Institutional strengthening
3. Develop an effective system for statistical development
4. Produce and supply a quality statistic
5. Utilize all kinds of data sources in the production and supply of statistics
6. Adopt the statistical methodology that ensure the quality of statistics
7. Reform the legal provisions for the statistical development
8. Improve the procedural provisions of statistical operations
9. Strengthen the institutional unit of statistical agencies
10. Statistical Human Resources development
11. Use of modern information technology in the statistical activities

With the above mentioned facts, the NSDS of Nepal is favorable and encouraging to develop the big data system in the country. Further, the NSDS has mentioned in its working policy that Big Data will be integrated in the Statistical System of the country. Hence, NSDS is one of the evidence of the government's commitment and policy document in opening the door to adapt big data for the production and supply of new statistical products.

5. Big Data in Fundamental Principles of Official Statistics (FPOS)

As mentioned earlier, a space has been created to initiate big data activities within the framework of NSS. CBS as a leading authoritative agencies in Nepal has been following the fundamental principles of official statistics in the statistical production process and dissemination in statistical system of Nepal. It is a matter of study that whether FPOS can fascilitate statistical production process with the utilization of Big Data. However, natioanl statistical organizations need to follow FPOS while producing statistics through big data system.

According to FPOS Principle 1, the official statistics generated with the utilization of big data should provide an indispensable element in the information system with relevant data. Such big data are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' right to public information. The principle 5 shows the possibilities of big data as one of the statistical data sources that can be for statistical purposes. It should be taken care of the quality, timeliness, costs of big data. With the use of big data, the burden on respondents will be minimum in terms of collection and compiling the data.According to principle 6, the big data as one of the eminent source are to be strictly confidential and used exclusively for statistical purposes only.

6. Challenges and Opportunities to accommodate Big Data in National Statistical System

Challenges and opportunities' exist while accommodating new system in the prevailing system with transformation or reformation of the whole statistical system. Likewise, NSS of Nepal has to face with the challenges in the context of prevailing statistical system of the country. The issue on big data accommodation has also created opportunities in advancing towards more modernised official statistical system of Nepal. Some of points has been mentioned as challenges as well as opportunities below:

Challenges:

- Reform current statistical infrastructures like legal provisions, policy formulations, organisational structure, extension of ICT friendly statistical human resources, standards and methods, classification systems, Coordination and Cooperation mechanism etc.,
- Develop Guidelines, methodology

- Sustainable investment in Big Data System
- Institutionalization or institutional set up of Big data system
- Natioanl Big data system policy formulation
- Capacity development of prevailing human resources
- Ensuring following the FPOS
- Ensuring the confidentiality of individuals
- Identifying and selecting standard data products
- Statistical Literacy or Awareness to data users to data producers

Opportunities:

- Existence of FPOS
- Existence of NSS
- National strategy for Statistical Development
- Statistical Acts/ regulations
- Prevailing Institutional set up
- Strengthened and modernised statistical system
- Cost-effective statistical products of quality and timeliness
- Access to statistical products in smart decision making

7. Some initiations of government services

Government of Nepal is transforming its manual based government services to digital systems. Digital based licenses, land management, voter list, cadastral mapping, hospital management, meteorology and hydrological systems etc. are some of the efforts. These sources are also the sources of Big Data. Beside these efforts, the management information system established in various institutions can be mix up with related Big Data.

Government has already established the National Information Technology Center. This is central data repository of GoN. This center would be instrumental to integrate Big Data generated by various sources.

CBS has come up with national data profile (NDP) portal. This portal is designed in the principle of designated system. There are 5 different themes. Some of the indicators will be highly relied on Big Data.

Nepal's new political frame is based on federal structure. There are 753 local level and 7 provinces. The constitution has explicitly given the right and responsibilities to these local and province level governments in statistical governance. Generation of data in high granularity is the major challenge at this moment. The usefulness of Big Data in this context is very relevant and the government has given priority to expedite and mix up of different data sources.

8. Conclusion

The notion of Big Data is now not only related to its volume, velocity and varieties but also the citizens' right to information. Trustworthy statistical

information would only be realized if the statistical community could make some efforts to develop scientific methods and standards. The major challenge at this moment is how to make scientific inference based on Big Data.

Nepal has initiated some measures to expedite the usefulness of Big Data. NSDS has clearly mentioned the role of the Big Data in Nepal's NSS. Central Bureau of Statistics is looking forward to use the data held by the telecommunication and the E-banking system. Now GoN is planning to prepare tourism satellite accounts. CBS is supporting the concerned ministry. Some of the information might be collected through ATM transaction, super market data and CDR. This will enable to make some consumption statistics.

The emergence of the Big Data has changed the traditional concept of triangulation in statistical governance. Majority of the Big Data are owned by private and quasi-public sector and they have become the data holder. This new approach is the combination of conventional parts (i.e data producers, respondents and users) and the big data data holders.

It is not uncommon that the majority of least developed and developing countries are lacking of suitable statistical and information infrastructures. The crunch of human resources is another limitation. The globe should take care of these matters for efficient use of Big Data.

References

1. Central Bureau of Statistics (2017). Statistical System of Nepal.
2. Central Bureau of Statistics (2018). National Strategy for the Development of Statistics (in Nepali language).
3. Central Bank of Nepal (2019). Monetary and financial statistics.
4. Nepal Telecommunication Authority (2019). Management Information System Report-NTA MIS143.
5. Pradhan P., Shakya S. (2018). Big Data Challenges for e-government services in Nepal
6. United Nations. www.un.org/en/sections/issues-depth/big-data-sustainable-development.html
7. United Nations Statistics Commision (2019). 50th session of the Statistical Commission.

Analysis of blockchain for healthcare applications

Asma Adnane[1], McSeth Antwi[1], Farhan Ahmad[2], Chaker Abdelaziz Kerrache[3], Mohsen Farid[2];

[1] Department of Computer Science, Loughborough University, UK
[2] College of Engineering and Technology, University of Derby, UK
[3] Department of Mathematics and Computer Science, University of Ghardaia, Algeria

Abstract

The health industry deal with highly sensitive data which must be managed securely and carefully. electronic health records (EHRs) hold various kinds of personal data such as names, addresses, social security numbers, insurance number… etc, which must never be released to the public. However, this kind of personal data is highly sensitive and valuable on the black market, which brings various security risks to the healthcare industry. The healthcare industry has been looking to adopt new technologies such as Blockchain to cover their inefficiencies. Indeed, blockchain has been considered as a solution to all the privacy challenges, and many believe that blockchain is the horizontal innovation needed to transform various industries. The purpose of this paper is to carry out a feasibility analysis of blockchain for healthcare scenarios and conclude whether it's a suitable technology for the healthcare industry.

Keywords

Blockchain; Healthcare; Security; Privacy; Hyperledger fabric

1. Introduction

In the health industry, such as medical institutions and insurance companies, the infrastructure running healthcare applications and managing the related data deals with highly critical assets which are the electronic health records (EHR). EHR hold various kinds of personal data such as names, addresses, social security numbers, Medicare numbers… etc which must never be released to the public. However, this asset has been the target of various cyberattacks. Indeed, various medical institutions have been hacked and millions of patients records have been accessed, this kind of personal data is highly valuable on the black market, as much as $50 per EHR, which increases the security risks to the healthcare industry (Adefala, 2018). Several laws and regulations, such as HIPAA and the Data Protection Act 2018 (GDPR), have been created to provide guidelines to healthcare applications on how personal data should be managed, processed and secured in order to avoid fraud and theft.

Despite this, the industry still seems to be an "easy target" for hackers and this is due to the lack of technological understanding within the industry. This can be seen by the number of attacks and vulnerabilities exploited within the targeted systems, such as phishing attacks, which are successful in retrieving personal data. Moreover, the high success of some attacks such as ransomware, has shown the lack of basic security measures such as backup and system updates. These attacks not only affect patients' privacy and security, but it also affect the institutions themselves, and many financial and reputation damage have been caused. As these attacks become more common and easy to perform, there is an urgent need for robust and reliable ways to ensure data security, confidentiality, integrity and availability to authorised users only. Various institutions have been looking into cloud-based technology and various kinds of encryption techniques. But In the last few years, Blockchain technology has been highly suggested and acclaimed as one of the best solutions to solve the security issues in healthcare applications and others. Blockchain is based on a peer-to-peer distributed and decentralised architecture which puts emphasis on value and trust rather than the exchange of information (Ahmad 2019).

The incorporation of blockchain seems to be in line with the GDPRs goal of protecting data by giving control to the users and using hashes and consensus to keep data integrity and consistency. This has been driven only by the success that Bitcoin has achieved, as no other applications has been fully deployed on Blockchain. Many researchers, and Blockchain developers believe that blockchain is the horizontal innovation needed to transform various industries. Even though blockchain has been heavily linked to the healthcare industry, there is a lack of research into what existing blockchains could be used for the industry. The purpose of this study is to carry out a deep analysis of blockchain and conclude whether it's a suitable technology for the healthcare industry via simple tests.

1.1. Security issues in healthcare applications

Healthcare is very sensitive as it directly involves personal data, which must be secured from unauthorized access from the attackers. In this section, we discuss different major security issues along with recent incidents in the healthcare industry. According to GDPR regulations, medical data should be held by data controllers due to the sensitive nature involved. Currently, medical data is passed on only to concerned department, if consent is gained via proper channel. The law defines these genuine needs as the notification of new births; potential epidemic or pandemic break out; an issue by formal court order and serious crime has been committed. Even though legislation is improving in term of data management, medical records are still at risk to security breaches. Personal data such as names, addresses, Medicare numbers and social security numbers are extremely wanted in the black market.

According to a panel of experts at the digital health conference in 2011, Electronic medical records (EHRs) are valued at $50 at the black market which is extremely high in comparison to $0.25 for a credit card number (Adefala, 2018). One incident occurred at Howard University Hospital in 2012, where the medical technician released the patients' names, addresses and Medicare numbers to black market, solely for financial gains. Other attack on healthcare industry is the CEO Phishing attacks, where the hacker masquerades as an authority to induce individuals to reveal personal data. These attacks have high risks as the reveal data can include patients' information, or employees' distinct details including social security number, addresses, salaries etc. One example of CEO Phishing attack is the attack on Magnolia Health Corporation (MHC), where the hacker was successful to obtain substantial information about its employees using a spoofed email from its CEO. Other recent attack incidents include the ransomware attack on National Health Service (NHS) in 2017, where, the hackers used malware to encrypt NHS files. In order to access the data, the hackers demand a ransomware of about 300$ in the form of bitcoins (Gayle, 2017). Further, these attacks result in the cancellation of over 6,900 NHS appointments.

1.2. Related work

In order to secure data and prevent attacks, various solutions are proposed to tackle such scenarios. We categorize these solutions into two major categories: (1) Cloud-based solutions, and (2) Blockchain-based solutions. Indeed, various cloud applications have been explored within healthcare industry specifically for managing EHRs and patient's information. Clouds can minimize the cost subsequently, thus motivating to improve different healthcare services. For instance, prescription expenses can be reduced by 80% while utilizing cloud-based services (Omar Alia, 2018). Due to centralized and ubiquitous nature of the clouds, it provides a fantastic opportunity to access data (patient or employee) at any time from any place. One such cloud-based system is proposed by Dhatri, which allowed physicians to access patients medical data at any given time (Vassiliki Koufi, 2010). On the other hand, blockchain is in its early stage of development, and there are few use-cases in healthcare industry. For example, applications such as ***BitHealth*** and ***MedRec,*** which are designed to support healthcare applications. BitHealth uses bitcoin for storing and securing healthcare data and focuses on privacy. Bitcoin is used for payments and for insurance companies to retrieve medical history. However, it uses proof of work algorithm and depending on the size of the blockchain it will be slow and energy inefficient. The other use-case, MedRed, is an EHR management system created by MIT which focuses on improving tracking of these records. Patients also have some degree of control with their information and permissions are given to the patients, so they can decide whether to share data with professionals. MedRed is based on

Ethereum, it uses the same algorithm for consensus (proof of work) as bitcoin which is extremely costly and energy inefficient. Personal data will be stored off chain so records cannot be determining whether the record is legitimate. So, authentication is legitimate, but the data may not be accurate.

2. Methodology

In this section, we will first present an overview of Blockchain, and its different platforms. We will then provide the different evaluation scenarios we have run on a selected blockchain platform. Each scenario presents a pseudo-function of a healthcare application and has different users and purposes.

1.3. Overview of Blockchain

Blockchain is a distributed ledger system which allows multiple users to securely store and share data. It gets its name from that fact that transactions are stored in blocks which are linked with each other to form a chain. A block contains various information about a transaction, including the time of the transaction; the sequences of transactions; and a pointer to the previous transaction (Hash value of the previous block). Indeed, linking transaction blocks using the hash will ensure the integrity of data on each transaction (previous block can't be altered), and prevent from inserting new blocks in-between 2 pre-existing blocks. In addition, to ensure full data integrity and availability in the blockchain, each blockchain platform guarantee the key blockchain concepts: Data immutability, Shared ledger, consensus, Permissions and data encryption.. Given that, there are two different types of blockchains which have significant distinctions: Public and Private. Public blockchains are the purest form of blockchain which allows any user to join the blockchain and does not discriminate between users (Smith T. D., 2018). However, private blockchains seem to be used more in the industry as there are more security and privacy constraints, and users need to request permission before becoming a member of the blockchain. Members of a private blockchain can be further restricted with different access privileges.

The main advantage of a public blockchain is its autonomy. All users have similar privileges, and no party can control stored data, which means that users don't have to trust and rely on a third party. However, public blockchains are extremely large and consume a large amount of energy as no user is restricted access. On the other hand, private Blockchains tend to be smaller and flexible as only limited number of users are allowed access, and they have different permissions and access privileges.

1.4. Evaluation methodology

In this study, we have used Hyperledger fabric for feasibility on healthcare applications. Hyperledger fabric has been selected due to various factors, including the use of a strong hashing algorithm like SHA-2, and providing

different levels of control to diverse range of users. This is only possible with a permissioned framework like Hyperledger fabric. Unlike Ethereum, Hyperledger allows nodes to have different roles within the blockchain. Nodes can be restricted on read, create, update and delete rights. Even though "delete" rights are offered to different nodes, no data is deleted on Fabric. A delete on Hyperledger is a transaction which simply marks certain data as "deleted". Moreover, we have used Hyperledger composer which is a development toolset to develop business networks. Hyperledger Composer has a UI for configuring, testing and deploying the business networks called "Playground" which is the main tool being used for implementation. Playground allows developers to simulate business networks by utilising assets (goods or services that are stored in the blockchain); participants (members of the blockchain) and transactions (methods which participants interact with assets). In order to discover whether blockchain should be adopted in the industry, it must solve the key issues related to security, regulation compliance, scalability and flexibility. In term of security, the blockchain platform must be able to implement integrity, confidentiality and availability of the data. In order to test whether the healthcare industry can utilise blockchain; the business network must take steps to comply with the GDPR as much as possible.

1.5. Test approach and scenario

Hyperledger composer offers 3 different types of tests for blockchain applications: interactive test, automated unit tests and automated system tests. This business network will be using interactive tests to assess whether the scenarios could be implemented into blockchain. As well as scenarios, interactive tests will be used to check validation, verification, permissions and the overall performance of the blockchain..

To test the blockchain environment, the following scenarios have been designed:

Scenario 1- Basic scenario : This scenario tests the different access control between a standard user and specified member of the blockchain (patients, medical institutions or medical practitioners). Specified member will be able to view data on the blockchain whereas a standard user will have no access. Further to this, this scenario will confirm the use of a strong hashing function and the concept of a shared ledger. The patient and the medical practitioner should have a copy of the same transaction.

Scenario 2 - Permissioned Scenario: This scenario tests the level of permissions utilised on Hyperledger regarding create, read, update and delete operations. The goal of this scenario is to explore whether Hyperledger's permissions could be used to restrict different types of participants to ensure an extra layer of security and minimise the number of security threats.

Scenario 3 - Purging data Scenario: To be GDPR/HIPPA compliant patients must have complete control over their EHRs, this includes both giving

patients the ability to remove read rights from reading the EHR and deleting the EHR. The GDPR states the user must have the right to be forgotten. Consequently, this scenario tests the removal of patient data.

Scenario 4 - Data type scenario: This scenario tests how Hyperledger blockchains interact with different kinds of data. Within this scenario, the blockchain will have to cope with images and text to mimic the data used within the healthcare industry such as X-Ray's and their annotations.

Scenario 5- Encryption Scenario: This scenario tests the cryptographic capabilities available on the Hyperledger. To ensure that connection to the blockchain is secure and protected from man in the middle attacks a level of security must be available.

Finally, to run these scenarios we have created different roles with different permissions. These permissions and roles will mirror some of the different roles used in the healthcare sector and will illustrate how a permissioned blockchain can be utilised in different use case scenarios.

- **Admin:** complete access to all users and system resources.
- **Member:** Create, delete, read and update their own participant information.
- **Medical institution:** Create, delete, read and update their own participant information. A medical institution such as a hospital can view their employees' participant information and manage medical practitioners such as doctors, pharmacists, surgeons…etc.
- **Medical practitioner:** Create, delete, read and update their own participant information, Read/ update permissioned HER (If authorised by the patient) or refer it to other practitioners (granting access rights for other practitioner on HER they have been authorised to manage).
- **Patient:** Create, delete, read and update their own participant information and HER, grant or remove access rights to practitioners on their HER.

3. Results

This section will compile all the results from the implemented blockchain environment.

2.1. Security

Throughout each scenario, validation has been used to increase the fault tolerance of the developed blockchain. Even though, Hyperledger Fabric is described as fault tolerant; it does not enforce any fault tolerance within chaincode leaving it up to the developer.

Basic Scenario used access control to restrict resource utilisation to named roles (patients, medical practitioners and medical institutions). This achieves a superficial level of confidentiality by keeps personal data private to blockchain participants. Further to this, Basic Scenario showcases 2 key

concepts of a blockchain: shared ledger and hashing which together achieve an acceptable level of integrity. Sha-2 was used to hash each transaction ensuring users that transaction is accurate. There is no known breach to SHA-2 making it near impossible for a hacker to replace or create a transaction that fits on to the blockchain. The concept of shared ledger ensures that data within the system is accurate and unaltered because each peer of the blockchain has their own copy. Basic Scenario alone shows 2 key concepts which are enough to achieve integrity but leaves much to be desired in regards for confidentiality.

Permissioned Scenario scaffolds from Basic Scenario and implements various access control providing confidentiality between different participants on the blockchain. By granting different permissions to different roles within the blockchain, the amount of users who have access to patients' personal data is significantly reduced, which will reduce the risk of data breach. In the previous scenario, it was demonstrated that Hyperledger is a shared ledger but to further increase confidentiality, transactions are hidden on the Composer level if the transaction doesn't affect the participant.

With ***Encryption Scenario***, confidentiality is fully achieved by protecting data outside of the blockchain. Basic Scenario and Permissioned Scenario achieved confidentiality on the blockchain but fails to protect any in-transit data. This scenario creates a bespoke REST API to encrypt and protect data being transmitted between the client and the blockchain. Elliptic-curve Diffie Hellman (ECDH) is used as the key exchange with the public-private key pair, and AES128 is being used as the symmetric encryption method.

2.2 Regulation compliance

Throughout blockchains short lifespan it has been heavily criticised for its lack of regulatory compliance. A key aspect of the developed blockchain was to assess whether blockchain could comply with the GDPR. ***Basic Scenario*** covers the GDPR's right to access. The GDPR states that individuals have the right to access their personal data and within the Basic Scenario, patients are able to access their information quickly and easily. However, Basic Scenario's results fail to exhibit any key aspects of the GDPR or HIPAA that the healthcare industry struggle with. A key aspect of health data regulations is to give control back to the patients which are achieved in ***Permissioned Scenario***. The introduction of different access control rules grants patients the ability to control who has access to their EHR. The GDPR states that individuals must have the right to restrict processing. Allowing patients to control who has access to their data is an alternative to removing data which is a large concern for blockchain. The very idea of data immutability is what makes blockchain infeasible in specific use cases. Personal data should not be kept longer than someone needs it. In this scenario, patients can control how long practitioners have access to their EHR.

Due to data immutability, data can remain accurate on each peer's ledger but fails to comply with the GDPR's right to be forgotten. Unlike HIPAA, the GDPR states individuals have the right to erasure which ***Purging data Scenario*** examines. Hyperledger Composer allows participants to "delete" data. Superficially, it seems that Hyperledger complies with the GDPR and can delete data. However, Hyperledger Composer is simply a higher-level toolset which runs on top of Hyperledger Fabric. Transactions are simply marked as deleted and appear that way in Composer but on the Fabric level, the transaction remains unchanged. If Fabric is the network level; Composer would be the application layer. To cooperate with regulations, Hyperledger-built applications must not store any sensitive data and it is recommended that all personal data should be stored in an off-chain database.

HIPAA and the GDPR enforce consent through, 'authorisation' and 'right to be informed' respectively. Though the GDPR takes it a step further and requires individuals to be notified if there are any changes regarding access or purpose. ***Data type scenario*** demonstrates this right by recording the reason for referral in each transaction. Patients can see why specific medical practitioners access to their EHR have by checking their transaction list. Without this feature implemented, patients who are able to track who has access to their EHR but be ignorant of the purpose behind it.

