STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 19, Number 1, March 2018

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, e-mail: m.wolan@stat.gov.pl
Secretary
Patryk Barszcz, e-mail: P.Barszcz@stat.gov.pl, phone number 00 48 22 — 608 33 66
Technical Assistant
Rajmund Litkowiec,

Address for correspondence

GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 1−3

FROM THE EDITOR

With this issue, the *Statistics in Transition new series* enters 25th anniversary of serving its mission as its first edition appeared in 1993 under the title *Statistics in Transition*. Although it took more than a decade the journal has developed to a widely acknowledged quarterly and its current version was formed about ten years ago under slightly extended title ('new series' was added for strictly formal reason during its repeated registration). The main goals of its initially defined mission remain firm and unchanged: to facilitate exchange of ideas and information amongst statisticians while contributing to building community of professionals, scholars and practitioners, world-wide. Since it happens that 25th anniversary of the *Statistics in Transition new series* takes place together with the 100th Anniversary of Statistics Poland being celebrated by the 2nd Congress of Polish Statistics to be held on July 10-12 in Warsaw, we are happy to announce that a special session will be organized during the congress to commemorate also this important moment in the journal development – read it below.

*

This issue is composed of XX articles distributed over the three sections, starting with Sampling Methods and Estimation, followed by Research Articles and Research Communicates.

In the paper entitled ***A Bayesian Inference of Multiple Structural Breaks in Mean and Error Variance in Panel AR (1) Model*****, Varun Agiwal, Jitendra Kumar,** and **Dahud Kehinde Shangodoyin** explore the effect of multiple structural breaks to estimate the parameters and test the unit root hypothesis in panel data time series model using Bayesian perspective. In particular, they obtain Bayes estimates for different loss functions using conditional posterior distribution, which is approximately explained by Gibbs sampling. For hypothesis testing, posterior odds ratio is calculated and solved via Monte Carlo Integration. The proposed methodology is illustrated with numerical examples. According to the authors, this model may be extended to panel AR (p) model with similar types of breaks as well as to VAR model.

In the next paper, ***Improved Rotation Patterns Using Two Auxiliary Variables in Successive Sampling*****, Jaishree Prabha Karna** and **Dilip Chandra Nath** discuss the role of two auxiliary variables on both the occasions to improve the precision of estimates at the current (second) occasion in two-occasion successive sampling. They use information on two auxiliary variables, which are positively correlated with the study variable, employing the exponential type structures and suggesting an efficient estimation procedure of population mean on the current (second) occasion. The proposed estimator has been compared with the sample mean estimator, when there is no matching from the previous occasion and natural successive sampling estimator. An optimal replacement strategy is also discussed along with justification of the use of the proposed sampling scheme. The conclusions are supported by results for real life data.

Seppo Laaksonen's and **Auli Hämäläinen's** paper on ***Joint Response Propensity and Calibration Method*** examines the chain of weights, beginning with the basic sampling weights for the respondents and converted next to reweights to reduce the bias due to missing quantities. In the case of availability of micro auxiliary variables for a gross sample, the authors suggest taking advantage of the response propensity weights, followed by the calibrated weights with macro (aggregate) auxiliary variables. They examined the calibration methodology that starts from the basic weights as well, employing simulated data for comparison. Eight indicators were examined and estimated leading to the main conclusion that the response propensity weights are the best starting weights for calibration.

The *Research Articles* section starts with the article by **Lahsen Bouchahed** and **Halim Zeghdoudi, *A New and Unified Approach in Generalizing the Lindley's Distribution with Applications*.** The authors propose a new family of continuous distributions with one extra shape parameter called the generalized Zeghdoudi distributions (GZD). They investigate the shapes of the density and hazard rate function, along with derivation of explicit expressions for some of its mathematical quantities. Various statistical properties like stochastic ordering, moment method, maximum likelihood estimation, entropies and limiting distribution of extreme order statistics are described. The results of the comparisons confirm the goodness of fit of GZ distribution.

Mirosław Krzyśko, Wojciech Łukaszonek and **Waldemar Wołyński** in the paper ***Canonical Correlation Analysis in the Case of Multivariate Repeated Measures Data*** present canonical variables applicable in the case of multivariate repeated measures data under the assumptions of (i) multivariate normality for the vector of observations and (ii) Kronecker product structure of the positive definite covariance matrix. These variables are especially useful when the number of observations is not large enough to estimate the covariance matrix, and thus the traditional canonical variables fail. Computational procedures for maximum likelihood estimates of required parameters are also provided.

Nicholas T. Longford's paper ***Searching for Causes of Necrotising Enterocolitis. An Application of Propensity Matching*** presents results of evaluation of the effect of changing the feeding regimen of infants in their first 14 postnatal days by analysing the data from the UK National Neonatal Research Database. The authors emphasize that they avoid some problems with drawing causal inferences from observational data by reducing the analysis to the infants who spent the first 14 postnatal days (or longer) in neonatal care and for whom NEC (necrotising enterocolitis, a disease of the gastrointestinal tract afflicting preterm-born infants) was not suspected in this period. Such limitation makes it possible to use summaries of the feeding regimen in this period as background variables in a potential outcomes framework. They emphasize the advantage of using a large size cohort and the usefulness of results for informing the design of a randomised clinical trial for preventing NEC, and the choice of its active treatment(s) in particular.

In the next article, Estimates of Trade Dependence of Ukraine: ***Indicators of International Trade Orientation of Ukraine in the Context of Assessment Of The Effectiveness of its Export Relations,*** **Nataliia Reznikova, Oleksandr Osaulenko** and **Volodymyr Panchenko** present the approach to analysing trade

relations between countries – especially trade dependence of Ukraine – by exploring economic vulnerability, economic sensitivity, symmetry and asymmetry of the established economic links. The estimated interdependence ratios for Ukraine and its largest trade partners – the EU, the Russian Federation, post-Soviet countries, China, the USA, Brazil and India – are compared with the respective ratios of Ukraine's dependence on these countries' markets. The analysed dynamics of Ukraine's GDP dependence on Ukraine's trade partners shows a growing relative weight of the countries that have not had a substantial role in the foreign trade of Ukraine. The proposed approach for estimating the quality of the established trade relations is supposed to contribute to transformation of Ukraine's foreign trade. The authors conclude that that the decreasing interdependence of partner countries, in parallel with establishing more diversified trade relations and/or reorientation to production of alternative goods/services with the respective growth in exports can be interpreted as a sign of economic development of a country.

In the paper ***Power Ishita distribution and its application to model lifetime data*** by **Kamlesh Kumar Shukla** and **Rama Shanker** the two-parameter power Ishita distribution (PID) is described and its important statistical properties – including shapes of the density, moments, skewness and kurtosis measures, hazard rate function, and stochastic ordering -are characterized, along with discussion of the maximum likelihood estimation of its parameters. An application of the distribution has been explained with a real lifetime data from engineering, and its goodness of fit shows better fit over two-parameter power Akash distribution (PAD), two-parameter power Lindley distribution (PLD) and one-parameter Ishita, Akash, Lindley and exponential distributions.

In the last section, Research Communicates and Letters, there are two papers related to different kinds of issues. In the first, **Jan Kordos** discusses ***Some results from the 2013 International Year of Statistics.*** Focusing on educational and other types of benefits brought about by the different conferences and workshops conducted occasionally to celebrate the International Year of Statistics, the author shares his observations on several challenges and problems involved in the development of statistics – as a science and as a field of experts' activities – taking the Workshop on the Future of Statistics as an example.

Shakti Prasad's article, ***Some product exponential methods of imputation in sample surveys,*** a product exponential method of imputation is presented together with discussion of a corresponding resultant point estimator proposed for estimating the population mean in sample surveys. The expression of bias and the mean square error of the suggested estimator has also been derived up to the first order of large sample approximations. The simulation studies show that the suggested estimator is the most efficient estimator.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 5

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and the Central Statistical Office of Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT-*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series.*

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl.,
GUS / Central Statistical Office
Al. Niepodległości 208, R. 287, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://stat.gov.pl/en/sit-en/guidelines-for-authors/

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 7–23, DOI 10.21307/stattrans-2018-001

A BAYESIAN INFERENCE OF MULTIPLE STRUCTURAL BREAKS IN MEAN AND ERROR VARIANCE IN PANEL AR (1) MODEL

Varun Agiwal[1], **Jitendra Kumar**[2], **Dahud Kehinde Shangodoyin**[3]

ABSTRACT

This paper explores the effect of multiple structural breaks to estimate the parameters and test the unit root hypothesis in panel data time series model under Bayesian perspective. These breaks are present in both mean and error variance at the same time point. We obtain Bayes estimates for different loss function using conditional posterior distribution, which is not coming in a closed form, and this is approximately explained by Gibbs sampling. For hypothesis testing, posterior odds ratio is calculated and solved via Monte Carlo Integration. The proposed methodology is illustrated with numerical examples.

Key words: panel data model, autoregressive model, structural break, MCMC, posterior odds ratio.