HIPAA requires safeguards to be put in place in order to protect patients from data leaks. However, the shared ledger displayed in the Basic Scenario potentially poses a problem for the requirement of physical safeguards. Is it possible to apply physical restrictions on each peer? In a full-scale implementation, it would be impossible to ensure that each device was physically secured.

2.3. Hyperledger test analysis – scalability and flexibility

Testing revealed that hyperledger is not very flexible and expects functionality to be carried out the application side. Hyperledger is designed to deal with only text-based data, there is no innate support for images or audio as shown ***in encryption scenario. Data type scenario*** shows that Hyperledger can cope with images but only with some outside interference. Within the scenario, a theoretical application on top of the blockchain converts the image to base64 which can then be stored on the blockchain. Hyperledger does not deal with data in any unique way and is unaffected by base64. Alternatively, an image could be stored in a database and the reference could be stored within Hyperledger but that adds another layer of complexity. Hyperledger's reluctance to support non-text-based data reinforces the notion that Hyperledger wants data to be stored in an external database. Ultimately, storing images in base64 isn't a major issue for healthcare. Base64 does not reduce any quality that physicians may need to see but simply changes the way the data is represented whilst compressing data. Hyperledger

doesn't require the capability to store images or audio directly as they're not being treated any different within the blockchain. Moreover, Hyperledger fails to solve the issue of limited computational resources. As a system scales so do the number of computation resources needed on each peer. For blockchain to be adopted by the healthcare industry energy consumption and computational resources will have to be evaluated.

It's become standard for blockchain platforms to offer some sort of encryption but Hyperledger allows developers to use what encryption methods they see fit. This is extremely beneficial within the industry as it allows hospitals to protect their data with the latest forms of encryption rather than waiting for Hyperledger to release an update. Despite the flexible encryption, Hyperledger offers no chaincode level encryption. It's obvious that Hyperledger expects all sensitive data to be stored off chain which is why they only offer encryption to in-transit data. This direction Hyperledger is taking seems to be an answer from all the criticisms blockchain has been getting regarding data immutability. If data is stored off-chain, then the data immutability of a blockchain becomes less of a significant problem.

4. Discussion and Conclusion

Throughout this report, multiple problems have been identified within healthcare and blockchain has been proposed as a solution. The healthcare industry has been identified as an 'easy' target for cyber-attacks but does blockchain reduce the security risks? As shown in the results, blockchain offers authenticity making it near impervious to impersonation attacks. Further to this, the encryption capabilities displayed by blockchain makes safe from man-in-the-middle attacks. Most importantly, the access control of blockchain restricts the number of people who can view an EHR. As a result, data breaches are less likely.

Does this mean that blockchain solves healthcare's security threat? If sensitive data is stored on the blockchain, the security benefits solve both the high and low-level cyberattacks. However, it's been shown that Hyperledger wants data to be stored off-chain. A blockchain in this form only protects reference data; reducing blockchain to a lookup table which doesn't provide any security on actual personal data. Sensitive data must be stored on the blockchain to benefit the healthcare industry resulting in applications taking more responsibility. The application must convert all data to text; provide script-level encryption and a degree of access control to accommodate for blockchain's weaknesses.

An issue highlighted in this project is blockchain's need for computational resources and energy when data is stored on-chain. Unlike blockchain, the cloud offers ubiquitous resources and is optimised for IoT. Despite this, blockchain still provides a better platform for healthcare as security is a necessity rather than an ideal function. It is essential that the scale of any

blockchain application must be managed; blockchain, in its current iteration, exponentially requires more resources as the size increases. Even though, many of the security features implemented are not exclusive to blockchain. Security is the responsibility of the vendor, allowing for less flexibility within the industry. Blockchain allows developers to easily change their encryption algorithm but this isn't the case for cloud computing. For example, AES 128 is currently believed to be secure but with the emergence of powerful supercomputers, collision will soon be found. Unlike cloud providers, blockchain allows developers to implement a stronger algorithm themselves.

If a blockchain were to be developed for healthcare, there is a trade-off between transparency and confidentiality that the industry should be aware of. Blockchain is intended to increase trust by sharing all data. Yet, access control offers limits the sharing of data and achieves a level of confidentiality. Any implementation must ensure only identifiable data is limited by access control and there is still a degree of transparency on the blockchain.

Blockchain is not often considered when discussing regulation compliance but this report has highlighted various areas where blockchain complies with regulations. The biggest issue stopping GDPR compliance is the inability to remove data. With current legislation, blockchain cannot be implemented effectively with current blockchain platforms. Using a blockchain with a relational database doesn't provide any security on personal data directly. As countries become more aware of the energy consumption of existing blockchains. The chance of legislation changing to accommodate the emergence of blockchain is extremely slim. The trade-off between computational overhead and security cannot be made if blockchain stores sensitive data. Ultimately, blockchain provides security benefits for the healthcare industry and would reduce the number of cyberattacks. Yet blockchain in its current iteration is not suitable for healthcare.

Blockchain presents significant security benefits but an even larger trade-off in the form of overheads and regulation compliance. Legislation takes a significant amount of time to adapt to technology; the data protection act stayed the same between 1997 when it was first proposed to 2018 when it was updated. Alternative technologies, such as cloud computing can completely comply with the GDPR as well as offering a surplus of resources cheaply. Whereas blockchain struggles to comply with the GDPR and requires a large number of resources as the system scales. Even though blockchain currently isn't suitable for the industry, it's only in its 2nd generation with bitcoin being the first generation.

References

1. **Adefala, L. (2018)**. Blog. Retrieved from Fortinet: https://www.fortinet.com/blog/business-and-technology/healthcare-

experiences- twice-the-number-of-cyber-attacks-as-othe.html2. (Tapscott)

2. **A. Dinh**, R. Liu, M. Zhang, G. Chen, B. C. Ooi and J. Wang, (**2018**). Untangling Blockchain: A Data Processing View of Blockchain Systems. IEEE Transactions on Knowledge and Data Engineering.
3. **Gayle, D**., Topping, A., Sample, I., Marsh, S. and Dodd, V., (**2017)**. NHS seeks to recover from global cyber-attack as security concerns resurface. The Guardian, 13.
4. **Kaspars Z.,** R. S. (**2012**). Blockchain Use Cases and Their Feasibility. Applied Computer Systems: The Journal of Riga Technical University, 12-20.
5. **F. Ahmad**, Z. Ahmad, C. A. Kerrache, F. Kurugollu, A. Adnane, E. Barka (2019) "Blockchain in Internet-of-Things: Architecture, Applications and Research Directions", 2019 International Conference on Computer and Information Sciences (ICCIS), 3-4 April 2019, Saudi Arabia
6. N**adezhda F. (2018).** Blockchain – an opportunity for developing new business models. In Business Management / Biznes Upravlenie. 2018 (pp. p75-92).
7. **Omar Alia, A. S. (2018).** Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review. International Journal of Information Management, 146-158.
8. **Peck, M. E. (2017).** Blockchain world - Do you need a blockchain? This chart will tell you if the technology can solve your problem. IEEE Spectrum, vol. 54, no. 10, 38-60.
9. **Sirer, A. E. (2017, Feb**). Miniature World: Measuring and Evaluating Blockchains. Retrieved from Hacking, Distributed: http://hackingdistributed.com/2017/02/10/miniature-world/
10. **S. K. Lo**, X. Xu, Y. K. Chiam and Q. Lu (**2017**). Evaluating Suitability of Applying Blockchain. 2017 22nd International Conference on Engineering of Complex Computer Systems (ICECCS), 158-161.
11. **Smith, T. D. (2017).** The blockchain litmus test. 2017 IEEE International Conference on Big Data (Big Data), 2299-2308.

Data science and the big data framework for development, and to benefit from disruptive technology advances

Fionn Murtagh
University of Huddersfield, UK

Abstract

Methodology is discussed and described here, as also are important data sources, that can be very relevant also for sustainable development. At issue is health and medical analytics, through analytical focus and contextualization, with new challenges and opportunities in the context of Big Data, Geometric Data Analysis: analytics of processes and behaviours. Foundational here is the geometry and topology of data and information for analytics of processes and behaviours. Also important is "homology" (i.e. associations that are determined in integrated data sources) and "field" (i.e. analytical focus) of eminent sociologist, Pierre Bourdieu, and addressing new societal challenges, and new themes and topics, problems and challenges, in medicine and in health and hence also in life sciences.

Keywords

Correspondence Analysis; Geometric Data Analysis; quantitative and qualitative analytics; health and wellbeing; developing economy countries.

1. Introduction

First to be noted is how Big Data can be availed of, to form the context for the patient's treatment.

The first chapter of Zhang et al. (2018), pages 1–6, entitled "Big data and clinical research: perspectives from a clinician" by Zhongheng Zhang, counterposes, as research, interventional analysis, which is, in fact, experimental research, relative to observational studies. An important release of a "data sharing platform for population and health" in China in 2017 is noted as: "historic leap in clinical research".

An important point made is that patients' treatments are usually complicated by the patients comorbidities. So it becomes so very important to have and to use ancillary and contextual observations also. This amounts to the practical setting for big data clinical trials.

Electronic medical records can be very important. Such big data may very well include also, demographic attributes, microbiology information on the patient, and other data sources. So it is noted that such observational data, encompassing what amounts to big data clinical trials, can be very relevant to the real-world, i.e. detailed and comprehensive information on the patient.

A repository, entitled "Medical Information Mart for Intensive Care III", MIMIC-III, with its data on over 40,000 patients is described. Access to that is described. An interesting statement is that the demographics of China will lead to very high quality big data sources in China. Description is provided of an important release at the beginning of 2017 of the "National scientific data sharing platform for population and health (NSDSPPH)" in China, comprising observed and recorded data with 280 million observations or records. This is noted as an "historic leap in clinical research".

2. Bias from Self-Selection of Behavioural Data

Keiding and Louis (2016), this most comprehensive survey (118 citations) sets out new contemporary issues of sampling and population distribution estimation. An interesting conclusion is the following. "There is the potential for big data to evaluate or calibrate survey findings ... to help to validate cohort studies". Examples are discussed of "how data ... tracks well with the official", far larger, repository or holdings.

Association with such data calibration, and following also the need to integrate data sources, is the importance of bridging and shared patterns and associations in the data. Hence, this is to benefit from the methodology of eminent social scientist, Pierre Bourdieu.

In Keiding and Louis (2016), it is well pointed out how one case study discussed "shows the value of using big data to conduct research on surveys (as distinct from survey research)". Limitations though are clear: "Although randomization in some form is very beneficial, it is by no means a panacea. Trial participants are commonly very different from the external ... pool, in part because of self-selection". This is due to, "One type of selection bias is self-selection (which is our focus)". Important points towards addressing these contemporary issues include the following. "When informing policy, inference to identified reference populations is key". This is part of the bridge which is needed, between data analytics technology and deployment of outcomes.

"In all situations, modelling is needed to accommodate non-response, dropouts and other forms of missing data". While "Representativity should be avoided", here is an essential way to address in a fundamental way, what we need to address: "Assessment of external validity, i.e. generalization to the population from which the study subjects originated or to other populations, will in principle proceed via formulation of abstract laws of nature similar to physical laws".

In Lebaron (2009), Bourdieu's analytics "amounted to the global (hence Big Data) effects of a complex structure of interrelationships, which is not reducible to the combination of the multiple (... effects) of independent variables". The concept of field, here, uses Geometric Data Analysis that is core

to the integrated data and methodology approach used in the Correspondence Analysis platform (Murtagh, 2017).

3. Analytical Focus and Contextualization

In this preliminary study of mental health, see Murtagh and Farid (2017), the following is described: Choice and selection of main and supplementary variables. Therefore: our main focus of analysis, and an explanatory context.

In this data source, "Adult Psychiatric Morbidity Survey, England, 2007", this reference is important, "HSCIC, Health and Social Care Information Centre (National Health Service, UK)", 2009. "National Statistics Adult Psychiatric Morbidity in England" – 2007, "Results of a household survey, Appendices and Glossary". 174 pp. Available at:
http://www.hscic.gov.uk/pubs/psychiatricmorbidity07

There are 1704 variables, including questioning of the subjects about symptoms and disorders, psychoses and depression characteristics, anti-social behaviours, eating characteristics and alcohol consumption, drug use, and sociodemographics, including gender, age, educational level, marital status, employment status, and region lived in.

An initial display of the neurotic symptoms and common mental disorders sought to have socio-demographic variables as supplementary. But these were projected close to the origin, therefore showing very little differentiation or explanatory relevance for the symptoms and disorders data.

It is found that factor 1 is explained as PTSDcom, "Trauma screening questionnaire total score" versus all other variables. Factor 2 is explained as "CISR-FOUR" versus "nosymp". These are, respectively, "CIS-R score in four groups, 0-5, 6-11, 12-17, 18 and over. (CIS-R = Common Mental Disorders questionnaire)"; and having no neurotic symptoms in the past week.

It was sought to characterize the socio-demographic data, and then to see if the neurotic symptoms and common mental disorders data could be explanatory and contextual for the socio-demographic data. But no differentiation was found for these supplementary variables, indicating no particular explanatory capability in this particular instance. It may be just noted how the main and supplementary variables were interchanged. Respectively, the symptoms and demographic variables were main and supplementary; then the main and supplementary variables were the demographic variables and symptoms. This was done in order to explore the data. It was seen to have age and education level counterposed to home region. It was also seen to have educational level coun-terposed to: employment status, gender, marital status, ethnicity.

Finally, it was checked whether neurotic symptoms and common mental disorders data should be jointly analysed with the socio-demographic data.

One summary interpretation is how factor 1 accounts for recorded trauma, and factor 2 accounts for region of the respondent.

The second analysis was to characterize the socio-demographic data, and then to see if the neurotic symptoms and common mental disorders data could be explanatory and contextualized. In the third analysis: It was checked whether neurotic symptoms and common mental disorders data should be jointly analysed with the socio-demographic data.

To be noted here, is how Big Data inputs are required to calibrate and validate, and Open Data sources are key.

In the Adult Psychiatric Morbidity in England, household survey, covered was: Common mental disorders; Posttraumatic stress disorder; Suicidal thoughts, attempts and self-harm; Psychosis; Antisocial and borderline personality disorders; Attention deficit hyperactivity disorder; Eating disorder; Alcohol misuse and dependency; Drug use and dependency; Problem gambling; Psychiatric comorbidity,

4. Health and Medical Data Sources for Developing Countries

In the "Atlas of the African Health Statistics" (WHO, 2017, see "Publications"), a 137 page document in 2017 from the African Health Observatory, World Health Organization (WHO), Regional Office for Africa, there are many comparative statistical evaluations. With data from World Health Organization, and from UNICEF, and with lots of coverage of morbidity and children, there is adolescent health coverage, and communicable diseases like HIV, and coverage of malaria, tuberculosis, hepatitis, and many other themes, including mental health, non-communicable diseases, accidents, etc. There is also: health financing, health workforce, and in the chapter entitled "Social determinants of health", there are sections on "Water and sanitation", and "Access to electricity".

From the African Development Bank Group (https://www.afdb.org/en), it is very clear how economic development has to be based on, and linked to, health and lifestyle, energy and environment.

Mathematics underpins, and is the basis for, all of Data Science and Big Data analytics, see Murtagh (2017). Here too, multidisciplinarity is essential, following the integration of data sources and of methodology. There will remain many research issues for the multiple source data integration, where there will be missing data and data with uncertainty, and the relevance of qualita¬tive and quantitative data encoding. Data curation is a very important current research challenge, see Murtagh and Devlin (2018), and also important disrup-tive technological advances, especially Internet of Things, Smart Cities, Smart Homes, these all provide important data sources, to be encoded, integrated and with deployment of optimal methodologies. Such will be playing a role in the developments and innovation in this project.

In general, and especially in regard to disruptive technological advances, e.g. Internet of Things and Smart Cities, our leading research will encompass the following. The role of ontologies is very central in qualitative analysis of research, cf. Murtagh et al. (2018). Context is so very important in Big Data analytics and in many domains, Murtagh and Farid (2017). In Murtagh and Farid (2017), it is described how analytical focus and ancillary and contextual information sources are to be well associated and/or well combined. Ultrametric regression, in Murtagh et al. (2011), the regression is based on the hierarchical structure of the predictor variables, for predicting the outcome or dependent variable.

We will list many sources of open data, an important aspect of development work is to have access to data sources, with good quality data curation, adhering to open data standards when appropriate relative to data rights and security, which will be fully taken into consideration.

5. Conclusion

Following also Allin and Hand (2017), here at issue has included: Qualitative and quantitative observing and monitoring of wellbeing: New statistical drivers, Big Data analytics, Open Data, geometry and topology of data and information, semantics, homology and field; Geometric Data Analysis and the Correspondence Analysis platform.

In Allin and Hand (2017), there is discussion of data sources for national health services, and the importance of Big Data to address bias of self-selected, social media or other data sources. Having Big Data to contextualize statistical analysis is at issue in Keiding and Louis (2016).

Open data sources are implying the essential need for integration of data sources; and in the future, from disruptive technology advances, such as Internet of Things (IoT), smart cities, etc. Other important current work is towards: health and medical management and policy making, to be based on association with, or integration with, many open data sources, and other data sources.

References

1. Allin, P. and Hand, D. (2017), "New statistics for old? – Measuring the wellbeing of the UK", *Journal of the Royal Statistical Society, Series A*, 180(1), 3–43, Including F. Murtagh comments.
2. Keiding, N. and Louis, T.A. (2016), "Perils and potentials of self-selected entry to epidemiological studies and surveys", *Journal of the Royal Statis¬tical Society, Series A*, 179 (2), 319–376. Including F. Murtagh comments.

3. Lebaron, F. (2009), "How Bourdieu 'quantified' Bourdieu: the geometric modelling of data", chapter 2 in K. Robson and C. Sanders, Eds., *Quantifying Theory: Pierre Bourdieu, Springer.*
4. Murtagh, F. (2017), *Data Science Foundations: Geometry and Topol¬ogy of Complex Hierarchic Systems and Big Data Analytics*, Chapman & Hall/CRC Press, Boca Raton Florida, USA.
5. Murtagh, F. and Devlin, K. (2018), "The Development of Data Science: Implications for Education, Employment, Research, and the Data Revolution for Sustainable Development", *Big Data and Cognitive Computing*, 2(2), 16 pp.
6. Murtagh, F., Orlov, M. and Mirkin, B. (2018), "Qualitative judgement of research impact: Domain taxonomy as a fundamental framework for judgement of the quality of research", *Journal of Classification*, 35(1), 5–28.
7. Murtagh, F. and Farid, M. (2017), "Contextualizing Geometric Data Analysis and related data analytics: A virtual microscope for Big Data analytics", *Journal of Interdisciplinary Methodologies and Issues in Science,* Special Issue on Digital Contextualization, Vol. 3, 19 pp.
8. Murtagh, F., Spagat, M. and Restrepo, J.A. (2011), "Ultrametric Wavelet Regression of Multivariate Time Series: Application to Colombian Conflict Analysis", *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, 41, 254–263.
9. WHO (2017), World Health Organisation, "Atlas of the African Health Statistics", http://www.aho.afro.who.int https://www.who.int
10. Zhongheng Zhang, Murtagh, F., Van Poucke, S. editors (2018), *Big Data Clinical Study and Its Implementation with R*, AME Publishing Company, Hong Kong.

Diet4You: A personal intelligent assistant for diets integrating data science

Karina Gibert[1,2,4], Beatriz Sevilla-Villanueva[1,2,4], Miquel Sànchez-Marrè[1,3,4]

[1] Knowledge Engineering and Machine Learning Group at Intelligent Data Science and Artificial Intelligence Research Center (KEMLG-at-IDEAI)

[2] Statistics and Operations Research Department

[3] Computer Science department

[4] Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona (Catalonia)

Abstract

Prevention and nutrition are key issues to guarantee a healthy lifestyle and are in the kernel of the new paradigm of patient-centered medicine. More and more, making right diets is becoming essential from the health point of view. However, nutritionists design diets on the specific needs of each patient, using their accumulated experience and there are no well-formalized support mechanisms for such activity.

The project Diet4You (financed by Spanish government) proposes the creation of an intelligent decision support System oriented to the adaptive and dynamic preparation of personalized diets for the specific individuals in general population, having, or not, one or more diseases. The diets are built by taking into account all the information available on the person, including the characteristics of the person itself, their health conditions, their habits and eventual drugs intake and their genòmic information. Authors are not aware of other Systems considering all these factors together for the menus recommendations. The Diet4You tool is a hybrid system interacting several components in a complex way. The paper shows how a data science approach helps to establish a reference set of prototypical dietary paterns as a starting point to establish the behaviour of the complete Diet4you System. Also, a CBR component, based on Knn with some especial distances helps to provide adaptive behaviour to the System, which includes as well, some knowledge based modules.

Keywords

Personalized Recommendation; Nutritional Plan Prescription; Case-based reasoning; Knowledge management; Contextual information; Healthy lifestyles; Personalized Medicine; Data science.