1. Introduction

Statistical inference of panel data time series model received great attention in the last several decades in both econometrics and statistics literature. The main idea behind the use of panel data time series model is to overcome the difficulty of unobserved variation in cross-section data sets over individual units as well as variation, which may change the structure also. It was assumed that this change was taken by some observations at a fixed and common time point in each series referred to as break point. Thus, structural break concept in panel data set-up is important to handle the permanent effects in the series and impacts other simultaneous variables. For this, an extensive literature concentrates on testing, estimation and detection of the existence of single or multiple structural breaks from univariate to multivariate time series. Bai and Perron (1998, 2003) considered the problem of estimation and testing for break point in linear model and determine the number of breaks using double maximum tests. They have also further addressed various issues such as estimation and testing number of

[1] Department of Statistics, Central University of Rajasthan, Bandersindri, Ajmer, India.
E-mail: varunagiwal.stats@gmail.com.
[2] Department of Statistics, Central University of Rajasthan, Bandersindri, Ajmer, India.
E-mail: vjitendrav@gmail.com.
[3] Department of Statistics, University of Botswana, Gaborone, Botswana.
E-mail: shangodoyink@gamil.com.

breaks, forming confidence interval related to multiple linear regression with multiple structural break. Altissimo and Corradi (2003) suggested an approach for detecting and estimation of the number of shifts in mean of a nonlinear process, which is having dependent and heterogeneous observations. They proposed a new estimator for long run variance, which was consistent in the presence of breaks and verified via a simulation exercise. Li (2004) applied quasi-Bayesian approach to detect the number and position of structural breaks in China's GDP and labour productivity data using predictive likelihood information criterion.

Apart from the above literature, which mainly dealt with the classical approach, a generalized form of estimation and testing the structural break by using Bayesian inference is less explored. Geweke and Jiang (2011) developed Bayesian approach to modelling in-sample structural breaks and forecasting out-of-sample breaks. Eo (2012) used Bayesian approach to estimate the number of breaks in autoregressive regressions with structural breaks in intercept, persistence, and residual variance. A model selection criterion was also considered to select the best model from U.S. GDP deflator data. Aue and Horvath (2013) discussed several approaches for estimating the parameter and locating multiple break points. They considered CUSUM procedure as well as likelihood statistic to adjust the serial dependence in presence of structural break. Recently, Melighotsiduo *et al.* (2017) suggested a Bayesian approach for autoregressive model allowing multiple structural changes in both mean and error variance of economic series occurring at unknown times, and Bayesian unit root testing is also proposed.

In current scenario, a growing literature on estimation and testing of multiple structural breaks in generalized univariate model such as panel data as well as multivariate time series model. A partial list of contributions in multiple structural breaks include Sugita (2006), Liu *et al.* (2011), Jin *et al.* (2013), Preuss *et al.* (2015) and Eo and Morley (2015) to analysis the procedure for detection and estimation of change point in vector error correction model, panel data model. In recent time, detection and estimation of multiple change points in panel data with interactive fixed effect and dynamic structure is introduced. Li *et al.* (2016) through penalized principal component (PPC) estimation procedure with an adaptive group fused LASSO.

An overview of the above description, this paper provides a general methodology to estimate and inference for panel data model under the presence of multiple change points in mean and error variance parameters. Our approach provides a flexible way to the interpretation of the result in real situation because in most economic and time series data are varying by trend and variance component. If one considers a break in mean also, then the impact of the series changes due to both type of break versus no break point. Thus, a Bayesian approach is introduced to capture the impact of break points in the panel data model. For Bayes estimation, we apply both symmetric and asymmetric loss function to posterior density in order to get better estimators and compare them with ordinary least square estimator. In addition, we also examine the model selection criterion to find the appropriate model, which may or may not contain multiple break points in a real data set.

2. Model Specification

Let $\{y_{it}$, t=1,2,…,T; i=1,2…,n$\}$ be a panel data time series model having *B* multiple break points in mean and error variance where breaks occur in both parameters in the same location. In that case our panel data model can be expressed as

$$y_{it} = \begin{cases} \rho y_{i,t-1} + (1-\rho)\mu_{i1} + \sigma_1 \varepsilon_{it}; & T_0 < t \le T_1 \\ \vdots & \vdots \\ \rho y_{i,t-1} + (1-\rho)\mu_{ij} + \sigma_j \varepsilon_{it}; & T_{j-1} < t \le T_j \\ \vdots & \vdots \\ \rho y_{i,t-1} + (1-\rho)\mu_{i,B+1} + \sigma_{B+1} \varepsilon_{it}; & T_B < t \le T_{B+1} = T \end{cases} \tag{1}$$

for j = 1,2,…, B and where n denotes number of cross sectional units, ρ is the autoregressive coefficient, μ_j is a (n × 1) vector of mean coefficients at j[th] division and ε_{it} are assumed to be independent and normally distributed with zero mean and division specific variance σ_j^2. This is a partial structural change model since the parameter ρ is not subject to shifts and is estimated using the entire sample space. The model in (1) can also be casted in the form of matrix notation with **.*** Kronecker delta product indicating element by element array multiplication, Z as the nT × (B+1) matrix whose j[th] column is equal to one if $T_{j-1} < t \le T_j$ and zero otherwise, and consider mean and residual variance parameters as a vector form.

$$y = \rho y_{-1} + (1-\rho) L\mu + \varepsilon \;; \quad \varepsilon \sim N(0, I_{nT} \cdot *S)$$

where

$$\underset{nT\times 1}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \underset{T\times 1}{y_i} = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix}; \underset{n(B+1)\times 1}{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{B+1} \end{pmatrix}; \underset{n\times 1}{\mu_j} = \begin{pmatrix} \mu_{1j} \\ \mu_{2j} \\ \vdots \\ \mu_{nj} \end{pmatrix}; \underset{(B+1)\times 1}{\sigma} = \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_{B+1} \end{pmatrix}$$

$$\underset{nT\times n(B+1)}{L} = \begin{pmatrix} I_n \otimes l_{T_1} & 0 & \cdots & 0 \\ 0 & I_n \otimes l_{T_2-T_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_n \otimes l_{T-T_B} \end{pmatrix}; Z = \begin{pmatrix} l_{T_1} & 0 & \cdots & 0 \\ 0 & l_{T_2-T_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{T-T_B} \end{pmatrix}; S = Z\sigma$$

Our study attempts to estimate the parameters in structural break model under Bayesian framework and test the unit root hypothesis by using posterior odds ratio. Under unit root case, model (1) reduces to a pure structural change model where all the model's coefficients are subject to change

$$\Delta y_{it} = \begin{cases} \sigma_1 \varepsilon_{it}; & T_0 < t \le T_1 \\ \vdots & \vdots \\ \sigma_j \varepsilon_{it}; & T_{j-1} < t \le T_j \\ \vdots & \vdots \\ \sigma_{B+1} \varepsilon_{it}; & T_B < t \le T_{B+1} \end{cases} \tag{2}$$

As mentioned, if we follow the usual approach defined in the literature to test for stationarity model reduces by (2) under the null hypothesis H_0: $\rho = 1$ is difference stationary with multiple breaks in error variance against the alternative hypothesis H_1: $\rho \in S$, series is stationary with multiple breaks in mean as well as error variance.

3. Bayesian Inference

In this section, we discuss issues related to the estimation and inference about the parameters and testing of unit root hypothesis. In order to perform Bayesian inference we need the likelihood function and specify prior distribution for the model parameters. Posterior probability is obtained by using sample information contained in the likelihood function combined with the joint prior distribution. The likelihood function for this model is

$$L(\rho,\mu,\sigma \mid y) = (2\pi)^{-\frac{nT}{2}} \prod_{j=1}^{B+1} \left(\sigma_j^{-n(T_j - T_{j-1})} \right) \exp\left[-\frac{1}{2} \sum_{j=1}^{B+1} \left\{ \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \left(y_{it} - \rho y_{i,t-1} - (1-\rho)\mu_{ij} \right)^2 \right\} \right] \tag{3}$$

For panel data model generally normal prior distribution is considered for $\mu_{ij} \sim N(\gamma_{ij}, \sigma_j^2)$, for error variance ($\sigma_j^2$) assume conjugate inverted gamma prior *IG(c_j , d_j)* and uniform prior is taken for autoregressive coefficient (ρ), see [Schotman and Van Dijk (1991) and Phillips (1991)]. The joint prior distribution is given as

$$\pi(\rho.\mu,\sigma^2) = \frac{(2\pi)^{-\frac{n(B+1)}{2}}}{1-l} \prod_{j=1}^{B+1} \left(\frac{d_j^{c_j}}{\Gamma c_j} \left(\sigma_j^2 \right)^{-c_j - 1 - \frac{n}{2}} \right) \exp\left[-\sum_{j=1}^{B+1} \frac{1}{\sigma_j^2} \left\{ d_j + \frac{1}{2} \sum_{i=1}^{n} \left(\mu_{ij} - \gamma_{ij} \right)^2 \right\} \right] \tag{4}$$

3.1. Bayesian Estimation via Gibbs Sampling

Given the likelihood function and prior density defined by eq^n (3) and eq^n (4), the posterior distribution is given by

$$\pi(\rho,\mu,\sigma\mid y)\propto\prod_{j=1}^{B+1}\left(\frac{d_j^{c_j}}{\Gamma c_j}\left(\sigma_j^2\right)^{-\left[\frac{n(T_j-T_{j-1}+1)}{2}+c_j+1\right]}\right)\exp\left[-\frac{1}{2}\sum_{j=1}^{B+1}\left\{\frac{1}{\sigma_j^2}\left(\sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-\rho y_{i,t-1}-(1-\rho)\mu_{ij}\right)^2+\sum_{i=1}^{n}\left(\mu_{ij}-\gamma_{ij}\right)^2+2d_j\right)\right\}\right] \tag{5}$$