1. Introduction

Prevention and nutrition are key issues to guarantee a healthy lifestyle and are in the kernel of the new paradigm of patient-centered medicine. A healthy diet protects against a large number of chronic diseases, but also contributes to delay the disease appearance in the general population. More and more, making right diets is becoming essential from the health point of view. However, nutritionists design diets on the specific needs of each patient, using their accumulated experience and there are no well-formalized support mechanisms for such activity.

The project Diet4You (financed by Spanish government) proposes the creation of an intelligent decision support System oriented to the adaptive and dynamic preparation of personalized diets for the specific individuals in general population, having, or not, one or more diseases. The diets are built by taking into account all the information available on the person, including the characteristics of the person itself, their health conditions, their habits and eventual drugs intake and their genòmic information. Authors are not aware of other Systems considering all these factors together for the menus recommendations. The Diet4You tool is a hybrid system interacting several components in a complex way. One of those components is a subsystem able to identify dietary patterns of the persons by using multiview clustering for heterogeneous variables, connected with some automatic class interpretation tools that provides a clear description of the patterns to nutritionists, who doesn't necessarily have technical skills. The paper shows how a data science approach helps to establish a reference set of prototypical dietary paterns as a starting point to establish the behaviour of the complete Diet4you System. Also, a CBR component, based on Knn with som especial distances helps to provide adaptive behaviour to the System, which includes as well, some knowledge based modules.

2. Methodology and structure of the system

2.1 Structure of the system

The Diet4You system is composed of two main blocks (see Figure 1):

1. A Nutritional Plan Generator (NPG). This part of the system is designed to return a nutritional plan given personal specifications. The NPG receives the following pieces of information as input:

- Dietary profiles that can be prescribed to certain types of persons with certain genetic characteristics and lifestyle habits, and the pattern of expected diet effects in those scenarios.
- Expert knowledge including Daily Reference Intakes (DRIs) of micronutrients and trace elements; the Recommended Daily Allowance (RDA) of macronutrients; and nutritional and institutional recommendations, such as those provided by the OMS.

- History of Nutritional Plans, which contains a case base made of all the personalized nutritional plans generated by the system in past experiences and feed the CBR that builds the nutritional plans. The nutritional plan may be accompanied by the nutritionist's evaluation of the outcome observed in past applications of the plan
- Characteristics of the person who wants to follow a nutritional plan. This means standard information of the health status of the person including biometrics and biochemical characteristics. Also, drugs intakes, medical history including current, past, and risk factors. Context data such as socio-demographic and habits such as tobacco, alcohol, physical exercise and diet. It can include genetic information when available.

With all these inputs the NPG provides a recommended nutritional plan for a given person to be followed along a certain period.

2. Personalized Menu Planner (PMP). Given a nutritional plan, either coming from the previous subsystem or directly provided by a nutritionist, this subsystem search on food databases for the menus better combining to fit the target nutritional plan. The result is a personalized menu for a given period of time. In Diet4You project, this system is also based on CBR. The system retrieves and adapts dishes to be included in the menu until it is considered that the plan is satisfied. This subsystem includes several components as well:

- Cultural styles: Is a knowledge-based component that can manage the different cultural eating styles regarding how meals distribute along day (Mediterranean, British, ...)
- Restrictions: Is a knowledge-based component that permits nutritionists to express nutritional restrictions (like no sugars for diabetic patients), specific allergies of the person that cannot be included in the proposal of dishes of the recommended nutritional plan
- Preferences: Including user limitations based on personal criteria that will be taken into account as far as possible, trying to avoid proposed menus containing the ingredients excluded by the user.

Thus, the Diet4You system is composed of two subsystems NPG and PMP both combined to give the person guidelines on how to eat for a certain period of time to achieve certain given goals (reduce cholesterol, feed better, improve healthy lifestyle, etc.).

2.2. Formalization of the Nutritional Plan

A nutritional plan contains the specification of the proportions of families of foods to be taken during a certain period of time. A nutritional plan is defined as the triplet $\mathcal{N} = < F, T, Q >$ where,

- F = (π_1, ..., π_N) is a vector containing the recommended balance of food families to be taken along a certain period of time.
- $\pi_{f, f\epsilon\{1:N\}}$ is the proportion of food family F_f recommended, such that $(\sum_{f=1:N} \pi_f = 1)$, given a set of N foods families $\mathcal{F} = \{F_1, \dots, F_N\}$ (i.e., fruits, proteins, etc.)
- $T = [t_0, t_f]$is the period of time where the diet must be followed
- Q indicates the total quantity of food in Kcal to be intaken along the whole diet.

The quantity Q is needed due to the fact that food proportions are different for children, young or elderly people

2.3. Hard Nutritional Restrictions Management

The Personalized Menu Planner must deal with additional restrictions to be satisfied by the final recommendation. According to health conditions and medical reasons, nutritionists may impose some restrictions in the design of the menu and Diet4you must preclude or force the proposal of restricted food families in the menus of a certain person. For example, for diabetic patients, sugars will be restricted, or refined flours might be mandatory, for persons with cholesterol, saturated fats might be mandatory. In addition, known allergies of the person can be managed as well by restricting the corresponding foods (nuts, strawberries, fish...). Finally, strict cultural features of the person, like avoiding meat for a vegetarian person, avoiding pork meat for a Muslim, can be managed as well under the mechanism of hard restrictions. Diet4You guarantees that the proposed menu will not include a single dish containing restricted ingredients or food families. This strategy is implemented by filtering the food database according to restriction and using the food and dishes ontology and building the personalized menu over the filtered database.

2.4. User Preferences

Diet4you can manage the user preferences as well, in an attempt to increase adherence to diet. A user preference expresses that a person likes or dislikes some kind of food. If a person dislikes fish or chicken, he can declare in Preferences section and the system minimizes the probability of including dishes that contain unpreferred ingredients for the final menu and maximizes probability of using preferred ingredients. In this case there is no guarantee that the disliked food is completely eliminated from the proposed menu, and the nutritional prescription can force the person to accept some specific food

in his menu, even in he dislikes, when there is no other alternative. This will be internally managed by introducing penalization or bonification of associated dishes.

2.5. Cultural Eating Styles

There is a cultural factor in the nutrition habits of a person according to where the person lives. For instance, in a Mediterranean country, the breakfast often is more caloric and less proteic, the lunch concentrates more proteins, and dinner can concentrate more vegetables or fruits. In the Anglo-Saxon-style breakfast tends to be more proteic, lunch is light with vegetables and not much calories, and dinner more proteic. Diet4you can manage this contextual knowledge in the composition of menus by getting Diet-Styles in form of tables with probability distribution of food or nutrients families conditioned to a variable number of meals per day. From this information the general nutritional plan is divided into sub-plans for each specified meal (breakfast, lunch and dinner for example). The personal menu planer is building local menus for each meal and guarantees that the whole resulting menu fits the global nutritional plan originally prescribed.

2.6. Personal Menu Planner

The personal menu planner (PMP) is mainly implemented following the cycle of Case-Based Reasoning. Given a nutritional plan for a certain individual i: vi = <Fi, Ti, Qi>, and considering that the Fi vector contains the N families of food resulting from a certain level of granularity determined in the reference food ontology:

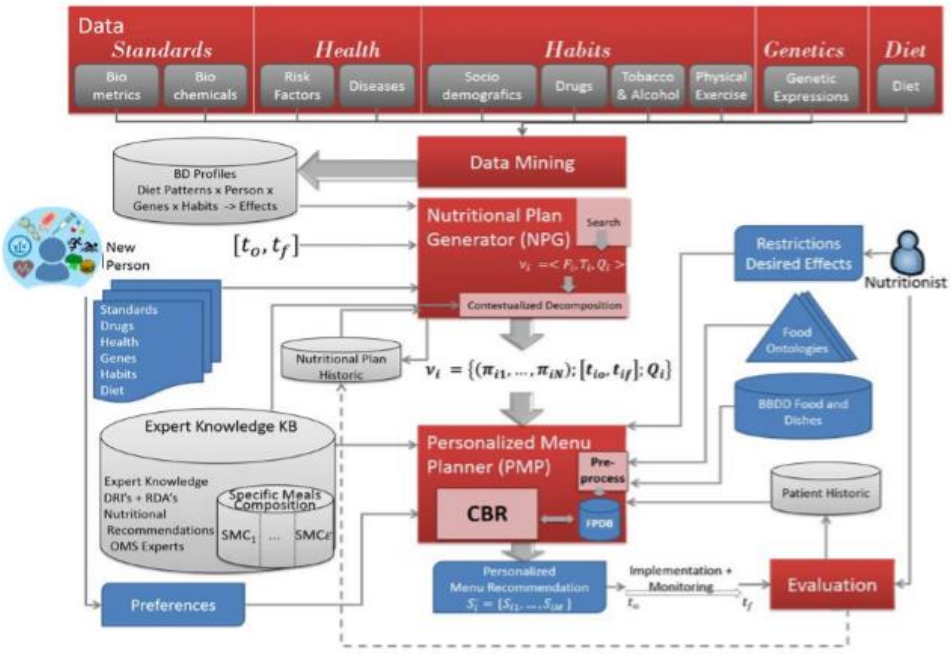
1. *Pre-processing step*. Pre-process the DB in order that all food families have equivalent units in Kcal and build the transformed data base Food Proportions Data Base (FPDB). The FPBD contains either prepared dishes or simple foods d with
   - $p_{\boldsymbol{d}} = (p_{d1}, \dots, p_{dN}), p_{df}$ being the proportion of food family F_f contained in one standard portion of dish $d(f = 1{:}N)$.
   - $q_{\boldsymbol{d}}$ is the quantity associated to one standard portion of dish d, in grams or cups or the corresponding measurement unit.
2. *Retrieval step.* p_d is a vector of proportions and thus, it is directly comparable with F_i. The Euclidean distance is suitable to compare composition of two dishes through their p_d. FPDB is used as the case base to identify candidate dishes with p_d close to F_i prescribed in the targeted nutritional plan. Sort the elements in FPBD into $FPDB^* = \{d_{(k)} | d_{(k)} \boldsymbol{\epsilon}\, FPDB \text{ and } d(p_{d(k)}, F_i) \leq d(p_{d(k+1)}, F_i)$. Candidates will be in the first positions of $FPDB^*$ and recommended menu is composed by using iCG strategy (iterative Candidates Generation) validated in previous works [CCIA 2017, CCIA 2018]. At each iteration, the candidate

solutions $d \in FPDB^*$ are sequentially included in the menu, and the corresponding q_d are subtracted from the Q_i associated to the nutritional plan. The plan is satisfied when card(S) is such that the $Q_i - \sum_{\forall d \in S} q_d$ is minimum.

3. *Reuse step*. Implement the reuse step of CBR by adapting the candidate solutions contained in $FPDB^*$ to be presented to the end-user in form of complete menus for the period T_i. Given the candidate solutions $d \in FPDB^*$, in general $q_d < Q_i$. Adaptation will guarantee $\sum_{\forall d \in S} q_d = Q_i$, by reducing or increasing the portion of last dishes in the menu. Currently complete menus for T_i are presented in a Word document that can be afterwards customized by the nutritionist. The resulting document contains the list of dishes organized by days for the whole T_i period, and sudvided by meals inside each day, according to the Cultural style indicated. Additional information about the portions (in gr) and the nutritional value of the meals is provided at different levels of granularity.

3. Discussion and Conclusions

In this work, we have presented a global view of the Diet4you system including the possibility of defining a nutritional plan at different levels of granularity, managing personal food preferences of the person, as well as hard restrictions able to manage allergies or health constraint according to medical reasons (diabetes, cholesterol....), and the cultural dietary styles to manage distribution of meals along the day by slicing nutritional prescription in sub-plans, which are locally optimised by the PMP through conditional distribution of nutrients fiven the meal of the day. While hard restrictions eliminate dishes and ingredients from the search, the PMP component applies penalties or bonifications for the probability of selecting or non-selecting preferences or unpreferences declared by the person, such biasing the proposed final menu towards personal preferences.



Specific strategies are used to avoid repetitions of same food in the same day. Also, the system provides the possibility to load and save all specifications of a certain prescription, in a personalized prescription including cultural context, restrictions and preferences, but also in guideline element like families of restricted foods for diabetics, guidelines of organization of meals in Mediterranean dietary style, guidelines of nutritional prescription for standard type of person (vegan, with diabetis, and hypertension...) providing a unique flexibility that can't be found in other menu planners.

Results are presented in an editable Word file with the proposed menu and additional nutritional information of the proposal where nutritionist can add extra information or customize at his best convenience before delivering to the person. The resulting document permits a wide scope usage of the system, not only for patients of chronic diseases, or acute post-chirurgical intervention, but for general population as well where healthy life styles can included balanced diets following international recommendations. Evaluation with the nutritionists, show good results, well adapted to the different personalization items considered into the system (nutritional constraints, preferences and cultural context). Specific results of this evaluation are detailed in some previous works and are not included here due to extension limitations [5][6][7][8][9][10].

Such, Diet4you is a well example of a real system assisting into a health aspect related to the new paradigm of personalized medicine and healthy lifeStyles which combines in an integral way Artificial intelligence with statistics in an intensive Data Science process embedded into an Intelligent Decision

Support system. Clustering, basic statistics, profiling analysis based on inferencial tests, conditional distributions management, probabilized distance based methods and statistical fitting are combined with ontology management, prior expert knowledge, automatic interpretation and case based reasoning to address the complexity of personalized diets recommendations.

To the best of our knowledge the current Diet4You systems is the first system integrating all these personalization items together in the recommendation of menus according to nutritional prescriptions which, in turn come from standard types of diets observed in population and tipified through clustering techniques by the Nutritional Plan Generator component. Currently the automatic generation of the Nutritional prescription according to the results of NPG component is being addressed and specific metrics to measure the nutritional quality of the recommendation regarding nutritional prescription are being developed.

Acknowledgement

This work has been partially supported by project Diet4You (TIN2014-60557-R), and the Consolidated Research Group Grant from AGAUR (Generalitat de Catalunya, cataalan government) IDEAI-UPC (AGAUR SGR2017-574).

References

1. The Automatic Meal Planner-Eat this much homepage, http://www.eatthismuch.com, last accessed February 2019.
2. DRI Calculator for Healthcare Professionals homepage. https://fnic.nal.usda.gov/fnic/dri-calculator/, February 2019.
3. S.A Bowman, J.C Clemens, et al. Food patterns equivalents database 2011-12: Methodology and user guide, 2014.
4. National Geographic. What the world eats. Accessed: May 14, 2018.
5. Gibert, Karina, Beatriz Sevilla-Villanueva, and Miquel Sànchez-Marrè. "The role of significance tests in consistent interpretation of nested partitions." *Journal of computational and applied mathematics* 292 (2016): 623-633.
6. B. Sevilla-Villanueva, K. Gibert, and M. Sànchez-Marrè. Generating complete menus from nutritional prescriptions by using advanced cbr and real food databases. In Recent advances in AI research and development, v 300: 166–175. IOSPress, Jan 2017.
7. B. Sevilla-Villanueva, K. Gibert, M. Sànchez-Marrè. Intelligent Management of measurement units equivalences in food databases. In procs CAEPIA 2018. LNAI 11160: 1-11. Springer, Amsterdam. Oct 2018.
8. Sevilla-Villanueva, K. Gibert, Sànchez-Marrè. Including hard restrictions into Diet4You Menu Planner. *Artificial Intelli-gence Research and*

Development: Current Challenges, New Trends and Applications 308 (2018): 190. IOSpress 2018.

9. K. Gibert, A. Valls, and M. Batet. Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. Knowledge and Information Systems, 40(3):559–593, 2014.

A survey of Machine Learning Algorithms for efficient biomarkers identification

Mahmoud Rafea[1], Passant Elkafrawy[2]*, Mohammed M. Nasef[2], Rasha Elnemr[1], Amani Tariq Jamal[3]*

[1] Central Lab of Agriculture Expert Systems, Giza, Egypt

[2] Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Shibin El Kom, Egypt,

[3] Computer Science Department, Faculty of Computing and InformationTechnology, Jeddah University, Jeddah, Saudi Arabia

Abstract

These days, health care industry generates a large amount of complex data about patients. This increase in data volume requires ways in which data can be extracted and processed efficiently. Machine learning algorithms could play an efficient role in the field of disease diagnosis and biomarker discovery. Machine learning attempts to tell how to automatically find a good predictor based on past experiences. The recent researchers in machine learning promise the improved accuracy of perception and diagnosis of disease. The accurate diagnosis of some serious diseases is a very crucial task in medical science. This paper discusses some of the machine learning algorithms in biomedical field.

Keywords

Biomarkers identification, Machine learning, Big data Analytics for disease prognosis

1. Introduction

Big data turns into Chunks due to multidisciplinary combined effort of machine learning (ML), databases and statistics. Today, in medical sciences disease diagnostic test is a serious task. It is very important to understand the exact diagnosis of patients by clinical examination and assessment. For effective diagnosis and cost effective management, decision support systems that are based upon computer may play a vital role. Health care field generates big data about clinical assessment, report regarding patient, cure, follow-ups, medication etc. However, this data is acquired from long periods of data collection, and clinical expertise for implementing the best diagnostic test and interpreting test results based on patient's history [1].

A large number of different systems for disease diagnosis were proposed in the early days as in [2, 3]. Then researchers developed their own representation methods guided by thoughts on how to handle a particular medical problem. The development of these early systems gave rise to the phrase knowledge-based system, or knowledge system, which is generally

employed to denote information systems in which some symbolic representation of human knowledge of a domain is applied, usually in a way resembling human reasoning, to solve actual problems in the domain. As this knowledge is often derived from experts in a particular field, and early knowledge-based systems were actually developed in close collaboration with experts. However, the new direction is extracting the knowledge and storing it to facilitate the decision. Rules are the most appropriate knowledge in medical problems. Thus, association rule mining algorithms play a vital role in solving these problems especially in disease diagnosis.

Some types of diseases are difficult to detect at early stage due to the lack of symptoms. Early detection of serious diseases is essential in reducing life losses. Earlier treatment, however, requires the ability to detect these diseases in early stages. Early diagnosis requires accurate and reliable diagnosis procedure. Automatic diagnosis is considered as a real world medical problem. Therefore, finding an accurate and effective diagnosis method is very important [4]. Thus, the new direction in disease diagnosis is based on proteomics. This is by using machine learning on proteomics data to extract biomarkers, and to identify the serious diseases.

The Association Rule Mining on Medical Applications is described in section 2. The Biomarkers Discovery is described in section 3. Applying machine learning algorithms on EDAS are described in section 4. Conclusion and future work are described in section 5.

2. Association Rule Mining on Medical Applications

Multiple papers reviewed the solution of medical problems as Association Rules. In [5], the researchers proposed a data mining technique based on Apriori Algorithm for generating the frequency of diseases that affect patients in the various geographical region and at various time periods. The analyses concluded that patients are affected by 4 different diseases in a particular geographical area during a particular year.

In [6], the researchers presented an association rule mining for medical data to anticipate heart diseases using Apriori algorithm. They used medical data which the diseased and healthy patient's details are categorized for the prediction of heart diseases.

In [7], it was focused on the implementation of the Apriori Algorithm to discover interesting patterns and association rules in chronic diseases. Percentage of possibility for chronic disease was calculated from each symptom of all considered chronic diseases. The higher number of symptoms lead to higher accuracy of calculating the disease possibility.

However, in [8], the researchers selected the topic of NED (No Evidence of Disease) and ED (Evidence of Disease) for the Breast Cancer problem. They experimented two association rule mining algorithms; Apriori and FP-Growth. They attempted to detect the relationships between different factors. Their

dataset shows the human hormones (ER, PR, and HER-2), the metastases of cancer (liver, brain, bone, and ovary), and treatment to the patient (chemotherapy and radiotherapy). They categorized NED and ED to detect the relationship between the different factors mentioned above. This paper is useful for knowing the main important factor which influences the patient.

In [9], it was provided a computational study, based on the Apriori algorithm to discover the associations among clinical traits and risk factors of asthma disease. The experiment was done on the original dataset collected from the asthma patients. They identified association among four attributes a cough, wheezing, running nose and stuffy nose. The Apriori algorithm was used to find the frequent symptoms and related causes of asthma disease from the dataset that was collected from the self-reported asthma patients. The remaining attributes like breathing shortness, dust allergy, skin allergy, fruit allergy and allergy to air conditioner may be considered for further analysis. Additional features such as the nature of foods, living environment, working environment, stress, and other related diseases may be considered.

3. Biomarkers Discovery

a. Proteomics

Proteomics-science identifies and characterizes protein expression in biological systems. Due to the limitations of studying DNA and RNA alone, Proteomics has been gaining full energy and trust. Gene sequences also can give little information about how much of its transcribed protein will be expressed and in what cellular states. By the usage of proteomics, studying gene expression at the protein level can achieve complementary knowledge at the nucleic acid level. Proteins are more diverse than DNA or RNA and therefore carry more information than nucleic acids, since alternative splicing and more than 100 unique post-translational modifications result in tens (and possibly hundreds) of species of protein from each gene [10].