The posterior distribution in (5) is very complicated and hence no closed form inference appears to be possible. For Bayesian estimation, we proceed via Gibbs sampler, a MCMC method, proposed by Geman and Geman (1984). The Gibbs sampler procedure, which we used, is described by Wang and Zivot (2000) in a time series regression model with multiple structural breaks. By means of this procedure, it gives a chain of estimated parameters values, which is frequently obtained by conditional probability distribution. Here, our aim is to generate a sequence of random variables from the conditional probability distribution using the current value of the parameters. For this we have derived the form of conditional posterior distributions given below:

$$\pi\left(\mu_{ij}\mid\rho,\sigma_j^2,\underline{y}\right)\sim N\left(\frac{(1-\rho)\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-\rho y_{i,t-1}\right)+\gamma_{ij}}{(1-\rho)^2\left(T_j-T_{j-1}\right)+1},\frac{\sigma_j^2}{(1-\rho)^2\left(T_j-T_{j-1}\right)+1}\right) \tag{6}$$

$$\pi\left(\sigma_j^2\mid\rho,\mu_{ij},\underline{y}\right)\sim IG\left(\frac{n\left(T_j-T_{j-1}\right)+n}{2}+c_j,d_j+\frac{1}{2}\sum_{i=1}^{n}\left(\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-\rho y_{i,t-1}-(1-\rho)\mu_{ij}\right)^2+\left(\mu_{ij}-\gamma_{ij}\right)^2\right)\right) \tag{7}$$

$$\pi\left(\rho\mid\mu_{ij},\sigma_j^2,\underline{y}\right)\sim TN\left(\sum_{j=1}^{B+1}\frac{\sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-\mu_{ij}\right)\left(y_{i,t-1}-\mu_{ij}\right)}{\sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it-1}-\mu_{ij}\right)^2},\sum_{j=1}^{B+1}\frac{\sigma_j^2}{\sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it-1}-\mu_{ij}\right)^2},l,1\right) \tag{8}$$

Using the generated samples from the above posteriors, Bayes estimates of the parameter are evaluated by different loss functions under Gibbs sampling algorithm. A loss function is a decision rule to select the best estimator and represent each of the possible estimates. Here, we consider squared error (symmetric) loss function as well as entropy (asymmetric) loss function for getting better understanding of the Bayesian estimation. Under squared error and entropy loss function, Bayes estimator are $E(\theta\mid\underline{x})$ and $\left[E\left(\theta^{-1}\mid\underline{x}\right)\right]^{-1}$.

3.2. Testing Unit Root Hypothesis via Posterior Odds Ratio

In a hypothesis testing problem, one is generally interested in testing the stationary condition of a model. Here, null hypothesis is used as a unit root hypothesis against the alternative of a stationary model. In Bayesian framework, testing is often convenient to summarize the information in terms of posterior odds ratio. The posterior odds ratio is the ratio of posterior probability under null versus alternative hypothesis with the product of prior odds, notation given as:

$$\beta_{01} = \frac{p_0}{1-p_0}\frac{P(y \mid H_0)}{P(y \mid H_1)} \tag{9}$$

Theorem: To test the null hypothesis that y_{it} is a non-stationary I(1) process, i.e. $\rho=1$ in equation (2), against the alternative hypothesis that y_{it} is a stationary I(0) process, i.e. $\rho \in S$ in equation (1). The posterior odds ratio can be constructed according to equation (9)

$$\beta_{01} = \frac{p_0}{1-p_0}\frac{\prod_{j=1}^{B+1}\left[d_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-y_{i,t-1}\right)^2\right]^{-\frac{n\left(T_j-T_{j-1}\right)+2c_j}{2}}}{\frac{1}{1-l}\int_l^1\prod_{j=1}^{B+1}\left[A_j(\rho)\right]^{-\frac{n}{2}}\left[C_j(\rho)\right]^{-\frac{n\left(T_j-T_{j-1}\right)+2c_j}{2}}d\rho} \tag{10}$$

where

$$A_j(\rho) = (1-\rho)^2\left(T_j - T_{j-1}\right)+1$$

$$B_{ij}(\rho) = (1-\rho)\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-y_{i,t-1}\right)+\gamma_{ij}$$

$$C_j(\rho) = d_j + \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-y_{i,t-1}\right)^2 + \gamma_{ij}^2 - \frac{\left[B_{ij}(\rho)\right]^2}{A_j(\rho)}\right)$$

Proof: The proof of the theorem is given in the appendix.

In the equation (10), closed form expression of posterior odds ratio is not obtained. Therefore, we use an alternative technique as Monte Carlo integration for approximately solving the integrals and get the value of posterior odds ratio.

4. Simulation Study

In this section, we conduct a set of simulated experiments to evaluate the performance of our model and compare different estimators based on Monte Carlo simulation. To estimate the model parameters, assume that the number of breaks and the location of break points are known so that the remaining objectives in equation (1) are estimated via an iterative procedure. In simulation experiment we have generated artificial time series from our model with varying

numbers of structural breaks at the same time points in mean and error variance parameters. We are starting with the initial observation y_{0i} = (10, 15, 20) to generate panel data time series from the suggested model having three panel (n=3) and each panel contains T observations. For better interpretation, we took different size of time series T = (50, 75, 100) and also varying autoregressive coefficients ρ = (0.9, 0.95, 0.99). The number of possible structural break (B) has been 3. Thus, the disturbances ε_{it} are generated as i.i.d. for all *i* and *j* with four different variance, namely $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (0.1, 0.2, 0.3, 0.4)$. For inverse gamma prior distribution with hyper parameters is to known. For numerical purpose we have taken as $c_j = d_j = 0.01$ for all break points. In the case of normal prior, hyper prior mean is equal to mean of the generated series at every break point interval (T_{j-1}, T_j) with parallel variance given in disturbances term. The true value of mean term for each panel having four partitions is written as (μ_{11}, μ_{12}, μ_{13}) = (14, 16, 18); (μ_{21}, μ_{22}, μ_{23}) = (20, 22, 24); (μ_{31}, μ_{32}, μ_{33}) = (26, 28, 30); (μ_{41}, μ_{42}, μ_{43}) = (32, 34, 36). All results are based on 5000 replications. From the generated sample, we obtained Bayes estimate of parameters and compared the performance with ordinary least square (OLS) estimate. We report the estimated value and its mean square error in Table-4.1 to 4.3.

Table-4.1. Estimators with varying time series at ρ = 0.9

	T=50			T=75			T=100		
	OLS	SELF	ELF	OLS	SELF	ELF	OLS	SELF	ELF
ρ	0.8607	0.9037	0.9036	0.8724	0.9049	0.9048	0.8763	0.9043	0.9043
	0.0025	0.0000	0.0000	0.0012	0.0001	0.0001	0.0009	0.0001	0.0001
μ_{11}	13.7381	13.9853	13.9785	13.8519	13.9959	13.9893	13.7567	13.9815	13.9749
	0.5840	0.0104	0.0107	0.5529	0.0098	0.0100	0.5706	0.0091	0.0094
μ_{12}	16.1184	15.9989	15.9929	16.0870	15.9938	15.9880	16.1338	16.0026	15.9969
	0.7121	0.0124	0.0125	0.5515	0.0106	0.0107	0.6236	0.0110	0.0111
μ_{13}	18.2558	17.9770	17.9717	18.2331	17.9833	17.9782	18.2917	17.9964	17.9912
	0.7394	0.0104	0.0106	0.7099	0.0108	0.0110	0.7042	0.0100	0.0101
μ_{21}	18.9938	20.0110	20.0014	19.4758	20.0186	20.0098	19.6677	20.0093	20.0015
	2.6156	0.0100	0.0099	1.2051	0.0235	0.0234	0.6942	0.0386	0.0385
μ_{22}	20.8791	21.9925	21.9838	21.4667	22.0115	22.0035	21.7912	22.0295	22.0224
	2.6572	0.0151	0.0154	1.5227	0.0293	0.0292	0.6849	0.0359	0.0356
μ_{23}	23.1734	24.0115	24.0035	23.5620	24.0168	24.0094	23.7996	24.0274	24.0209
	2.2536	0.0116	0.0114	1.2840	0.0278	0.0275	0.5568	0.0325	0.0322
μ_{31}	24.9505	26.0408	26.0295	25.5998	26.0509	26.0410	25.7780	26.0333	26.0244
	4.4370	0.0257	0.0248	1.7431	0.0381	0.0373	0.7004	0.0381	0.0376
μ_{32}	26.7468	28.0131	28.0026	27.6054	28.0537	28.0445	27.7914	28.0448	28.0364
	4.8589	0.0230	0.0228	1.7379	0.0385	0.0377	0.9369	0.0557	0.0550
μ_{33}	28.6660	30.0059	29.9961	29.5123	30.0316	30.0230	29.6804	30.0104	30.0027
	5.1925	0.0218	0.0218	1.7849	0.0359	0.0355	0.8159	0.0444	0.0443
μ_{41}	30.9830	32.0348	32.0234	31.1368	31.9948	31.9832	31.4541	32.0210	32.0093
	3.6570	0.0363	0.0358	3.9274	0.0448	0.0450	2.8567	0.0384	0.0382
μ_{42}	32.9950	34.0339	34.0232	33.5021	34.0500	34.0391	33.6176	34.0352	34.0243
	4.4146	0.0376	0.0369	2.5846	0.0370	0.0361	2.8885	0.0380	0.0374
μ_{43}	34.8637	36.0233	36.0132	35.1567	36.0032	35.9929	35.4250	36.0242	36.0139
	3.7348	0.0342	0.0340	3.4651	0.0386	0.0388	2.5830	0.0353	0.0349
σ_1^2	0.1071	0.1085	0.0988	0.1042	0.1053	0.1014	0.1038	0.1046	0.1005
	0.0007	0.0007	0.0000	0.0004	0.0005	0.0000	0.0005	0.0005	0.0000
σ_2^2	0.2104	0.2094	0.1961	0.2065	0.2069	0.2052	0.2031	0.2037	0.2030
	0.0040	0.0034	0.0000	0.0017	0.0015	0.0001	0.0006	0.0006	0.0000
σ_3^2	0.3121	0.3189	0.3028	0.3002	0.3031	0.2920	0.3054	0.3060	0.2984
	0.0075	0.0071	0.0003	0.0034	0.0034	0.0001	0.0018	0.0018	0.0000
σ_4^2	0.4058	0.4116	0.4027	0.4140	0.4193	0.3987	0.4252	0.4241	0.4221
	0.0055	0.0055	0.0004	0.0078	0.0078	0.0002	0.0100	0.0094	0.0011