Clinical proteomics is the application of proteomic techniques to the field of medicine with the aim of solving a specific clinical problem. The study of clinical proteomic may provide us with opportunities in more effective strategies for early disease detection and monitoring, more effective therapies, and developing a better understanding of disease pathogenesis [11]. Such studies may aim at earlier or more accurate diagnosis, improvement of therapeutic strategies, and better evaluation of prognosis and/or prevention of the disease. Although clinical proteomics currently mainly focuses on diagnostics and biomarker discovery, it includes the identification of new therapeutic targets, drugs, and vaccines for better therapeutic outcomes and successful disease prevention. [12].

The application of clinical proteomic research is growing rapidly in the field of biomarker discovery, especially in the area of cancer diagnostics. Clinical proteomics holds the potential of taking a snapshot of the total protein

complement of a cell, or body fluid, and identifying proteins as potential biomarkers for the differentiation of disease and health.

Clinical proteomics is an extremely large field consisting of a different collection of platforms. Mass spectrometry (MS) technology is an essential device in these platforms. MS has a powerful use for protein identification and profiling experiments [13, 14, 15, 16, 17]. The goal of protein identification is to find out the amino acid sequence of extracted proteins; whereas the goal of protein profiling is to quantify the expression level of proteins.

b. Biomarkers Discoveryfrom EDAS

New clinical tests are conducted to identify novel biomarkers from MS based proteomics. The Mass spectrometry has a powerful role in proteomics [18]. Biomarker discovery which uses MS techniques necessitate sensitivity, mass accuracy, and reproducibility. Construction of a comprehensive biomarker pipeline is a trend supported by the advances in technology, based on five essential process components: candidate discovery, quantification, verification, research assay optimization, and biomarker validation [19].

Biomarkers discovery is depending on the comparison of different physiological states, phenotypes done during controlling (diseased) patient groups. The discovery of biomarkers adopts the idea of considering the molecular species such as genes, and proteins possible biomarkers as these species can show the changes that are done within phenotypes [20, 21].

There are many definitions of biomarker [22, 23, 24]. Meanwhile, we will state the definition of the National Cancer Institute which defines the biomarker as "a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease" [25]. A biomarker is also called molecular marker or signature molecule.

There is over a thousand single candidate cancer biomarkers have been known for several years [26]. However, the US Food and Drug Administration (FDA) approved that none of these is routinely used for early clinical diagnosis, except a few of them for example, CA125 (also known as mucin 16) for ovarian cancer, prostate-specific antigen (PSA) for prostate cancer and CA19 9 for pancreatic cancer have been proposed to be useful for longitudinal disease monitoring [27, 28, 29].

Biomarkers have the ability to indicate the different changes that can be happened into organs and cells. These changes could be connected with specific disease causes or with the different implications of normal and pathological cases. So biomarkers are considered very necessary in prediction and monitoring the molecular variations related to the development of diseases and their reactions. Biomarkers can play a very important role into representing therapeutic objectives. For example their important role into disease cases is very obvious. They can provide more clearance into the

biological systems understanding regarding the domain of health and disease. Biomarkers are considered a great director in the development of new treatment strategies [30].

One of the most important applications of specific biomarkers is to find the tumor at an early stage even before clinical symptoms are developed. Patients with cancerous disease can be treated efficiently when detected in early stages [31]. This would certainly increase overall survival. The World Health Organization (WHO) proposed that "millions of cancer patients could be saved from premature death if early detection and treatment were available" [32].

In addition to early diagnosis, evidence-based medicine can profit from biomarkers knowledge providing; selection of the optimal therapy and improving prognosis of diseases. The greatest potential for enabling biomarkers for cancer lies in improving the technology for protein biomarker discovery. Protein biomarkers should be found in a minimally invasive liquid biopsy such as a simple blood sample. Nevertheless, does blood contain enough information? Numerous researchers in this era work to find protein cancer biomarkers of clinical utility [33, 34, 35, 36].

In [37] they conclude that the antigens exist in the RBC cytoplasm have inversely proportional to immune tolerance. When the antigens in RBC increases the antibodies in plasma decreases and vice versa, where Ag represents the antigens in RBC, and Ab represents antibodies in plasma. Also, they conclude that RBC has a dynamic store of: body antigens (Tissue Specific Antigens (TSA)), food antigens, environment antigens, bacterial commensal antigens, and disease antigens whether microbial, viral, or tumors. This store is known as: **Erythrocytes Dynamic Antigens Store (EDAS).**

4. Applying Machine learning algorithms on EDAS

In order to make maximum benefit from this discovery, computer processing capability and computer knowledge processing capability can help to profit this discovery. To this endeavor, a random generation of EDAS was described in [38]. Meanwhile, the generation of EDAS model was very simple and did not reflect the real EDAS. It was based on classifying proteins into normal and abnormal, only, without specifying the nature of these proteins. Also, in [37, 38] they proposed a technique to discover biomarkers of diseases based on EDAS. However, these algorithms are very simple, and do not match with reality EDAS. Also, they did not show which disease or set of diseases can be applied on? They used one category of diseases. And, they do not make any experiment to verify the model.

Recently, in [4], researchers developed the random generation of EDAS. This developing is based on proposing a new mathematical model to abstract the problem computationally with richer knowledge. This new random generation of EDAS data simulate reality. The generated EDAS data consists

of normal proteins and disease proteins. The normal proteins consist of Environment proteins, Food proteins, Commensal proteins, Tissue proteins. The disease proteins are malignant proteins or pathogenic proteins. Then, they proposed machine learning algorithms to analyze EDAS data. The aims of such investigation have been to identify the minimum set of proteins that can be used as biomarkers for a particular disease (Pathogens, and Malignancies). This work moves from identifying single biomarker to discovering multiple biomarkers foe each disease separately. Lastly, they represent their results (detected biomarkers) in the form of rules. Based on these rules they proposed a diagnostic model. This diagnostic model can diagnose any new case (new EDAS) and determine if this case has a specific disease or not. If this new case is diseased, the model can predict the ratio of the biomarkers at this case.

5. Conclusion and future work

This survey provides the brief description of machine learning techniques for disease diagnosis and discovering biomarkers. The mentioned related work concentrated on discovering the relationship between diseases and their symptoms. However, no one of them try to detect the relationship between the normal proteins and the diseased proteins. In the future, it is vital to predict the relations between the diseases proteins and environmental actors present in RBC.

References

1. Fatima, M. and Pasha, M. (2017) Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, 9, 1-16. https://doi.org/10.4236/jilsa.2017.91001
2. Pople, H.E.: Heuristic methods for imposing structure on ill-structured problems: the structure of medical diagnostics. In: Szolovits, P. (ed.) Artificial Intelligence in Medicine .AAAS Selected Symposium, vol. 51, pp. 119–190. Westview Press, Boulder (1982)
3. Szolovits, P.: Artificial Intelligence in Medicine. AAAS Selected Symposium, vol. 51. Westview Press, Boulder (1982).
4. Rafea M, Elkafrawy P, Nasef MM, Elnemr R and Jamal AT (2019) Applying Machine Learning of Erythrocytes Dynamic Antigens Store in Medicine. Front. Mol. Biosci. 6:19. doi: 10.3389/fmolb.2019.00019
5. Gitanjali, J., Ranichandra, C., & Pounambal, M. (2014) APRIORI algorithm based medical data mining for frequent disease identification. IPASJ International Journal of Information Technology (IIJIT), 2(4), 1-5.
6. Said I. , Haruna A. , Garko A. (2015) Association rule mining on medical data To predict heart disease. International Journal of Science Technology and Management, 4(8), 26-35.
7. Karthiyayini, R., & Jayaprakash, J. (2015) Association technique on prediction of chronic diseases using Apriori algorithm. International

Journal of Innovative Research in Science, Engineering and Technology, 4(6), 255-259.
8. Fahrudin, T. M., Syarif, I., & Barakbah, A. R. (2017, September). Discovering patterns of NED-breast cancer based on association rules using apriori and FP-growth. In Knowledge Creation and Intelligent Computing (IES-KCIC), 2017 International Electronics Symposium on (pp. 132-139). IEEE.
9. Poorani, S., Balasubramanie, P., & Kumar, D. V. Apriori algorithm for identifying the association rules between clinical traits of asthma.
10. Aebersold, R., Anderson, L., Caprioli, R., Druker, B., Hartwell, L. and Smith, R., 2005. Perspective: a program to improve protein biomarker discovery for cancer. Journal of proteome research, 4(4), pp.1104-1109.
11. Hanash, S., 2004. Moving forward with clinical proteomics. Clinical Proteomics, 1(1), p.3.
12. Mischak, H., Apweiler, R., Banks, R.E., Conaway, M., Coon, J., Dominiczak, A., Ehrich, J.H., Fliser, D., Girolami, M., Hermjakob, H. and Hochstrasser, D., 2007. Clinical proteomics: a need to define the field and to begin to set adequate standards. PROTEOMICS–Clinical Applications, 1(2), pp.148-156.
13. Barnes, M.R. and Gray, I.C. eds., 2003. Bioinformatics for geneticists. John Wiley & Sons.
14. Timms, J.F., Hale, O.J. and Cramer, R., 2016. Advances in mass spectrometry-based cancer research and analysis: from cancer proteomics to clinical diagnostics. Expert review of proteomics, 13(6), pp.593-607.
15. Wang, H., Shi, T., Qian, W.J., Liu, T., Kagan, J., Srivastava, S., Smith, R.D., Rodland, K.D. and Camp, D.G., 2016. The clinical impact of recent advances in LC-MS for cancer biomarker discovery and verification. Expert review of proteomics, 13(1), pp.99-114.
16. Pasini, E.M., Mann, M. and Thomas, A.W., 2010. Red blood cell proteomics. Transfusion clinique et biologique, 17(3), pp.151-164.
17. Bryk, A.H. and Wiśniewski, J.R., 2017. Quantitative analysis of human red blood cell proteome. Journal of proteome research, 16(8), pp.2752-2761.
18. Jain, K.K. and Jain, K.K., 2010. The handbook of biomarkers (pp. 23-72). New York: Springer.
19. Rifai, N., Gillette, M.A. and Carr, S.A., 2006. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nature biotechnology, 24(8), p.971.
20. Vasan, R.S., 2006. Biomarkers of cardiovascular disease: molecular basis and practical considerations. Circulation, 113(19), pp.2335-2362.
21. Gerszten, R.E. and Wang, T.J., 2008. The search for new cardiovascular biomarkers. Nature, 451(7181), p.949.

22. Naylor, S., 2003. Biomarkers: current perspectives and future prospects. 525–529. doi: 10.1586/14737159.3.5.525
23. http://www.biomarkersconsortium.org.
24. World Health Organization, 2001. Biomarkers in risk assessment: Validity and validation.
25. Aebersold, R., Anderson, L., Caprioli, R., Druker, B., Hartwell, L. and Smith, R., 2005. Perspective: a program to improve protein biomarker discovery for cancer. Journal of proteome research, 4(4), pp.1104-1109.
26. Polanski, M. & Anderson, N. L. A list of candidate cancer biomarkers for targeted proteomics. Biomark. Insights 2007; 2, 1–48.
27. Füzey, A. K., Levin, J., Chan, M. M., Chan, D. W. Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges. Clin. Proteomics 2013; 10, 13–27.
28. Menon, U. et al. Risk algorithm using serial biomarker measurements doubles the number of screen-detected cancers compared with a single-threshold rule in the United Kingdom Collaborative Trial of Ovarian cancer screening. J. Clin. Oncol. 2015; 33, 2062–2075.
29. Pavlou, M. P., Diamandis, E. P., Blasutig, I. M. The long journey of cancer biomarkers from the bench to the clinic. Clin. Chem. 2013; 59, 147–157.
30. Kamel, H.F.M. and Al-Amodi, H.S.B., 2016. Cancer Biomarkers. In Role of Biomarkers in Medicine. IntechOpen.
31. Schiffman, J.D., Fisher, P.G. and Gibbs, P., 2015. Early detection of cancer: past, present, and future. Am Soc Clin Oncol Educ Book, 35(1), pp.57-65.
32. World Health Organization, 2008. Cancer Control: Knowledge Into Action: WHO Guide for Effective Programmes. Policy and Advocacy. Module 6 (Vol. 6). World Health Organization.
33. Brennan, D.J., O'connor, D.P., Rexhepaj, E., Ponten, F. and Gallagher, W.M., 2010. Antibody-based proteomics: fast-tracking molecular diagnostics in oncology. Nature Reviews Cancer, 10(9), p.605.
34. Neagu, M., Constantin, C., Tanase, C. and Boda, D., 2011. Patented biomarker panels in early detection of cancer. Recent patents on biomarkers, 1(1), pp.10-24.
35. Vlahou, A., 2013. Network views for personalized medicine. PROTEOMICS–Clinical Applications, 7(5-6), pp.384-387.
36. Frantzi, M., Bhat, A. and Latosinska, A., 2014. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. Clinical and translational medicine, 3(1), p.7.
37. Rafea, M. and Souchelnytskyi, S., 2012. Rediscovering Red Blood Cells: Revealing Their Dynamic Antigens Store and Its Role in Health and Disease. In Blood Cell-An Overview of Studies in Hematology. IntechOpen.
38. Rafea, M., Zaki, H. and Sultan, T., 2010, November. Bioinformatics data mining tool using data collected from red blood cells hemolysate. In

2010 2nd International Conference on Computer Technology and Development (pp. 485-489). IEEE.

Measuring rice yield from space: The case of Thai Binh Province, Viet Nam

Lakshman Nagraj Rao[1], Kaiyu Guan[2,3], Ngo The Hien[4], Zhan Li[5]

[1] Asian Development Bank, Manila, Philippines.

[2] Department of Natural Resources and Environmental Sciences, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA.

[3] National Center for Supercomputing Applications, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA.

[4] Center for Informatics and Statistics, Ministry of Agriculture and Rural Development, Viet Nam.

[5] School for the Environment, University of Massachusetts Boston, MA 02125, USA.

Abstract

Despite a growing interest in using satellite data to estimate paddy rice yield in Southeast Asia, significant cloud coverage has led to a scarcity of usable optical data for such analysis. In this paper, we study the feasibility of using surface reflectance fusion data which integrates Landsat and Moderate Resolution Imaging Spectroradiometer (MODIS) images to circumvent the cloud cover problem and estimate yield in Thai Binh Province, Viet Nam during the second growing season of 2015. Our findings indicate that although Landsat–MODIS fusion data are not necessarily beneficial for paddy rice mapping when compared with only using Landsat data, fusion data allows us to estimate the peak value of various vegetation indices and derive the best empirical relationship between these indices and yield data from the field.

Keywords

Crop cutting; yield, fusion; paddy area; remote sensing

1. Introduction

Traditionally, crop area and yield are estimated using administrative data or sample surveys (Asian Development Bank 2016). However, measurement related concerns persist in both cases as data collection officers, farmers, and others involved in the process may have the tendency to systematically over or underestimate production and area in their assigned areas (Dillon and Rao, 2018). An alternative to using administrative data or conducting surveys is the application of satellite remote sensing techniques, which has been ongoing for the past several decades with some progress achieved for paddy rice (Kuenzer and Knauer 2013; Mosleh, Hassan, and Chowdhury 2015).

From a methodological perspective, substantial progress has been made on remote sensing techniques to identify rice areas. However, estimating rice yield in the remote sensing context is still at a very nascent stage. There are several challenges associated with satellite-based crop yield estimation. First, there is a lack of reliable ground-truth crop yield data for model calibration

and testing at regional scales. Field-level crop cutting data is usually costly and labor-intensive, and district-level crop statistics are either not easily accessible or of low quality in developing countries (Asian Development Bank 2016). Second, satellite data with both high temporal and spatial resolutions are limited in terms of availability and cost. Given that the majority of paddy rice fields in Southeast Asia are smallholder farms, there is a need for high spatial resolution data down to 10–30 meters (m), and high-frequency time series data during the peak growing season to develop an advanced crop yield algorithm (Lobell et al. 2015, Sibley et al. 2014).

The objective of this paper is to build a prototype to map paddy rice fields and estimate crop yield in Thai Binh, using two satellite data sources - Landsat, and MODIS, alongside field data collected through crop-cutting activities during the rainy season of 2015. This study contributes to the growing literature on yield estimation using remote sensing techniques in its innovative employment of data fusion of Landsat– MODIS for crop yield estimation, which makes it possible to obtain high resolution data in both space than the individual sources themselves, which is critical for estimating rice area and yields in settings where smallholder farms are prevalent.

2. Methodology

A. Study Area

The study area includes the province of Thai Binh, located in northeastern coastal Viet Nam. Thai Binh is a key paddy rice production area in the Red River Delta region which is the second largest paddy rice-producing region in Viet Nam. Paddy rice is grown twice a year – during summer (mid-June to early October) and winter (mid-December to late May). Thai Binh has one key rainy season which starts in May and ends in October. Our study focuses on the summer growing season of 2015.

B. Landsat–MODIS Fusion

To overcome the challenge of availability of satellite data with high spatial and temporal resolution, we fuse the surface reflectance data from Landsat (16-day, 30 m) and MODIS (daily, 250–500 m) to generate a product that has both high spatial and high temporal resolution. We employ a mature Landsat–MODIS fusion algorithm, the Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) (Gao et al. 2006). STARFM model blends Landsat and MODIS data to generate synthetic daily surface reflectance products at Landsat spatial resolution based on a deterministic weighting function computed by spectral similarity, temporal difference, and spatial distance. The algorithm requires Landsat and MODIS pair images for the same date with clear-day quality. This posed several challenges for our study (Chen 2011). First, no single completely clear Landsat scene was available in the study area due to cloud contamination and the SLC-off problem, which limited the

selection of Landsat–MODIS pair images. To address this issue, we used a gap-filling algorithm called Geostatistical Neighborhood Similar Pixel Interpolator (GNSPI) (Zhu, Liu, and Chen 2012). Second, the ratio of valid pixels of MODIS images from both Terra and Aqua were also limited due to clouds. Thus, we use a pair of MODIS and Landsat images to train the STARFM algorithm and apply it to the rest of the MODIS images when MODIS surface reflectance data is available.

C. Paddy Rice Mapping and Land-Cover Classification

To identify paddy rice area from satellite images, we classify the land cover of Thai Binh into six categories, namely croplands, barren, built-ups, water, wetlands, and other vegetation. These were based on the International Geosphere-Biosphere Programme (IGBP) classification scheme used by MODIS global land cover product (Friedl et al. 2002). The six classes were selected based on our visual interpretation of high-resolution images on Google Earth and the knowledge from local field crew. Since paddy rice is the predominant crop grown in Thai Binh during the rainy season, the category "Croplands" refers to paddy rice in our study.

Our land cover classification uses a random forest classifier (RFC) algorithm (Breiman 2001), which has been widely tested and proved robust and efficient in the classification of remote sensing images (Hansen et al. 2000; Pal 2005; Zhu et al. 2016). The training pixels were selected as evenly as possible across the spatial extent of the images and excluded from pixel sampling during the assessment of classification accuracy. For the classification accuracy assessment, we follow the protocol set up by Olofsson et al. (2014). We obtain the conjectured overall accuracy and user's accuracy from the cross validation of RFC and prescribe the expected standard errors of user's accuracy for the six classes as 0.01 for croplands, 0.05 for barren, 0.05 for built-ups, 0.02 for water, 0.05 for wetlands, and 0.10 for other vegetation.

D. Crop Yield Estimation

A three-stage stratified sampling methodology was employed for the crop cutting survey, using an area frame that was constructed based on the expected likelihood of finding paddy rice area. Training of field staff on crop-cutting activities was conducted in September 2015 with significant attention paid to the creation of the survey instruments and methodology for crop cutting (Durante et. al 2018). The actual fieldwork took place between late September 2015 and early November 2015, covering the period associated with rice harvesting in Thai Binh. Crop cutting was implemented in random 2.5m x 2.5m square sub-plots in selected rice plots.

Usually, two variables are needed to predict yield – Aboveground Biomass (AGB), and Harvest Index. AGB usually can be approximated by the peak vegetation index, which can be derived from the Landsat–MODIS fusion data

through a curve fitting from the fused data points. Harvest Index requires spatially variable weather data and/or multiple-season data to capture the impact of climatic conditions. In this study, we only have crop cutting data for one growing season. Also, given the relatively small area of Thai Binh, there may not be significant variation in climatic variables across the province. Thus, we primarily focus on approximating AGB for yield estimation, under the assumption that all rice fields in the province share the same harvest index for the current growing season.

To overcome the large gaps and the noises of both positives and negatives in our time series data, we use a simple quadratic curve fitting method to derive peak vegetation indexes of the second growing season. The quadratic curve is centered at DOY 250, which was determined by visually inspecting many time series of crop pixels distributed over the study area. To reduce the impact of noises in the time series to our peak estimation, we calculate the standard deviation of the fitted curve and remove vegetation index values beyond three standard deviations from the mean. Then a new curve is fitted to the remaining vegetation index values. This procedure is repeated iteratively until all the vegetation index values for the curve fitting are within the confidence interval of the curve fitting. The derived peak vegetation index values of the pixels of all the representative field subplots are then regressed against the crop cutting yield data. We use NDVI, EVI, and GCVI peak values respectively to derive univariate linear regression models.

3. Results

A. Landsat–MODIS Fusion

Figure 1 shows a typical example of a 30 m by 30 m pixel time series from both the Landsat–MODIS fusion data (Figure 1 top panel) and original Landsat data (Figure 1 bottom panel). We can see clearly two growing cycles from the NDVI data from the two sources of this example pixel, with the first growing season ending around DOY 190, and the second growing season peaking around DOY 250. It is worth noting that if we only rely on Landsat data, we will not have a clear-day scene during the peak growing season around DOY 250 as shown in Figure 1. Only through the fusion approach can we recover the information during the peak value of NDVI for the second growing season.

B. Paddy Rice Mapping

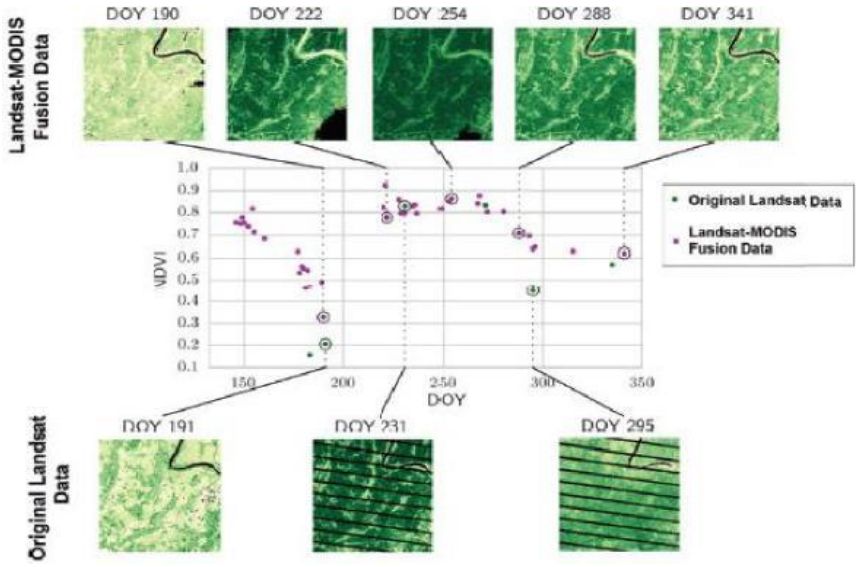
The overall accuracy associated with classifying landcover of Thai Binh using four different inputs are ranked from high to low as: () "Landsat + ALOS-2", ② "Landsat Only", ③ "Fusion NDVI SG Fit", and ④ "ALOS-2 Only". The difference between the first two inputs, "Landsat + ALOS-2" and "Landsat Only" is small, 0.77±0.02 versus 0.76±0.02. For the class of our main interest here, paddy rice, user's accuracy follows the same ranking order across the four inputs. The producer's accuracy of paddy rice is the highest for the input

"Fusion NDVI SG Fit" despite the lower user's accuracy and the relatively low overall accuracy. By merging all the maps from the four inputs, the derived final land cover map shows accuracy levels similar to the "Landsat + ALOS-2" input (Figure 2). Figure 3 shows the final land cover classification map produced by merging the four inputs.

C. Paddy Rice Mapping

The overall accuracy associated with classifying landcover of Thai Binh using four different inputs are ranked from high to low as: () "Landsat + ALOS-2", ② "Landsat Only", ③ "Fusion NDVI SG Fit", and ④ "ALOS-2 Only". The difference between the first two inputs, "Landsat + ALOS-2" and "Landsat Only" is small, 0.77±0.02 versus 0.76±0.02. For the class of our main interest here, paddy rice, user's accuracy follows the same ranking order across the four inputs. The producer's accuracy of paddy rice is the highest for the input "Fusion NDVI SG Fit" despite the lower user's accuracy and the relatively low overall accuracy. By merging all the maps from the four inputs, the derived final land cover map shows accuracy levels similar to the "Landsat + ALOS-2" input (Figure 2). Figure 3 shows the final land cover classification map produced by merging the four inputs.

Figure 1. Normalized Difference Vegetation Index Time Series



DOY = date of year, MODIS = Moderate Resolution Imaging Spectroradiometer.

Note: The series shows a 30 m by 30 m pixel that combines the original Landsat data (in green points) and the Landsat–MODIS fused data (in purple points). The top and bottom rows show the image data (3,000 m by 3,000 m) that correspond to different time stamps, and the corresponding DOY and NDVI

values at the central of the image. The second rice growing cycle starts around DOY 200.

Figure 2. Results from Using Four Different Inputs

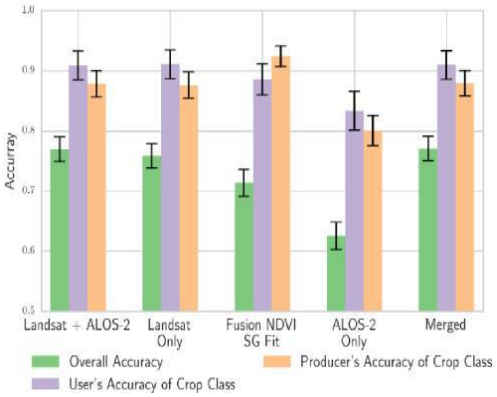
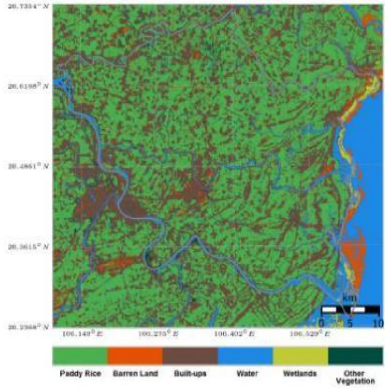


Figure 3. Classified Land Cover Map Resulting from Merging Four Inputs



ALOS = Advanced Land Observing Satellite, NDVI = normalized difference vegetation index, SG = Savitzky-Golay.

D. Crop Yield Estimation

All the three vegetation indices (NDVI, EVI, and GCVI) show some contribution to estimating crop yield at field or pixel level, as suggested by the low values of F test against the null hypothesis of an intercept-only model (Figure 4). The NDVI-based model gives the best performance, as indicated by the R2 of 0.40 for all the representative field subplots (Figure 6, black solid line), the highest among the three vegetation indexes. If we only include the dominant rice variety, BC15, in the regression, accounting for 58% of the representative subplots, the R2 increases significantly for all the three vegetation indexes (Figure 4, purple solid line). This increase in the R2 value suggests that different crop varieties may lead to different relationships between vegetation indices and crop yield, making the collection of crop variety information a crucial input.

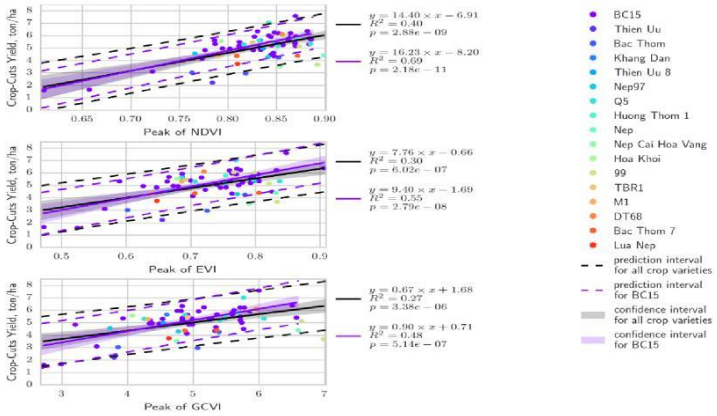
E. Scaling up to the Whole Province and Regional Validation

We apply the best yield estimation model, i.e., using peak NDVI for all the crop varieties, to the whole province of Thai Binh, shown in Figure 5. The figure clearly shows a large spatial heterogeneity in crop yield from 3 t/ha to 6.5 t/ha, with the northern part of the province having the lowest crop yield, which is consistent with the local survey data.

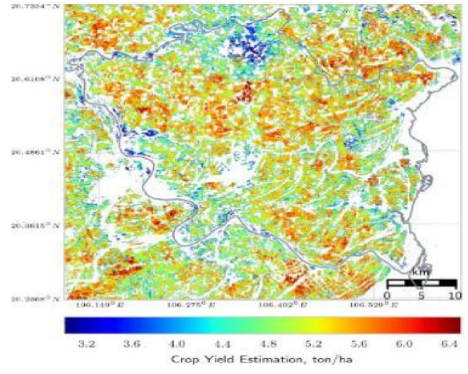
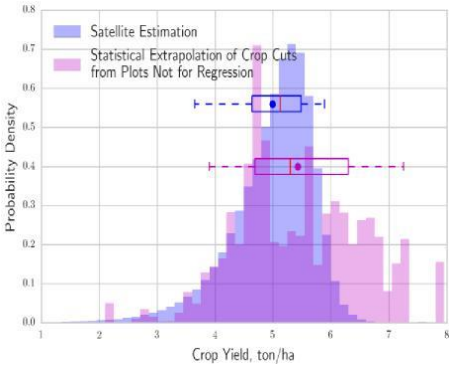
The probability density distribution of crop yield from the NDVI-based regression model within Thai Binh (not the whole image extent) is a near-normal distribution with a slight skew toward the low tail (Figure 6, blue bars). We derive the probability density distribution of crop yield (Figure 6, purple

bars) from those field subplots that are not used in our regression and their area weights within the province given by the statistical extrapolation of field samples. The field-sampling-based distribution is bimodal at approximately from 5.0 t/ha to 5.70 t/ha, if we assume a normal distribution of crop yield.

Figure 4. Linear Regression Model between the Peak of Vegetation Indexes and Crop Yield



EVI = enhanced vegetation index, GCVI = green chlorophyll vegetation index, NDVI = normalized difference vegetation index. Note: The vegetation indexes are NDVI, EVI, and GCVI. Crop yield for all the crop varieties are represented by the black line and BC15 by the purple line. Colors of the dots refer to different crop varieties.

Figure 5: Spatially Explicit Yield Map based on Normalized Difference Vegetation Index (NDVI) Crop Yield Estimates	Figure 6. Probability Density Histogram of Satellite-Based and Spatially Explicit
	

4. Discussion and Conclusion

In this study, multiple satellite data sources (including optical and L-band radar data) were used to map the paddy rice in Thai Binh, Viet Nam. Fused Landsat–MODIS data and crop cutting data were used for estimating field-

level yield data for Thai Binh. Results show that while the Landsat–MODIS fused data does not necessarily show benefits for paddy rice mapping, it has provided great benefits for crop yield estimation. Only through the fusion data from Landsat and MODIS can we recover the peak growth trajectory of vegetation indexes. This information is the most critical input for our current algorithm. Our results also confirm the value of optical data for crop yield estimation if the cloudiness issue can be alleviated or overcome to some degree. We recognize that the current fusion approach still has room for improvement as has been reviewed by Gao et al. (2015), and as is being further improved by Zhu et al. (2016).

One possible issue here is how to best utilize the Landsat–MODIS fused data and original Landsat data. More advanced smoothing or weighted regression approaches are needed to deal with the possible discrepancy between the fused and original data. Meanwhile, emerging new datasets of surface reflectance, such as Sentinel-2 (20 m resolution, 16-day revisiting frequency) and Project for On-Board Autonomy - Vegetation (PROBA-V) from Satellite Pour l'Observation de la Terre- VEGETATION (SPOT-VGT) (100 m resolution, 16-day revisiting frequency), can further improve the temporal and spatial samplings to alleviate cloudiness issue in tropics. New fusion algorithms thus should consider multiple sources of data for fusion, instead of only for Landsat and MODIS.

References

1. Asian Development Bank. 2016. *Results of the Methodological Studies for Agricultural and Rural Statistics*. Manila.
2. Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32.
3. Chen, J., Zhu, X., Vogelmann, J.E., Gao, F., & Jin, S. (2011). A Simple and Effective Method for Filling Gaps in Landsat ETM+ SLC-off Images. *Remote Sensing of Environment, 115*(4), 1053–1064.
4. Dillon, A. & Rao, L.N. (2018). Land Measurement Bias: Comparisons from Global Positioning System, Self-Reports, and Satellite Data. *ADB Economics Working Paper Series*. No. 540. Manila: Asian Development Bank.
5. Durante, A.C.D., Lapitan, P., Megill, D., & Rao, L.N. (2018). Improving Paddy Rice Statistics Using Area Sampling Frame Technique. *ADB Economics Working Paper Series*. No. 565. Manila: Asian Development Bank.
6. Friedl, M.A., Mciver, D.K., Hodges, J. C. F., Zhang, X.Y., Muchoney, D.M., & Strahler, A.H. (2002). Global Land Cover Mapping from MODIS: Algorithms and Early Results. *Remote Sensing of Environment, 83*, 287–302.

7. Gao, F., Hilker, T., Zhu, X., Anderson, M., Masek, J., Wang, P. & Yang, Y. (2015). Fusing Landsat and MODIS Data for Vegetation Monitoring. *IEEE Geoscience and Remote Sensing Magazine, 3*(3), 47–60.
8. Gao, F., Masek, J., Schwaller, M., & Hall, F. (2006). On the Blending of the Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance. *IEEE Transactions on Geoscience and Remote Sensing, 44*(8). 2207– 2218.
9. Hansen, M. C., Defries, R.S., Townshend, J. R. G., & Sohlberg, R. (2000). Global Land Cover Classification at 1 km Spatial Resolution Using a Classification Tree Approach. *International Journal of Remote Sensing, 21*(6). 1331–1364.
10. Kuenzer, C. & Knauer, K. (2013). Remote Sensing of Rice Crop Areas. *International Journal of Remote Sensing, 34*(6). 2101–2139.
11. Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B.B. (2015). A Scalable Satellite-Based Crop Yield Mapper. *Remote Sensing of Environment*.
12. Mosleh, M.K., Hassan, Q.K., & Chowdhury, E.H. (2015). Application of Remote Sensors in Mapping Rice Area and Forecasting Its Production: A Review. *Sensors, 15*(1). 769–791.
13. Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., & Wulder, M.A. (2014). Good Practices for Estimating Area and Assessing Accuracy of Land Change. *Remote Sensing of Environmen,. 148*, 42–57.
14. Pal, M. (2005). Random Forest Classifier for Remote Sensing Classification. *International Journal of Remote Sensing, 26*(1), 217–222.
15. Sibley, A.M., Grassini, P., Thomas, N.E., Cassman, K.G., & Lobell, D.B. (2014). Testing Remote Sensing Approaches for Assessing Yield Variability among Maize Fields. *Agronomy Journal, 106*(1), 24.
16. Zhu, X., Liu, D., & Chen, J. (2012). A New Geostatistical Approach for Filling Gaps in Landsat ETM+ SLC-off Images. *Remote Sensing of Environment, 124*, 49–60.
17. Zhu, Z., Fu, Y., Woodcock, C.E., Olofsson, P., Vogelmann, J.E., Holden, C., & Yu, Y. (2016). Including Land Cover Change in Analysis of Greenness Trends Using All Available Landsat 5, 7, and 8 Images: A Case Study from Guangzhou, China (2000-2014). *Remote Sensing of Environment, 185*,243–257.

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

How much better is better? Quantifying the CAPI advantage using Viet Nam's Labor Force Survey

Lakshman Nagraj Rao[1], Dave Pipon[2], Jude David Roque[*3]
[1] Asian Development Bank, Manila Philippines
[2] Asian Development Bank, Manila Philippines
[3] Asian Development Bank, Manila Philippines

Abstract

Labor statistics published by government agencies rely on data from Labor Force Surveys (LFS), which in most countries are conducted using the pencil-and-paper interviewing (PAPI) technique. More recently, there has been a concerted effort for countries to switch to computer-assisted personal interviewing (CAPI), wherein a handheld device is used during the interview process. CAPI not only eliminates the need to manually re-enter the data, but also automates questionnaire navigation and flags inconsistent responses on the fly. While these features may lead to improvements in data quality and timeliness, it is unclear to what extent, and whether these improvements affect estimates during data analysis.

This paper presents results from a randomized experiment, designed specifically to compare CAPI and PAPI using data from July 2017 - September 2017 for Ho Chi Minh in Viet Nam. Within each of a total of 180 sample enumeration areas, 15 households were randomly selected and interviewed using PAPI, while another 15 households were randomly selected and interviewed using CAPI. This design allows for a detailed comparison of errors, interview times, and costs between the two methods. In addition, we test the hypothesis whether these errors are non-random, which may lead to differences in estimates for basic labor force statistics between the two groups.

Keywords

Computer-assisted personal interviewing; data quality; randomized experiment; survey; labor statistics

1. Introduction

Datasets matter in statistical analyses and tend to be scrutinized and dissected down to minute details to generate meaningful results. Yet, not much attention is given towards how such datasets are brought about in the first place. The underlying data collection process along with the tools used, associated errors, and implications for analysis are seldom considered. This is important to consider because most data collection processes in developing countries are still reliant upon traditional pencil/pen-and-paper interviewing

(PAPI) techniques which are prone to numerous data issues, carrying huge implications for the quality of data analysis.

PAPI records data onto paper forms, which are then manually compiled and entered to come up with a data set for analysis. While a lot of surveys have benefitted from PAPI over time, the limitations of the method are potentially compromising for data quality. With surveys becoming more comprehensive and complex, PAPI might suffer from issues in data accuracy because enumerators would need to exert additional effort to accurately navigate through the more complicated logic and skips built into the surveys. At some point, more complicated surveys could cause enumerator burden to set in and potentially compromise the data gathered. Further, PAPI involves dealing with heaps of paper and reprints, entailing tedious manual data encoding, which is prone to human error. The encoding process also involves time costs, which could cause delays in data availability and analysis.

The advent of information technology has brought forth an alternative that could potentially address the limitations of PAPI in the form of computer-assisted personal interviewing (CAPI). With CAPI, interviewers use a handheld device instead of a paper questionnaire to record interview responses. The main advantages of CAPI include the elimination of paper forms and associated print outs, increased data accuracy due to automated skipping mechanisms and logical checks, and faster data availability virtually eliminating manual data entry allowing for almost immediate data analysis. Further, CAPI arguably may lower costs for larger sample sizes, which is beneficial for national statistical systems conducting multi-topic nationally representative surveys. CAPI also possesses additional features such as the ability to integrate images, video and audio recordings, timestamps, and global positioning system (GPS) information into the questionnaire.

In theory, CAPI is expected to address the limitations of PAPI when it comes to improved data quality, timeliness, and costs. Yet there is very limited literature that rigorously and empirically looks at the cost and benefits of transitioning from CAPI to PAPI in in developing economies, particularly in Asia and the Pacific. Much of the earlier research on the implications of transitioning to CAPI has usually focused on developed economies (Couper and Burt 1994; Nichols and de Leeuw 1996; Banks and Laurie 2000). Only two studies have made systematic and rigorous attempts at studying the impact of transitioning to CAPI. Caeyers et al. (2012) empirically assessed CAPI's benefit in terms of data quality, cost, and timeliness from a randomized control trial done on Tanzanian households. Meanwhile, Fafchamps et al. (2012), attempted to quantify CAPI's advantage for data quality when it comes to collected data on sales and profits for microenterprises in Ghana. Both studies only provide an African context and have contrasting findings, thereby calling for more research in this area, especially with external validity across different contexts and types of surveys.

Context of Study and Setting: Labor Force Survey (LFS) in Viet Nam

This study builds on empirical evidence from Viet Nam on how a switch to CAPI would affect interview length, costs, and data quality. The study was conducted as part of an ADB statistics capacity building project that aims to improve data collection and management of national surveys in support of the Sustainable Development Goals (SDGs) using information and communication technology tools, such as CAPI.

The study focuses on one of the major surveys implemented by Viet Nam's General Statistics Office (GSO), the LFS. The survey is implemented on a quarterly basis to obtain estimates of the country's labor market and serves to lay the groundwork for labor policies. The LFS survey instrument builds on previous versions and is continuously updated based on the recommendations of the International Labor Organization (ILO). The survey consists of 71 questions grouped into three main sections, namely (a) household and resident information (b) respondent characteristics, and (c) questions for classifying economic status. The sampling frame of this survey is based on the 2014 Intercensal Population and Housing Survey and the sample is drawn from a two-stage stratified sample, while the enumeration areas (EAs) are selected proportional to the size of the two independent sub-sample frames (urban and rural). The sample is nationally representative for 63 provinces/cities and can be disaggregated quarterly down to 6 major economic regions, the cities of Hanoi and Ho Chi Minh,
and rural and urban areas.

Research Objective

Using a randomized roll out of CAPI and PAPI LFS surveys, with a focus on Ho Chi Minh, the study aims to provide answers to the following questions:

1. What effect does CAPI have on interview time?
2. Does CAPI help reduce the number of errors made in the questionnaire and are there implications for recruitment of enumerators when switching from PAPI to CAPI?
3. What are the cost implications of switching to CAPI?

Relevance

Data collection methods have a bearing in the data sets provided for analysis. The implications are even greater for national statistics systems or offices where quality of data is crucial for them to come up with meaningful insights to serve long term policy. As such, it matters to look at ways to improve data collection. This study sheds light on the practicality and advantage of transitioning to CAPI by looking at tangible benefits such as interview length or time and the reduction in errors. Furthermore, it also potentially provides policy input on other variables involved in data collection such as the important characteristics to consider for the hiring enumerators.

Of course, another practical point that is considered in the study is whether CAPI is even a cost-effective move.