Table-4.2. Estimators with varying time series at $\rho = 0.95$

	T=50			T=75			T=100		
	OLS	SELF	ELF	OLS	SELF	ELF	OLS	SELF	ELF
ρ	0.9277	0.9511	0.9511	0.9239	0.9511	0.9510	0.9231	0.9515	0.9515
	0.0010	0.0000	0.0000	0.0012	0.0000	0.0000	0.0011	0.0000	0.0000
μ_{11}	14.2813	14.0055	13.9979	14.3639	14.0031	13.9955	14.2784	14.0008	13.9934
	2.8644	0.0026	0.0026	2.8140	0.0026	0.0027	2.4136	0.0021	0.0021
μ_{12}	17.0431	16.0055	15.9989	16.8423	15.9981	15.9916	17.0321	16.0007	15.9942
	4.9016	0.0027	0.0027	4.3718	0.0029	0.0029	3.7536	0.0022	0.0022
μ_{13}	19.2725	17.9970	17.9910	19.2403	17.9869	17.9811	19.5766	17.9947	17.9889
	5.8300	0.0026	0.0027	5.7341	0.0021	0.0022	6.0293	0.0021	0.0022
μ_{21}	19.1892	20.0027	19.9902	19.1421	20.0079	19.9981	19.0485	20.0086	19.9988
	18.5998	0.0029	0.0030	6.4266	0.0088	0.0087	3.9955	0.0082	0.0082
μ_{22}	21.1867	21.9990	21.9876	21.3103	22.0017	21.9928	21.2755	21.9997	21.9907
	10.5883	0.0025	0.0027	4.6515	0.0079	0.0080	3.2040	0.0079	0.0080
μ_{23}	23.8031	24.0070	23.9966	23.6473	24.0044	23.9962	23.7322	24.0062	23.9980
	11.7370	0.0029	0.0028	3.8351	0.0074	0.0074	3.0148	0.0089	0.0088
μ_{31}	23.5217	25.9969	25.9830	23.9895	25.9922	25.9803	24.5652	26.0025	25.9912
	30.2210	0.0041	0.0043	9.4940	0.0073	0.0076	6.4169	0.0145	0.0146
μ_{32}	26.4880	28.0077	27.9947	26.7673	28.0165	28.0056	26.7426	28.0081	27.9977
	22.0991	0.0037	0.0037	9.3507	0.0119	0.0117	5.5741	0.0151	0.0151
μ_{33}	28.4393	29.9977	29.9856	28.7290	30.0130	30.0027	28.7848	30.0002	29.9904
	25.9163	0.0038	0.0039	7.4535	0.0120	0.0119	5.2416	0.0146	0.0148
μ_{41}	29.9873	31.9863	31.9745	30.8228	32.0145	32.0025	31.0121	32.0101	31.9987
	10.5213	0.0195	0.0201	8.4309	0.0336	0.0335	4.5547	0.0305	0.0305
μ_{42}	32.9922	34.0512	34.0402	32.9787	34.0263	34.0151	33.0487	34.0116	34.0008
	8.6453	0.0235	0.0224	5.3532	0.0246	0.0242	4.1663	0.0297	0.0296
μ_{43}	34.7339	36.0139	36.0035	35.2679	36.0439	36.0333	35.0919	36.0133	36.0032
	8.8365	0.0193	0.0192	5.1718	0.0286	0.0278	4.1124	0.0292	0.0291
σ_1^2	0.1087	0.1058	0.1019	0.1077	0.1048	0.1017	0.1065	0.1047	0.0999
	0.0011	0.0010	0.0000	0.0007	0.0007	0.0000	0.0008	0.0008	0.0000
σ_2^2	0.2566	0.2515	0.2244	0.1889	0.2040	0.1978	0.2057	0.2030	0.2000
	0.0126	0.0100	0.0019	0.0016	0.0017	0.0002	0.0015	0.0014	0.0000
σ_3^2	0.3739	0.3679	0.3372	0.3182	0.3137	0.2980	0.3094	0.3062	0.3036
	0.0265	0.0211	0.0016	0.0049	0.0048	0.0003	0.0026	0.0026	0.0000
σ_4^2	0.3961	0.4001	0.3946	0.4132	0.4091	0.4032	0.3940	0.4016	0.3984
	0.0043	0.0043	0.0002	0.0035	0.0034	0.0001	0.0024	0.0024	0.0000