2. Methodology

Experimental Design

A randomized experiment was designed to identify the effects of CAPI vis-à-vis PAPI. This was implemented during the third quarter of 2017 in the Ho Chi Minh enumeration phase of the LFS with field operations conducted during the first fifteen days of July, August, and September 2017, respectively. From each enumeration area, 30 households were randomly selected, from which 15 were randomly assigned to CAPI, while the remainder was assigned to PAPI. The experiment enlisted separate enumerators for CAPI and PAPI. The main respondent for the survey was the household head, although some household members were requested feedback on certain questions. The CAPI arm of the experiment was implemented using CSPro for Android, which is a free platform developed by the US Census Bureau and widely used among national statistical systems in Asia and the Pacific.

A total of 174 enumerators were recruited, with 73 of them interviewing for CAPI and 101 assigned to PAPI. To the extent possible, the enumerators were randomly assigned to each group, which comes across in the similarity of enumerator characteristics shown in Table 1. Across both modes, over 60% of the enumerators were males. The average ages of enumerators for the two groups are similar. Moreover, more than three fourths of the enumerators across both methods had reported completing a university degree. The average number of years of enumerator experience was between 5 to 6 years.

Table 1. Enumerator Characteristics

DESCRIPTION	Method					
	CAPI		PAPI		Total	
	Freq	%	Freq	%	Freq	%
Total number of Enumerator	73	42	101	58	174	100
Sex						
Male	51	69.86	63	62.38	114	65.52
Female	22	30.14	38	37.62	60	34.48
Average Age	34.92		38.41		36.94	
Educational Attainment						
Lower secondary school	0	0.00	2	1.98	2	1.15

Mid-term professional school	4	5.48	7	6.93	11	6.32
University	56	76.71	77	76.24	133	76.44
Upper secondary school	13	17.81	15	14.85	28	16.09
Average Experience	5.04		6.01		5.60	

One of the metrics in the experiment used as a benchmark to assess data quality is the total number of errors committed by the mode of data collection. Errors were classified across four categories: skip, validation, logical, and missing. Skip errors are those where an enumerator failed to implement the skip rules incorporated into the questionnaire based on responses to preceding questions. Data validation errors are those that are committed based on the condition that answer to a question is restricted to certain values. For example, when a question only expects numeric answers up to 100 but the enumerators inputs a 200, it will be flagged as a validation error. Logical errors refer to those committed that fail to meet cross-sectional logic. An example of this is a respondent answering "Male" to a gender question and "Currently pregnant" to another question. Finally, missing errors refer to those fields that were unanswered but were required based on previously set conditions within the questionnaire.

Empirical Strategy

To investigate the first research question looking at the effect CAPI has on interview time, the following Ordinary Least Squares model was applied:

$$Y_{ijc} = \alpha + \beta \times CAPI_i + \boldsymbol{\delta} \times \boldsymbol{HH} + \boldsymbol{\phi} \times \boldsymbol{ENUM} + \epsilon$$

where, Y_{ijc} refers to the Survey Duration in minutes in the questionnaire of household i interviewed by enumerator j in cluster c. Meanwhile, $CAPI_i$ = 1 if the questionnaire was implemented using CAPI and 0 if otherwise. Moreover, $\boldsymbol{HH}$ are the household characteristics and $\boldsymbol{ENUM}$ are the enumerator characteristics. The specification also uses enumeration area and enumerator code for fixed effects.

To answer the second research question which is to identify if CAPI led to a reduction in errors and if enumerator characteristics matter, the following model was employed:

$$Z_{ijc} = \alpha + \beta \times CAPI_i + \boldsymbol{\delta} \times \boldsymbol{HH} + \boldsymbol{\phi} \times \boldsymbol{ENUM} + \epsilon$$

where, Z_{ijc} in this case is the total number of errors (sum of skip, logical, validation, and missing) committed in the questionnaire of household i interviewed by enumerator j in cluster c. Again, $CAPI_i$ = 1 if the mode of collection was CAPI and 0 if otherwise. $\boldsymbol{HH}$ and $\boldsymbol{ENUM}$ are the household characteristics and enumerator characteristics, respectively. The fixed effects employed for the model include enumeration area and enumerator code.

Finally, to assess how the costs of CAPI stack up against traditional modes, back of the envelope calculations were done looking at fixed and variable costs for CAPI and PAPI to determine the break-even point. The simple arithmetic for such is as follows:

$$FC_{CAPI} + (VC_{CAPI} \times No.Questionnaire) = FC_{PAPI} + (VC_{PAPI} \times No.Questionnaire)$$

where FC is the fixed cost and VC is the variable cost to conduct the survey.

3. Results

The result of the first model investigating CAPI's effect on interview time is summarized in Table 2. The model strongly suggests that CAPI reduces interview times by about 28.5 minutes. This finding is significant and robust even when accounting for enumerator and household characteristics.

The model also finds that across both modes it seems that females' enumerators tend to conduct shorter interviews. Meanwhile, it is interesting to note that enumerators that possess higher education (college) seem to be associated with shorter interviews. Duration seems to decrease with age, while more enumerator experience seems to be linked to longer interviews. A possible reason for this could be that more experienced enumerators might just be more meticulous and detail oriented in administering the questions and probing for answers. Also, as expected, the number of adult household members contributes to the duration of the interview positively.

The results of the second model are presented in Table 3. Our study finds that errors are statistically significantly reduced with CAPI (Table 3). To put into perspective, moving from PAPI to CAPI is associated with reducing the error incidence by about 1.5 per questionnaire. Our study also suggests that regardless of whether CAPI or PAPI is used, household characteristics still do matter as far as the reduction of errors, while only the gender of the enumerator has implications for data quality. Most importantly, using CAPI does not mean that errors are completely random, thereby producing unbiased statistics. This has implications for data analysis and policy.

Table 2. CAPI and Interview Time

VARIABLES	Duration
Method of Data Collection	
CAPI (base = PAPI)	-28.52***
	(6.342)
Household Characteristics	
Sex of HH Head (base = Male)	-0.0994
	(0.347)
Age of HH Head	-0.0346***
	(0.00731)
Wage (in thousand VND)	2.66e-05**
	(1.11e-05)
No. of adults in the HH	2.956***
	(0.0791)
Enumerator Characteristics	
Sex (base = Male)	-17.52***
	(3.031)
Age	-0.784***
	(0.0966)
Experience (in years)	5.048***
	(1.574)
Squared Experience (in years)	-0.0438
	(0.163)
Education (base = Below	-12.30***
	(2.233)
Interaction Effect	
Enumerator and HH head Sex	0.167
	(0.415)
Constant	61.23***
	(7.558)
Observations	5,332
R-squared	0.788
EA FE	Yes
Enumerator FE	Yes

Dependent variable: Survey Duration (in minutes)
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3. Errors and Enumerator Characteristics

VARIABLES	Error Model
Method of Data Collection	
CAPI (base = PAPI)	-1.576**
	(0.798)
Household Characteristics	
Sex of HH Head (base = Male)	-0.0394
	(0.0437)
Age of HH Head	-0.00605***
	(0.000921)
Wage (in thousand VND)	2.05e-05***
	(1.40e-06)
No. of adults in the HH	0.186***
	(0.00996)
Enumerator Characteristics	
Sex (base = Male)	-1.529***
	(0.382)
Age	0.0151
	(0.0122)
Experience (in years)	0.237
	(0.198)
Squared Experience (in years)	-0.0139
	(0.0205)
Education (base = Below	-0.188
	(0.281)
Interaction Effect	
Enumerator and HH head Sex	0.0547
	(0.0522)
Constant	2.108**
	(0.952)
Observations	5,332
R-squared	0.258
EA FE	Yes
Enumerator FE	Yes

Dependent variable: Total number of errors
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

It is imperative that before scaling up any technology to ascertain whether the costs make sense. Implementing either CAPI or PAPI involves fixed and variable costs. The cost analysis involves using these cost categories to identify the breakeven point where CAPI becomes more cost effective than PAPI. First, the fixed costs involved include software costs and programming costs. The software costs are both free for PAPI and CAPI, since we are utilizing CSPro which is open source. On the other hand, programming costs involved hiring a programmer/developer for a total of 15 consultancy days for programming PAPI, and 29 days for CAPI. With a consultancy daily rate of $400 the programming costs would be $6,000 and $11,500 for PAPI and CAPI, respectively. In terms of variable costs, these include tablets, data entry costs,

printing costs, and miscellaneous costs (e.g. storage, bags, internet, electricity, etc.). Tablet costs are not applicable for PAPI, while the cost is $195 for CAPI. With a lifespan of 2 to 3 years, the cost of using a tablet per day is $0.21. If an enumerator completed 5 interviews a day, the per questionnaire cost of a tablet comes out at $0.04. Meanwhile, PAPI data entry and printing costs were estimated to be at $0.65 and $1.50, respectively. CAPI is not expected to incur these costs, as these variable costs are virtually eliminated with the technology. Finally, miscellaneous cost estimates came out to be $1.50 and $0.5 per questionnaire for PAPI and CAPI, respectively. This brings the total fixed costs for CAPI and PAPI at $11,500 and $6,000, respectively. Meanwhile the variable costs for CAPI are $0.54 and $3.65 for CAPI and PAPI, respectively. Computing for the breakeven point using the arithmetic equation stated in the methodology results in a breakeven point of 1,769 questionnaires. In other words, CAPI becomes more cost-effective for survey operations involve 1,769 or more questionnaires.

4. Conclusion

The study is an attempt to systematically look at the benefits of a transition to CAPI from a developing Asia and the Pacific perspective, particularly looking at its implications for data quality, interview and costs. The results corroborate the literature in terms of the perceived benefits of CAPI for data quality. First, CAPI seems to reduce the number of errors committed per questionnaire. The reduction in error is also correlated with household characteristics, although out of the enumerator characteristics only female enumerators tend to reduce the number of errors. Furthermore, the findings strongly suggest that CAPI has much shorter interview durations, which has huge implications for the turnaround time for data to be processed from field to headquarters. Female enumerators tend to take less time to accomplish the interviews. Finally, the cost analysis showed that costs were in favour of CAPI for medium to large scale surveys of more than 1800 households. This works well for statistics offices that cater to larger national scale surveys. These findings come together to strengthen the case for CAPI's adoption in national statistical systems worldwide and build a case for female enumerators to be given priority given that they are able to accomplish less erroneous questionnaires in lesser time than men.

References

1. Banks, R. & Laurie, H. (2000). From PAPI to CAPI: The Case of the British Household Panel Survey.*Social Science Computer Review*, **18**(4), 397-406.
2. Caeyers, B., Chalmers, N., & De Weerdt, J. (2012). Improving Consumption Measurement and Other Survey Data Through CAPI: Evidence from a Randomized Experiment. *Journal of Development Economics*, **98**(2), 19-33.

3. Couper, M. & Burt, G. (1994). Interview Attitudes Toward Computer-Assisted Personal Interviewing (CAPI). *Social Science Computer Review,* **12**(1), 38-54.
4. Fafchamps, M., McKenzie, D., Quinn, S. & Woodruff, C. (2012). Using PDA consistency checks to increase the precision of profits and sales measurement in panels. *Journal of Development Economics*, **98**(1), 51-57.
5. Nichols, W. & de Leeuw, E. (1996*).* Factors in Acceptance of Computer-Assisted Interviewing Methods: A Conceptual and Historical Review. *Proceedings of the Section of Survey Research Methods, American Statistical Association*, 758-763.

On the identification and handling of outliers in composite index data

Ali S. Hadi
Department of Mathematics and Actuarial Science,
The American University in Cairo, Egypt. E-mail ahadi@aucegypt.edu

Abstract

Composite indices data often need editing before the computation of the indices. Variables may need some transformation due to their high skewness or kurtosis coefficients. Also, outliers are commonly found in composite index data. These outliers can drastically affect the results of composite indices. Identification of outliers improves data quality and reliability, hence it improves the quality of the decisions drawn from the data and analysis. Three important steps in constructing composite indices are (1) determining the variables that need transformation, (2) identifying outliers when they exist in index data, and (3) what to do with the outliers once they are identified? We discuss these steps in constructing composite indices.

Keywords

BACON; Kurtosis; MCD; Mahalanobis distance; Min-Max Normalization; Outliers; Robust, Skewness

1. Introduction

Numerous composite indices are computed on an annual basis. For example, the Global Knowledge Index, the corruption perception index (CPI), The Human Development Index (HDI), the Ibrahim Index of African Governance (IIAG), the Gender Inequality Index (GII), and the Climate Change Performance Index (CCPI) to mention only a few. Bandura (2008) provides a survey of the current composite indices around the world. At that time Bandura (2008) found 187 indices.

Most recently, the International Knowledge Index (IKI) was computed for the first time and published in 2017 by the United Nations Development Program (UNDP). The IKI extended the Arab Knowledge Index (AKI) which was computed for the first time in 2015 by the Al Maktoum Foundation (http://www.mbrf.ae/) to measure knowledge in the Arab countries.

Composite indices data are high-dimensional data, where the number of variables sometimes exceeds the number of observations. A composite index is a single number summary for each observation in the data. The quality of a composite index cannot exceed the quality of the data that are used to construct the index. Variables can be highly skewed and/or have severe kurtosis. The data may also contain univariate and multivariate outliers. These

violations of the expected norm can have a severe influence on the computed index numbers as they maylead to biased index values.

The rest of this paper is organized as follows: Sections 2 and 3 discuss ways of identifying and treating the above mentioned violations. Section 4 discusses what to do with the outliers once they are identified.

2. The Univariate Approach

Each variable in a composite index data should be examined individually for the presence of skewness, kurtosis, and/or the presence of univariate outliers before the composite index is computed.

2.1 Skewness and Kurtosis

The skewness coefficient is a measure of the lack of symmetry in data distribution about the mean. Let x1, x2, ..., x_n denote the n observation of a variable X. A common definition of the skewness coefficient is given by, see, e.g., Groeneveld and Meeden, (1984) and Johnson, Kotz, and Balakrishnan (1994),

$$SC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}, where\ \bar{x} = n^{-1}\sum_{i=1}^{n} x_i\ and\ s = \sqrt{n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Values of 0, negative, positive SC indicates that the distribution of the variable is symmetric, negatively skewed, and positively skewed distribution.

The Kurtosis is a measure of the thickness of the tail of the data distribution relative to the Normal distribution. A definition of the Kurtosis coefficient is given
by

$$KC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

A zero value of KC indicates that the distribution has a tail similar in thickness as that of a normal distribution. A positive or negative value indicates that the distribution has a heavy-tails or light-tails relative to the tails of the Normal distribution, respectively.

Other measures of skewness and kurtosis are discussed in Joanes and Gill (1998). See also, Hair et al. (2015).

When making statistical inference based of index data, the data are assumed to be Normally-distributed. If the Skewness and/or Kurtosis coefficients are far from zero, they indicate departure from the Normality assumptions. Variables with significant Skewness and/or Kurtosis coefficients may require special treatment before the index is computed. As an

experimental rule, a variable has a sever skewness if its absolute SC is greater than 2 and severe kurtosis if its absolute value of KC is greater than 0.5.

What to do with variables that have severe skewness and/or kurtosis? One way out here is to use the Box-Cox power transformation to make the variable that have severe skewness and/or kurtosis closer to the Normal distribution. To be specific, one can replace the i-th value, xi, by $y_i(\lambda) = \sum_i^{\lambda} -1)/\lambda$. The parameter λ is chosen such that the distribution of the variable Y(λ) is close to normal. One way to achieve this is draw the Normal Probability Plot of Y(λ) and choose the value of λ that makes the graph as linear as possible. Techniques such as the use of sliders (see, e.g., the software package Data Desk) can be used to achieve this goal. Alternatively, the function "BoxCoxLambda" in the R package "DescTools" automatically detects the optimal parameter A. Note that if the optimal value of l turns out to be zero, this indicates that the optimal transformation of the log transformation, that is, y(0) = log(x).

For example, Figure 1(a) shows the histogram of a variable X, which shows clear departure from Normality as indicated by SC = 5.079 (significantly positively skewed) and KC = 25.495 (significantly heavy right tail distribution). The variable is highly skewed and has a relatively heavy tail. The variable needs transformation to achieve Normality. Figure 1(b) shows the Normal Q-Q plot of the variable X. Here λ = 1 means no transformation is taken. The scatter of points do not resemble a straight line and the correlation between the sample quantiles and the theoretical quantile (under Normality) is low (correlation = 0.544). Consistent with the histogram in Figure 1(a), this graph in (b) shows clear departure from Normality.

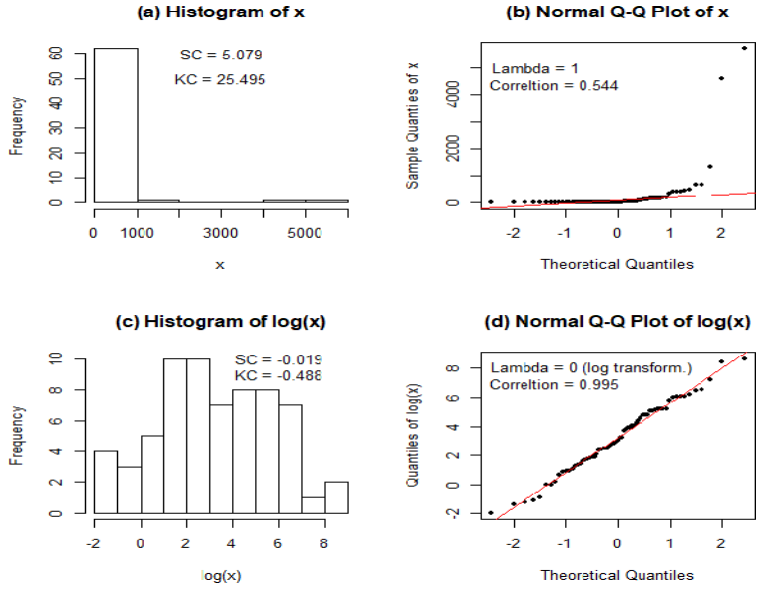


Figure 1. Box-Cox Transformation of the variable X

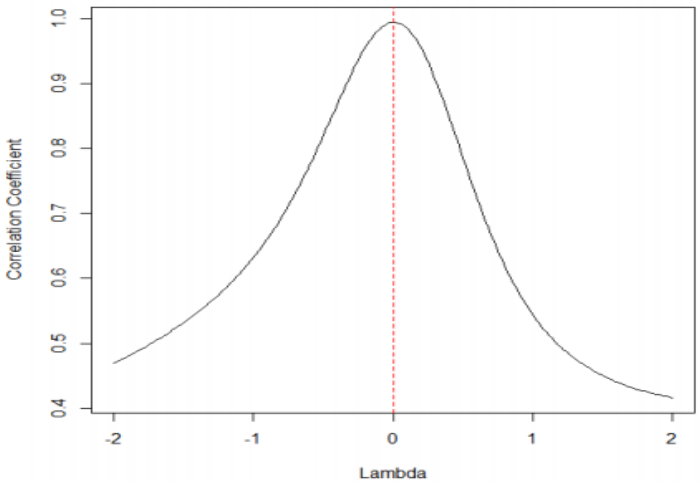


Figure 2. The graph of the correlation between Y(λ) and the Normal scores versus λ.

The optimal value of power transformation parameter λ is zero, indicating a log transformation is needed. Indeed. Figure 2, which is the graph of the correlation between Y(λ) and the Normal scores versus λ, for λ between —2 and 2, shows that the optimal value of λ is zero indicating that log(X) is much closer to a Normal variable. The histogram of log(X) is shown in Figure 1(C), which indicates that the assumption of the Normality of log(X) is supported by the data. Here SC = -0.019 and KC = -0.488 compared with SC = 5.079 and KC = 25.495 before transformation. Figure 1(d) is the Normal Q-Q Plot of log(X), which shows strong linearity and a very high correlation of 0.995. The power transformation here succeeded in transforming a highly skewed and heavy tailed distribution to a nearly symmetric variable.

2.2. Univariate Outliers

One way to identify outliers in the composite index data is to plot a box plot for each variable in the data. Points that fall outside the boxplot limits are declared outliers. The boxplot limits are given by

Lower Limit = Q1 – 1.5 (Q3 – Q1) and Upper Limit = Q3 + 1.5 (Q3 – Q1),

where Q1 and Q3 are the first and third quartiles of the data, respectively. Accordingly an observation xi is declared as an outlier if either xi < Lower Limit or xi is greater than the Upper Limit. Outliers are then treated by replacing them by the Lower Limit (if they are on the low side) or by the Upper Limit (if they are on the high side). This rule is used by composite indices such at the Global Knowledge Index; see the Al Maktoum Foundation Web site at: http://www.mbrf.ae/.

Another approach for the identification of univariate outliers is as follows: The a% trimmed mean and trimmed standard deviation are computed based on the central 100(1 – α)% of the values. Commonly used choices of a are 5% and 10%. Then all observations more than 3 trimmed standard deviations away from the trimmed mean are declared outliers. Outliers are then treated by replacing them by

Trimmed mean – 3:1 x Trimmed standard deviation,

if they are in the lower tail, or by

Trimmed mean + 3:1 x Trimmed standard deviation,

if they are in the upper tail. This rule, which is simple and also strikes a balance between efficiency and robustness, is adopted by other well-established indices, such as the Ibrahim Index of African Governance (IIAG); see the MIF foundation Web site at: mo.ibrahim.foundation.

3. The Multivariate Approach

Although the methods for the identification of univariate outliers are simple and easy to compute, they may fail to identify outliers in higher-dimensional spaces. For example, Figure 3 shows that scatter plot of two variables Y and X. Here we can see clearly that there are two outliers in the two-dimensional space. Neither one of the univariate methods for identifying the outliers will identify these two observations because each one of then falls near the mean of X and the mean of Y. An early method for the identification of multivariate outliers is to compute the Mahalanobis distances (Mahalanobis, 1936)

$$M(\boldsymbol{x_i}, \overline{\boldsymbol{x}}, \boldsymbol{S}) = \sqrt{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})}, \text{ for i=1,2,....,n,}$$

where $\boldsymbol{x_i}$ is a vector representing the i-th observation in the p-dimensional space, p is the number of variables in the index data, $\overline{\boldsymbol{x}}$ is the mean vector, and $\boldsymbol{S}$ is the p x p covariance matrix of the data. Here M measures the elliptical distance between $\boldsymbol{x_i}$ and $\overline{\boldsymbol{x}}$ relative to the covariance matrix $\boldsymbol{S}$. Values of M larger than $x^2_{p,\alpha/n}$ are declared multivariate outliers, where $x^2_{p,\alpha/n}$ is the α/n upper quantile of the x^2 distribution with p degrees of freedom, and α is the significance level. Here we divide α by n as a way of Bonferrouni adjustment.

Mahalanobis distances are easy to compute but they are not robust because they depend on $\overline{\boldsymbol{x}}$ and $\boldsymbol{S}$, which are not robust. One may replace $\overline{\boldsymbol{x}}$ and $\boldsymbol{S}$ by a robust version of them, say $\overline{\boldsymbol{x}}_r$ and $\boldsymbol{S}_r$.This gives a robustified version of the Mahalanobis distances, that is,

$$M_r(\boldsymbol{x_i}, \bar{\boldsymbol{x}}_r, \boldsymbol{S_r}) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_r)^T \boldsymbol{S}_r^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})}, \text{ for i=1,2,....,n,}$$

Several methods exist for obtaining x r and S r. Two of the most common ways are the Minimum Covariance Determinant (MCD) proposed, e.g., in Rousseuw and Van Driessen (1999), and the Blocked Adaptive, Computationally-Efficient outlier Numerator (BACON), proposed by Billor et al. (2000). These two methods are implemented in R using the functions "CovMcd" in the package "rrcov" and BACON in the Package "robustX".

Consider for example two variables X and Y. Figure 3(a) shows a boxplot for each of the two variables. The boxplot rule does not declare any outliers. The scatter plot of Y versus X, shown in Figure 3(b) shows clearly that there are four outliers in the data. These four observations are bivariate outliers. Usually, outliers in high dimensional space are not easily detected by examining lower dimensional spaces. When applying multivariate outlier detection methods, like the three mentioned above, the outliers clearly stand out in the Index Plot of the distances as shown in Figure 4, where the Index Plots of the Mahalanobis distances and the robust distances obtained by using the BACON and the MCD methods are displayed. Even the non-robust Mahalanobis distances is able to identify the four observations as outliers.

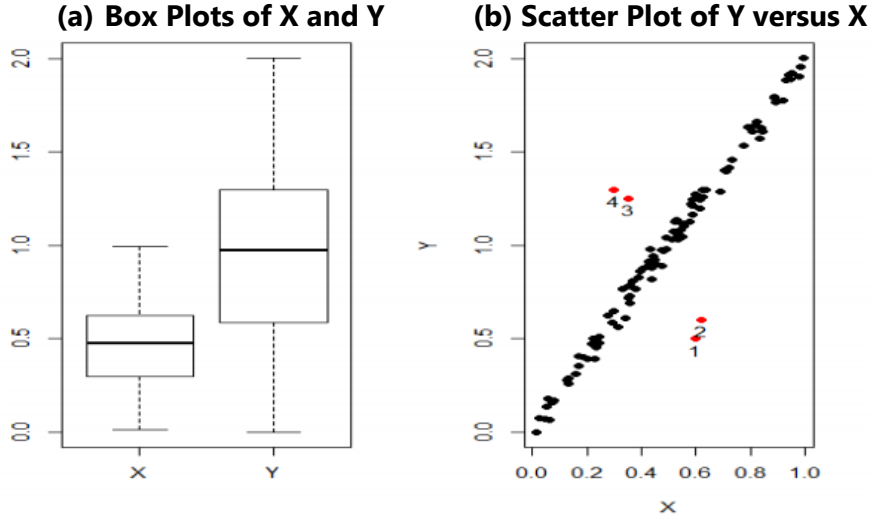


Figure 3. (a) Boxplots of two variables, X and Y and (b) their scatter plot

4. Discussion and Conclusion

Composite indices data often need editing before the computation of the indices. Highly skewed variables and variables with fat tails may need some transformation to achieve symmetry or normality. In addition, univariate and multivariate outliers in the data need to be identified and dealt with before

the computation of the index numbers. In this article, we discussed methods for editing the data that deal with these problems.

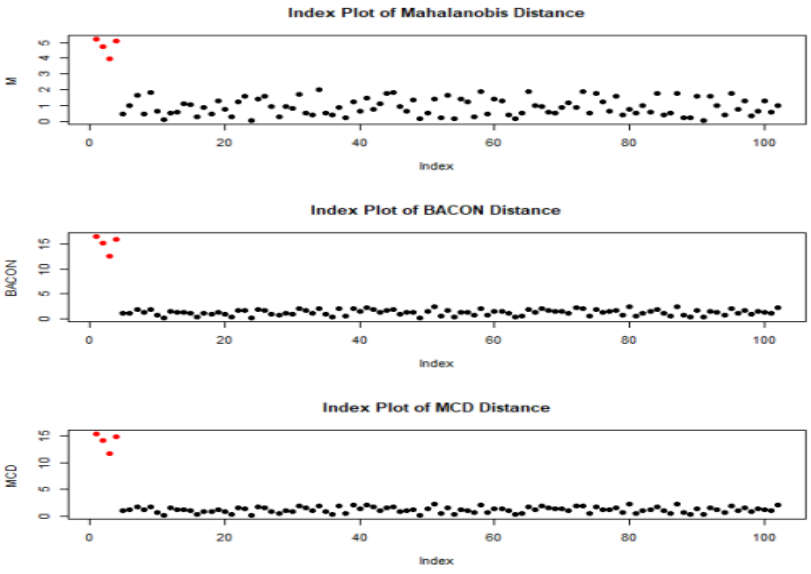


Figure 4. Index Plot of multivariate distances using Mahalanobis, BACON, and MCD methods.

References

1. Al Maktoum Maktoum Foundation http://www.mbrf.ae.
2. Bandura R. (2008), "A Survey of Composite Indices Measuring Country Performance," The United Nations Development Program, Office of Development Studies.
3. Billor, N., Hadi, A. S., and Velleman, P. F. (2000), "Blocked Adaptive, Computationally-Efficient
4. Groeneveld, R.A., and G. Meeden (1984). Measuring Skewness and Kurtosis. *The Statistician,* Vol.33, pp. 391–99.
5. Hair, J., R. Anderson, R. Tatham and W. Black (2015). *Multivariate Data Analysis*. 7th ed. Upper Saddle River, NJ: Prentice-Hall International.
6. IIAG: http://www.moibrahimfoundation.org.
7. Joanes, D. N., Gill, C. A. (1998): Comparing measures of sample skewness and Kurtosis. *The Statistician*, 47, 183–189.
8. Johnson, NL, Kotz, S, Balakrishnan, N (1994) *Continuous Univariate Distributions,* Vol 1, 2nd Edition Wiley.
9. Mahalanobis, P. C. (1936), *"On the Generalized Distance in Statistics," Proceedings of the National Institute of Sciences of India*.**2** (1): 49–55.
10. Organisation for Economic Co-operation Development (OECD), (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Paris.

11. Rousseuw, P. J., and Van Driessen, K., (1999), "A Fast Algorithm for Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223.

Composite indices and traditional data – The global knowledge index

Mohamed A Ismail
Department of Statistics, Cairo University, Egypt

Abstract

The Global Knowledge Index (GKI) consists of six sectoral indices and one supporting index focusing on the general enabling environment. Each of these seven indices was constructed in accordance with standard international methodologies for the design of composite indicators. The GKI is structured with a hierarchy of five levels: index, constituting indices (also referred to as sectoral indices), pillars, sub-pillars and variables. Each of the six sectoral indices has a weight of 15 percent, while the general enabling environment has a weight of 10 percent.

The selection of variables (individual variables) included in the construction of each of the seven indices was based on a well-defined scientific methodology drawn from an extensive review of relevant local and international literature, as well as the experiences and concepts of international organizations and agencies. It also relied on an intensive consultation process that engaged experts.

Principal components analysis was used to confirm the consistency of the selected variables and the structure of their classification into the various sub-indices, further supporting the consistency of the broader conceptual context across the variables and their classification in the sub groups - for which the explained variance ratio in most cases exceeded 50 percent. The results of the in-depth correlation analysis and Cronbach's Alpha coefficient (exceeding 0.70 in most cases) confirmed the validity of the selection and classification of the variables.

Weights assigned for the seven constituent indices range from equal weighting and budget allocation to factor analysis. The weights produced by using both the budget allocation and factor analysis methods were consistent with each other and with the initial weights estimates, based on the intellectual and conceptual framework. The arithmetic aggregation formula was used to calculate all composite indicators of the Index.

Keywords

Knowledge Index; Education; TVET; ICT; Budget Allocation

1. Introduction

The Global Knowledge Index (GKI) aims to measure the multidimensional concept of knowledge. The concept is a fluid one, often linked to related

concepts such as 'knowledge economy' or 'knowledge society'. It is also sometimes restricted to a narrow understanding that limits the focus to education or technology. Given the variations in its use and meaning, the GKI aims to introduce a more systematic understanding of knowledge in two respects. First it breaks down the concept into its constituent components—i.e. education, economy, research and technology. Therefore, it recognizes the multidimensional nature of knowledge systems in all contexts and applications relating to economic and social structures. Second it also enables a more scientific and evidence-based linkage between development and a multidimensional concept of knowledge, in keeping with the notion of human development as applied by the UNDP as well as the concept of sustainable development agreed by world leaders in 2015 at the 2030 Agenda for Sustainable Development. The GKI is a joint initiative between the United Nations Development Programme and the Mohammed bin Rashid Al Maktoum Knowledge Foundation (MBRF).

Attempts to bridge knowledge gaps between countries cannot depend on improvizational processes or interpretations based on unreliable data and analysis. Rather, it is necessary to gather a precise and objective description of the reality of those gaps in their different manifestations. This requires a systematic assessment process based on scientific indicators that take into consideration the multidimensional nature of knowledge and its functional links to sustainable human development. The Global Knowledge Index has been developed in response to this specific need.

Several attempts to measure knowledge in the context of understanding knowledge within the broader context of the economy or institutional structures may include:

1. World Bank Institute (2008) introduced a Knowledge Assessment Methodology (KAM). This methodology is based on two indices: the Knowledge Index (KI) and the Knowledge Economy Index (KEI). The first is the simple average of three pillars: innovation; education; and ICT infrastructure. The second represents an arithmetic mean of four pillars, with the addition of the economic and institutional system. The World Bank stopped producing the index as of 2013.
2. World Economic Forum (2010) proposed the Lisbon Scorecard as a tool for comparing the progress made by EU member states towards developing knowledge economies to that achieved by the USA and East Asian countries. The last modified format covered five pillars: innovation; liberalization; enterprise; employment and social inclusion; and sustainable development and the environment.
3. The European Innovation Scoreboard includes 27 indicators under four main groups: framework conditions, investments, innovation activities and impacts, see Leon (2017).
4. United Nations Development Programme and Mohammed bin Rashid

Al Maktoum Foundation (2015) launched the first Arab initiative to build six composite indicators to measure the state of knowledge in the Arab region from a development perspective, while focusing on the sectors of education; economy; ICT; and research, development and innovation (RDI). United Nations Development Programme and Mohammed bin Rashid Al Maktoum Foundation (2016) released the second version of the Arab Knowledge Index, alongside a special index on reading in the Arab region as a knowledge-related component.

2. Methodology

Due to the multidimensional nature of knowledge, a composite index was constructed consisting of six sectoral sub-indices. Composite indices allow for a single value that gives a fuller picture of the phenomenon being measured, especially if this phenomenon is multidimensional as in this case. Composite measurements also better reflect possible connections between different dimensions and their internal interactions, and allow for standard comparisons between countries.

The structure of the GKI covers the most important dimensions of development. The sectoral indices that form the pillars of the GKI are: (1)pre-university education; (2)technical vocational education and training (TVET); (3) higher education; (4) research, development and innovation (RDI); (5) information and communications technology (ICT); (6) economy. A seventh pillar was added to support the sectoral indices, as these sectors do not operate in isolation from their surroundings, but rather in a space governed by a range of contextual factors—political, socio-economic, health-related and environmental. Each of these seven indices was constructed in accordance with standard international methodologies for the design of composite indicators, OECD (2008).

The GKI is structured with a hierarchy of five levels: index, constituting indices (also referred to as sectoral indices), pillars, sub-pillars and variables. Each of the six sectoral indices has a weight of 15 percent, while the general enabling environment has a weight of 10 percent.

Selection of variables

The selection of variables (individual variables) included in the construction of each of the seven indices was based on a well-defined scientific methodology drawn from an extensive review of relevant local and international literature, as well as the experiences and concepts of international organizations and agencies. It also relied on an intensive consultation process that engaged experts -from different countries, including Canada, Egypt, Jordan, the UAE, the United Kingdom and the United States- each of them specialized in fields related to the sectors of the GKI. Experts expressed their agreement, rejection or proposed additions or amendments

to a selected list of variables and aggregations. Based on this feedback, and that of the core team who prepared the report, a final list of variables was produced.

Principal components analysis was used to confirm the consistency of the selected variables and the structure of their classification into the various sub-indices, further supporting the consistency of the broader conceptual context across the variables and their classification in the sub groups—for which the explained variance ratio in most cases exceeded 50 percent, see Hair et al. (2015).

The results of the in-depth correlation analysis and Cronbach's Alpha coefficient (exceeding 0.70 in most cases) confirmed the validity of the selection and classification of the variables. Furthermore, the correlation matrix for normalised variables was analysed to ensure that they followed the same direction as the composite index, confirming the need to include variables with high correlation coefficients (above 0.9) with other variables.

Data collection

The 133 variables employed in the 2018 GKI were obtained from sources including the United Nations Educational, Scientific and Cultural Organization (UNESCO); the World Bank; the International Telecommunication Union (ITU); the World Economic Forum (WEF); the International Monetary Fund (IMF); the Organisation for Economic Co-operation and Development (OECD); the International Labour Organization (ILO) and other UN and international agencies. The team reviewed the data multiple times to ensure no errors had occurred during data entry; consequently, data was processed on the assumption that it was error-free. In the cases where those variables were linked to other size-dependent variables – such as population or GDP – results were recalculated after adjusting for the effect of the size. Variables included are in the form of hard data, composite indicators and survey questions/responses. The most recent data for each variable within the period 2007–2018 was used.

As a prerequisite, data employed in the construction of the composite indices have met certain statistical criteria. For example, each country was required to have at least 50 percent of the figures for variables in each sectoral index for it to be included in the general index (GKI). The team had to ensure these criteria were met before calculating the composite index. The methods used to identify and treat outliers, severe skewness and severe kurtosis are outlined below.

Data treatment

A variable was considered to have severe skewness if its absolute skewness coefficient was above 2.25, while an absolute kurtosis coefficient above 3.5 indicated that the variable had severe kurtosis. Conditions were relaxed due

to the small sample size (134 countries), but those variables with severe skewness and/or severe kurtosis required statistical treatment before they could be employed, see Groeneveld, R.A., and G. Meeden (1984). Variables with one to five outliers were winsorized, whereby those values considered outliers were assigned the next highest value until the skewness and kurtosis were brought into acceptable ranges. However, five variables with more than five outliers required additional calculation and were treated using logarithm and the square root transformations, see Cornell University et al. (2017).
The value of a variable was considered an outlier if its instance fell outside the range of the specific data fence defined as follows:

$$\textit{Lower bound} = \textit{first quartile} - 1.5\ X\ \textit{interquartile range}$$
$$\textit{Upper bound} = \textit{third quartile} + 1.5\ X\ \textit{interquartile range}$$

Outliers were treated by replacing each outlier with the second highest value in the case of high values, or the second lowest value in the case of low values.

Normalization

The rescaling or 'maximum–minimum' method was used for normalization. The values of variables were normalized into the 0–100 range, in which higher values indicated better results. The normalization criterion depends on whether the variable is good (has a positive relation with the overall Index) or bad (has a negative relation with the overall Index). The good variables were normalised using the following formula:

$$\text{Normalized value} = \frac{Country\ value - minimum\ sample\ value}{Maximum\ sample\ value - minimum\ sample\ value} \times 100$$

In the case of bad variables (i.e. those with an inverse relation) the formula was adjusted to:

$$\text{Normalized value} = \frac{Maximum\ sample\ value - country\ sample\ value}{Maximum\ sample\ value - minimum\ sample\ value} \times 100$$

For survey data or composite indices, the original series' range of values was retained in the form of minimum and maximum values; for instance, in the case of the 1–7 range for the World Economic Forum Executive Opinion Survey variables.

Index weighting

It should be noted that weighting across the index components (indices, pillars and sub-pillars) was not unified, and varied according to the nature of the components and their relative importance. Weightings identified for the seven constituent indices range from equal weighting and budget allocation

to factor analysis. Equal weights were used in the absence of any clear evidence of a diversity of significance among variables, as well as in the absence of sound and complete information concerning the existence of causal relationships, or where a lack of consensus exists on a classical method for estimating weights.

The budget allocation method was also used for weighting where a group of specialists and experienced experts were invited to attend a workshop for each of the knowledge sectors. Each expert was given a budget consisting of 100 points to award to the variables. If the variable was believed to have greater relative importance, it was allocated a greater number of points. Subsequently, the weights were calculated according to the average of the total points allocated to each variable.

The weights were also assessed using factor analysis, which is based on aggregating the linked sub-indicators to form a single factor containing as much information as possible that is shared between these linked indicators. The weights produced by using both the budget allocation and factor analysis methods were consistent with each other and with the initial weights estimates, based on the intellectual and conceptual framework, OECD (2008).

Index calculation

The 2018 GKI was calculated for 134 countries in this second edition, using the most recent and best available data to calculate the variables for each country, with 2007 as a cut-off year and 2006 being exceptionally used for specific countries that required additional data to qualify for inclusion in the Index. The values of the composite sub-index were calculated by applying a series of successive aggregations starting with the (more detailed) variables and ending with the production of the index.

Owing to the lack of availability of data covering all the components for each country, and in view of the need to maintain an adequate level of accuracy, the composite index was calculated in a bottom-up pattern, where the upper level index is calculated only where at least two thirds of its components are available. This applies to all knowledge sub-indices and for all countries. In cases where data for a variable was not available for at least half of the countries, this variable was excluded from the calculation of the overall composite indicators (i.e. excluded from the index structure).

The arithmetic aggregation formula was used to calculate all composite indicators of the Index. The composite indicator (CI) is calculated by aggregating its sub-components (SCj) as:

$$CI = \sum_{j=1}^{n} w_j \times SC_j$$

CI is the proposed composite indicator to be computed; wj is the relative weight of the sub-component SC; and n is the number of sub-components aggregated to form the composite indicator.

3. Results

GKI scores are distributed on a scale from 0 to 100. Higher scores indicate greater progress towards meeting the knowledge requirements of development. Switzerland (73), , Finland (69), Sweden (69), the United States (68) and Luxembourg (68) obtained the highest rankings, with scores ranging between 68 to 73. The majority of the high-scoring countries either belong to the European Union or are located in East Asia. In the Arab region, the top-scoring country was the UAE, which was ranked 19th. Despite their relative high scores compared to others, this group of countries is yet to achieve maximum knowledge efficiency. The lowest scoring countries on the GKI scale were mostly Sub-Saharan countries. However, they also included three Arab countries and a number of countries in South and West Asia.

Comparing these results to those of the 2018 Sustainable Development Index reveals significant correlation between the two indices in terms of the countries at the top and bottom of the rankings. Furthermore, correlation coefficient analysis of the GKI and the Human Development Index 2018 showed a very high correlation of around 0.8. An equally high correlation between the GKI and the Sustainable Development Goals Index reinforces the conceptual foundation and assumption of the GKI that there is a strong correlation between knowledge and sustainable human development.

At the sectoral indices level, the lowest scores were recorded in the RDI sector and ranged from 8 to 61.9. The highest scores were in the General Enabling Environment (between 25.7 and 84.6). The same trend was noted when the analysis focused on the best-performing 10 countries and the worst-performing 10 countries, although average scores varied (Table 1).

Table 1. Sectoral average scores of top 10 ranked and lowest 10 ranked countries

	GKI	Pre-University	TVET	Higher Education	RDI	ICT	Economy	Enabling Environment
Top10	68	51	68	58	56	64	63	81
Lowest	30	28	36	28	15	27	32	34

Country scores in the GKI as a whole, as well as in the sectoral indices, reflect large gaps between nations. These gaps also vary between sectors – with the widest being in the ICT, RDI and Enabling environment sectors.

4. Discussion and Conclusion

The Global Knowledge Index was created, comprising six sectoral composite indices and a seventh index measuring the availability of a general enabling environment. The importance of this initiative stems from the need to support sound, evidence-based and effective decision- and policymaking. However, the Index alone will not be sufficient to devise new policies or interventions. The true function of the Index is to shed light on the strengths and weaknesses of a given system, not to provide suggestions for how to respond to these strengths and weaknesses. Additional, tailored analysis by relevant national, regional and international actors is required to identify appropriate actions and interventions. Furthermore, it is important to ensure continuity in data collection in order to amass information over a sufficient period of time to allow accurate analysis of positive and negative trends, rather than providing a snapshot of the situation at a single point in time. Another intended benefit of the Index is the facilitation of partnerships and exchanges between relevant actors such as governments, decision makers, scholars and all other concerned authorities at the local, regional and international levels. In this regard, the Index may facilitate collaborative projects among institutions and organizations that seek to build their own indices to fill the gaps in existing knowledge systems.

While recognising the strength of the methodology and structure, the Global Knowledge Index will be subject to regular updates, revisions and refinement. This will allow us to develop an increasingly credible and relevant product that can adapt to global transformations and remain responsive to evolving development requirements. In its next edition, more space will be allocated to the environment component, which constitutes an important pillar in achieving sustainable human development in its contemporary sense. The process of refining the Index will not be free from challenges, such as creating comprehensive, accurate and regularly updated databases, expanding data sources, developing methodologies that allow accurate utilization of big data and checking the robustness of the index using global sensitivity analysis, Saltelli el al. (2008).

References

1. Cornell University, INSEAD and WIPO (2017). *The Global Innovation Index 2017: Innovation Feeding the World*. Tenth Edition. Ithaca, Fontainebleau, and Geneva. Available from: https://www.globalinnovationindex.org/gii-2017-report
2. Groeneveld, R.A., and G. Meeden (1984). "Measuring Skewness and Kurtosis". *The Statistician*, vol. 33, no. 4, pp. 391–99.
3. Hair, J., R. Anderson, R. Tatham and W. Black (2015). *Multivariate Data Analysis* (7th ed.). Upper Saddle River, NJ: Prentice-Hall International.

4. Leon, R.D. (2017). Measuring the Knowledge Economy: A National and Organizational Perspective. *Management Dynamics in the Knowledge Economy*, vol. 5, No. 2. Available from:http://www.managementdynamics.ro/index.php/journal/article/view/212/173
5. Organisation for Economic Co-operation Development (OECD), (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Paris.
6. Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana and S. Tarantola (2008). *Global Sensitivity Analysis: The Primer*. Chichester: John Wiley & Sons.
7. United Nations Development Programme and Mohammed bin Rashid Al Maktoum Foundation (2015). *Arab Knowledge Index 2015*. Dubai: Al Ghurair Printing and Publishing. Available from: http://www.knowledge4all.com/uploads/files/AKI2015/PDFEn/AKI2015_Full_En.pdf
8. United Nations Development Programme and Mohammed bin Rashid Al Maktoum Foundation (2016). *Arab Knowledge Index 2016*. Dubai: Al Ghurair Printing and Publishing. Available from: http://knowledge4all.com/admin/uploads/files/AKI2016/En/ArabKnowledgeIndex2016.pdf
9. World Bank Institute (2008). *Measuring Knowledge in the World's Economies. Knowledge Assessment Methodology and Knowledge Economy Index. Knowledge Development Program*. Available from: http://web.worldbank.org/archive/website01030/WEB/IMAGES/KAM_V4.PDF
10. World Economic Forum (2010). *The Lisbon Review 2010: Towards a More Competitive Europe?*. Geneva. Available from: https://www.weforum.org/reports/lisbon-review-2010

Moving from traditional data to big data in assessing knowledge societies

Bruno André RodriguesCoelho[1], Stamatis Kalogirou[1], Iakov Frizis[1], Anthony Fakhoury[2]
[1]PwC Luxembourg, [2]United Nations Development Programme

Abstract
After years of measuring the current state of knowledge based on data from national statistical and data systems, and given the exponential growth of knowledge creation, a new knowledge measurement tool using big data is being designed to better understand the future of knowledge societies. This report presents a pilot study, covering 20 countries, on the future fields of knowledge that will shape the future of knowledge societies. The purpose of the study is to have a better understanding of today's strong and weak signals in the next wave of (technology) disruption by testing a new way to capture and analyse real-time data associated with five key future fields of knowledge: Artificial Intelligence (AI), Cybersecurity, Blockchain, Biotechnology, and Future Skills and to accelerate knowledge development by helping country leaders to benchmark their performance against that of front-runners.
To ensure a reliable representation of the variation in readiness for the future across countries, it is essential to use tools that enable the gathering of data as close to real time as possible. This helps us identify and discuss the latest technological advances, the future impact of technological change on the economy and society, and the level of technological awareness that characterizes labour markets.

Keywords
Real-time data

1. Introduction

The Fourth Industrial Revolution (also called Industry 4.0) is unfolding before our eyes. It is the era of creative convergence, where a myriad of technologies that span themes such as the Internet of Things (IoT), cloud computing, big data analytics and artificial intelligence (AI) begin to connect, (PriceWaterHouseCoopers, 2016) creating an ecosystem in which each technology both exploits and fosters the development of the others (OECD, 2017). This revolution creates unprecedented opportunities and challenges for businesses and societies alike. It is distinct from prior revolutions, given its intensity, complexity and scope, and it is rooted in a new technological phenomenon – digitalization, i.e. the integration of digital technologies – that is penetrating the infrastructure of every business, organization and

government with unprecedented speed (Schwab, 2016). In this working paper, we propose a new method for measuring where countries stand in terms of future fields of knowledge using real-time data, and thus offer government leaders and supporting stakeholders a tool for anticipating what may come next.

The purpose of this paper is to present the methodology that allowed us to have a better understanding of today's strong and weak signals in the next wave of technological disruption by testing a new way to capture and analyse real-time data associated with five key future fields of knowledge: Artificial Intelligence (AI), Cybersecurity, Blockchain, Biotechnology, and Future Skills.

When we talk about the future of societies, technological change cannot be considered in isolation. This report focuses on "key technologies for the future", which are technologies we believe will help overcome most of the challenges associated with globalization, sustainability, demographic shifts and urbanization. The European Commission calls these technologies "Key Enabling Technologies" (KETs) (European Commission, Februrary, 2018). Harnessing the opportunity offered by these technologies will require investment in five key dimensions, which we call the five knowledge dimensions:

- Education
- Research, Development and Innovation (RDI) and Science
- Technology
- Economy
- Enabling Environment

2. Methodology

The present methodology is based on recent developments in the community of impact evaluation practitioners, scientists and policymakers regarding the use of alternative metrics for impact assessment. The availability of webometrics [1] has enabled the increasing use of publicly available information for assessing the societal impact of an object of evaluation.

The European Commission has recently set an Expert Group on Altmetrics with the purpose of discussing and providing evidence on alternative metrics for impact assessment (applied to the impact of science and innovation) and of formulating recommendations for their future utilization. The group's final report "Next-generation metrics – Responsible metrics and evaluation for open science" (2017), (European Commission, 2017) lays the groundwork for the use of alternative metrics as complementary to traditional metrics in impact evaluation. Within this context, our team has had

[1] Examples of webometrics include: simplistic counts and content analysis of web pages, counts and analyses of outgoing links from web pages or "outlinks," and links pointing to web pages, called "inlinks" (Björneborn L. and Ingwersen P., 2001).

the opportunity to implement the logic behind the use of alternative metrics for evaluation under the prominent European project "Digital Entrepreneurship Monitor," where PwC developed and applied a methodology for assessing technology uptake using "real-time big data" extracted from publicly available sources (European Commission, 2018).

Research design and data collection

We selected three types of metrics commonly used in social monitoring and listening for measuring future knowledge development:

- The number of mentions of a specific topic (i.e. number of times a specific set of keywords assumed to define a specific topic are mentioned online);
- The level of engagement on a specific topic (i.e. number of times an online publication has been forwarded, shared or commented on);
- Sentiment concerning a specific topic (i.e. overall mood associated with the context in which a specific set of keywords appears, which can be either positive, neutral or negative).

Nevertheless, existing knowledge can only be transformed into new knowledge when two interrelated processes take place, socialization and combination. Socialization, allows the sharing of tacit knowledge and combination involves the conversion of explicit knowledge into more complex sets of explicit knowledge. The amount of socialisation and the intensity of communication and rate of dissemination of (explicit) knowledge occurring within a community can therefore be used as a proxy of that community's capacity for future knowledge creation.

In an innovative attempt to quantify these key processes for knowledge creation, this pilot study uses the number of mentions, the level of engagement and metrics from a sentiment analysis as representative measures of the current, real-time knowledge socialization, communication and dissemination within a country.

In order to collect web data, a social listening tool was required. The tool had to be able to crawl all public web pages and public social media sites, across the globe in a wide variety of languages.

The selected tool, the Digital Intelligence Platform, collects data from 150 million public sources and covers sources in over 180 languages. 20 countries were selected for the pilot study based on their rankings on the Global Knowledge Index 2017. Only data from these countries were retrieved, extracted and analysed:

Figure 1. Countries included in the study

Sources that were used for scraping are:

Online News	Newspapers	Magazines	Blogs	Forums	Twitter	Other Social Media

Apart from Germany, the main source was Twitter. One year's worth of data was extracted from the platform for each country. All data was published online between September 4, 2017 00:00:00 GMT+1 and September 2, 2018 23:59:59 GMT+1.

Construction of the "Future of Knowledge Model"

Figure 2. The "Future of Knowledge Model"

We first created five Future Field Readiness Indices corresponding to the five future fields of knowledge:

- Four Technology Readiness Indices, corresponding to the four key technologies for the future i.e. AI, Cybersecurity, Blockchain and Biotechnology; and
- The Future Skills Readiness Index.

To create these indices, we aggregated the raw data vertically (across knowledge dimensions), as shown in the illustration above. We calculated these indices separately for each of the 20 countries. Similarly, we aggregated the data horizontally, across all four technologies for the future, in order to create an index for each of the five knowledge dimensions (Education, RDI and Science, Technology, Economy and Enabling Environment). In addition, in order to portray the overall performance of each country across the four technology readiness dimensions, we constructed the **Global Technology Readiness Index (GTRI).**

For the calculation of each index, we first aggregated the raw data (mentions and engagement) in the level of aggregation defined by the index, and then built the index.

Future Field Readiness Indices

To construct each of the five Future Field Readiness Indices i, the score for country j is based on the aggregation of the raw data across the five substitutable knowledge dimensions of equal importance (k= {Education, RDI and Science, Technology, Economy and Enabling Environment}) and determines the performance of the country within each of the five indices.

$$Future\ Field\ Readiness\ Index_{j,i} = \sum_{k} \quad Knowledge\ Dimension_{k,j,i}$$

Each composite Index is the result of a linear combination of two types of social media metrics: the number of mentions and the level of engagement.

Global Technology Readiness Index

The Global Technology Readiness Index is a composite index that refers to the four Technology Readiness Indices that are conceptually different to the Future Skills Readiness Index. We derived the ranking of each country j by aggregating the raw data and computing a score for all four technologies (i= {AI, Blockchain, Biotechnology and Cybersecurity}). This form of ranking shows

that we treat the components of the GTRI as substitutable and of equal importance. This means that, for instance, a deficit in AI can be compensated by a surplus in Blockchain.

$$GTRI \quad _j = \sum_{i} \quad Technology \quad _{j,i}$$

Calculation of composite indices

First, we standardized the number of mentions by dividing them by the number of Internet users, which we derived from the data included in Table 2, to calculate the mention density.

$$Mention\ density = \frac{Mentions}{Internet\ Users} * 1.000.000$$

In a similar vein, to be able to compare different levels of engagement, we standardized engagement by dividing total engagement with the number of mentions for each country to compute the engagement density.

$$Engagement\ Density = \frac{Engagement}{Mentions}$$

Based on the above, we calculated each composite index as follows:

$$Composite\ Index = \frac{V1+V2}{2}$$

where V1 is the normalised value of mention density and V2 is the normalised value of engagement density, respectively. The formula to calculate the normalized values of mention density and engagement density is a standard min-max normalization that is commonly used in calculating composite indices:

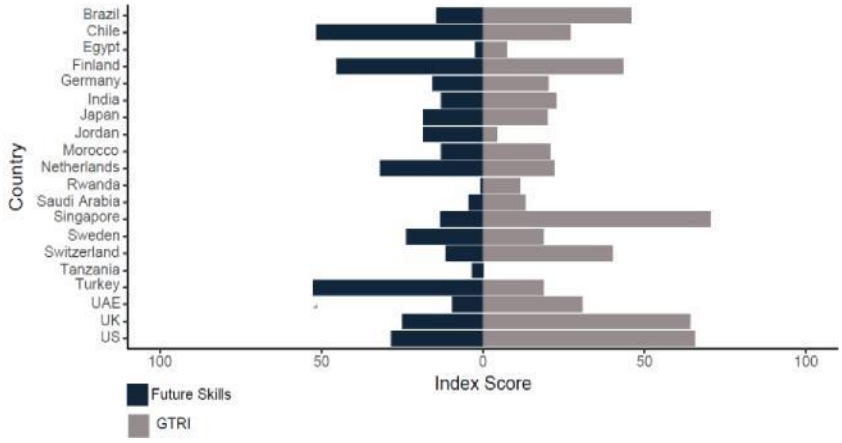
$$Normalized\ value = \frac{Actual\ Value - Min.value}{Max\ Value - Min.Value} * 100$$

3. Results

The Global Technology Readiness Index (GTRI) and the Future Skills Readiness Index are both appropriate indicators for the comparison of future readiness across countries. Figure 1 illustrates the scores on both indices.

The analysis of GTRI scores, aggregated over the period of observation for each of the 20 countries of interest, shows clear differences in the prominence of emerging technologies in the public debate. The scores are mainly concentrated in Singapore, the United States and the United Kingdom. Arab countries are the ones with the most distance to cover in terms of raising awareness and having experts/practitioners engage in discussion with the rest of the Internet using population. Tanzania is the least well-performing country in our sample. The analysis of the Future Skills Readiness Index scores across countries shows a significant concentration of scores in three countries: Turkey, Chile and Finland. A more in-depth analysis of online activity reveals that a public debate on the teacher deficit drives the observed mentions upwards for Turkey, and to a lesser extent for Finland.

Figure 3. Future Skills Readiness Index and GTRI across the 20 countries



4. Discussion and Conclusion

This study comes with certain limitations that we wish to detail in this discussion.

The occurrence of a major event that does not relate to technology uptake will drive online activity (for a country) upwards. During the sampling period, we identify two such events that may drive results. For example, the implementation of the General Data Protection Regulation (GDPR) on 25 May 2018. which was one of the most significant changes in data privacy regulation of the past 20 years. Second, the recent spike in the price of Bitcoin (together with other cryptocurrencies) has attracted attention in cryptocurrencies. Cryptocurrencies use Blockchain technology, which results in a higher frequency of Blockchain-related vocabulary use in online discussions. This heightened level of activity however, does not relate directly to technology adoption. Instead, it reflects a speculative environment of early-stage developments along the hype cycle of the technology. In terms of validity of results, we observe that Blockchain has a largely global impact in driving results, while the GDPR topic features more prominently in Europe, Brazil and in countries where international press has a strong presence. The sampling period introduces a bias that inflates results for certain countries above their true value. We recognise two factors that may drive the score of a country upwards: elections and teacher shortages. Elections, local or national, also tend to drive online activity upwards for a country. Limited access to the Internet impacts the validity of our results. In the instance where only a small subgroup of the population has access to the Internet, our sample ceases to be representative. As is the case of Rwanda, where only 20 percent of the total

population has access to Internet, access to Internet functions as proxy of upper socio-economic status.

Noisy data due to linguistic idiosyncrasies may inflate results upwards. We find little evidence of linguistic idiosyncrasies in our sample that present a challenge in terms of performing text mining. The main linguistic challenge that we face is with regard to Future Skills. High use frequency of relevant keywords in everyday discussions inflates our results for this field across all countries. However, as our key metric is an index score that we interpret in comparative terms (ranking), we expect this drawback to have little impact on the validity of our findings.

The revolutionary times we are living in constitute a great opportunity for visionary leaders – policymakers, business leaders, training providers and individuals – to realize the benefits of the fields of knowledge that will shape the society, economy, science and the education of tomorrow. However, the road ahead is not an easy one. Visionary leaders will need to become experts in strategic foresight, conduct visionary exercises, engage in experimentation and prototyping and develop flexible monitoring tools for leading and coordinating the process of future knowledge development. Luckily, the many new technologies that are emerging at present can help us to develop solutions that support leaders in carrying out these new tasks. Our monitoring approach, based on big data collected through a single Digital Intelligent Platform, is one example of such a solution, but there are many other technological tools and products being developed by creative individuals, start-ups and companies that may be leveraged. All we need to do is put our future in focus, and be open and collaborative, to build collective new knowledge and develop our learning to acquire new skills.

References

1. PriceWaterHouseCoopers(2016)
2. OECD (2017)
3. European Commission (February 2018)
4. European Commission (2017)
5. European Commission (2018)
6. Björneborn, L. & Ingwersen, P. Scientometrics (2001) 50: 65.

Index

A

Aditya Joshi, 194
Alex Khor, 17
Ali S. Hadi, 384
Amani Tariq Jamal, 357
Amrit Lakra, 31
Anastasiia Rytova, 172
Anna Lin, 56
Anthony Fakhoury, 401
Arvind Shrivastava, 8
Asma Adnane, 331

B

B. Bruijnis, 89
Beatriz Sevilla-Villanueva, 348
Brian Blankespoor, 236
Bruce Tracey, 39
Brunero Liseo, 65
Bruno André RodriguesCoelho, 401

C

Candido J. Astrologo, Jr, 311
Cecile Paris, 194
Chaker Abdelaziz Kerrache, 331
Chan Foo Keong, 17

D

Daniel Clarke, 228
Daria Balashova, 164
Dave Pipon, 375
Davide Di Cecco, 65
Devika Balan, 31

E

Elena Yarovaya, 156, 172

F

Farhan Ahmad[2], 331
Fionn Murtagh, 342

G

Guangwu Chen, 279

H

Haniza Yon, 31
Hannes I. Reuter, 244
Hataichanok Puckcharern, 317
Hui Wei, 261
Hyun Song Shin, 108

I

Iakov Frizis, 401
Iluminada T. Sicat, 82

J

Jaanus Kroon, 117
James Houran, 39
Jude David Roque, 375
Júlia M Pavan Soler, 210
Julien Gaffuri, 244
Justin Angelo O. Bantang, 311

K

K. Prokopenko, 89
Kaiyu Guan, 366
Karina Gibert, 348
Kok Mun Yee, 31

L

Lakshman Nagraj Rao, 366, 375

M

M. Symotiuk, 89
Magued Osman, 295
Mahdi Roozbeh, 143
Mahmoud Rafea, 357
Marco Di Zio, 65
Mariana Kotzeva, 244
Mariza de Andrade, 219
Mary Everett, 108
Mazlina Muhammad, 31
McSeth Antwi, 331
Md. Kamrul Islam, 150
Md. Sabiruzzaman, 150
Michael Wild, 236
Miquel Sànchez-Marrè, 348
Mohamad Shukor Mat Lazim, 180
Mohamed A Ismail, 392
Mohammed M. Nasef, 357
Mohd Bakri Adam, 1
Mohd Shafie Mustafa, 1
Mohsen Farid, 331

N

Nadim Ahmad, 99
Nazaria Baharudin, 180
Ngo The Hien, 366
Nikolaos Roubanis, 244
Nitin Kumar, 8
Norhaslinda Ali, 1
Nur Ayu Johar, 31
Nurul Fatin Shakira Helmi, 31
Nurul Nisa' Khairol Azmi, 1

O

Ossi Nurmi, 135

P

Pasi Piela, 135
Passant Elkafrawy, 357
Patricia Anne R. San Buenaventura, 311
Patrick Graham, 56
Pete Jones, 48
Peter van de Ven, 99
Philip Lane, 108
Purnendu Kumar, 8

Index

R
Rasha Elnemr, 357
Reem Ismail Mohamed Elsybaey, 286
Rense Lange, 39
Richard Finlay, 73
Ronald W. Jansen, 304
Ross Sparks, 194
S
Sachvinder Singh, 25
Sarpono, 127
Sarvnaz Karimi, 194
Shujian Xiang, 268
Siobhan Murray, 236
Stamatis Kalogirou, 401
Stefan Avdjiev, 108
Suman Raj Aryal, 323
Suzira Daud, 180
T
Talip Kilic, 236
Titi Kanti Lestari, 127
V
Vassilis P. Plagianakos, 201
W
Wenjun Wu, 268
Y
Yafei Wang, 261
Yakob Mudesir Seid, 253
Yingmei Xu, 268
Z
Zhan Li, 366

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 978-967-2000-66-2

9 789672 000662

#ISIWSC2019