Table-4.3. Estimators with varying time series at $\rho = 0.99$

	T=50			T=75			T=100		
	OLS	SELF	ELF	OLS	SELF	ELF	OLS	SELF	ELF
ρ	0.9887	0.9900	0.9900	0.9896	0.9900	0.9900	0.9896	0.9900	0.9900
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
μ_{11}	14.0541	13.9998	13.9995	13.8541	13.9996	13.9993	13.9496	13.9991	13.9988
	2.5280	0.0000	0.0000	2.2032	0.0000	0.0000	1.4473	0.0000	0.0000
μ_{12}	15.7119	15.9994	15.9992	16.1693	16.0004	16.0002	16.1484	16.0006	16.0004
	2.7599	0.0000	0.0000	1.7276	0.0000	0.0000	1.9185	0.0000	0.0000
μ_{13}	18.0323	18.0016	18.0014	17.9392	18.0001	17.9998	18.0924	18.0009	18.0006
	2.8263	0.0000	0.0000	2.4351	0.0001	0.0001	2.0293	0.0000	0.0000
μ_{21}	20.0894	20.0014	20.0011	20.1550	20.0008	20.0005	19.7357	19.9988	19.9985
	15.3991	0.0001	0.0001	5.6088	0.0001	0.0001	4.1294	0.0001	0.0001

Table-4.3. Estimators with varying time series at $\rho = 0.99$ (cont.)

	T=50			T=75			T=100		
	OLS	SELF	ELF	OLS	SELF	ELF	OLS	SELF	ELF
μ_{22}	22.1373	21.9999	21.9996	21.8658	22.0008	22.0005	22.0746	22.0001	21.9998
	17.3947	0.0001	0.0001	6.3083	0.0001	0.0001	3.4212	0.0001	0.0001
μ_{23}	24.1136	24.0001	23.9998	23.8898	23.9995	23.9992	24.1808	24.0002	23.9999
	20.6617	0.0001	0.0001	3.6968	0.0001	0.0001	4.7187	0.0001	0.0001
μ_{31}	25.4894	26.0006	26.0002	25.9236	26.0004	26.0001	25.8274	25.9991	25.9988
	50.4575	0.0001	0.0001	9.9667	0.0001	0.0001	9.8900	0.0001	0.0001
μ_{32}	27.7662	27.9993	27.9991	27.8165	28.0000	27.9997	27.5075	28.0013	28.0010
	37.2489	0.0001	0.0001	15.0037	0.0001	0.0001	11.5979	0.0001	0.0001
μ_{33}	29.7095	30.0006	30.0003	29.7752	30.0009	30.0006	30.0822	29.9988	29.9985
	36.9404	0.0001	0.0001	12.4540	0.0001	0.0001	7.6263	0.0001	0.0001
μ_{41}	31.7953	32.0001	31.9998	32.1870	32.0006	32.0004	31.7316	31.9987	31.9985
	72.6862	0.0001	0.0001	7.7369	0.0001	0.0001	2.9958	0.0001	0.0001
μ_{42}	33.7731	34.0002	33.9999	33.9305	34.0000	33.9998	33.8511	34.0027	34.0024
	72.3302	0.0001	0.0001	6.8616	0.0001	0.0001	2.8177	0.0001	0.0001
μ_{43}	35.9221	35.9988	35.9985	35.9559	35.9998	35.9995	35.8189	36.0004	36.0002
	75.6086	0.0001	0.0001	6.9744	0.0001	0.0001	2.9922	0.0001	0.0001
σ_1^2	0.1191	0.1088	0.1006	0.1221	0.1121	0.0993	0.1153	0.1122	0.1017
	0.0024	0.0020	0.0000	0.0034	0.0017	0.0003	0.0024	0.0016	0.0001
σ_2^2	0.2396	0.2082	0.1962	0.2306	0.2288	0.1845	0.2405	0.2236	0.2025
	0.0088	0.0077	0.0002	0.0097	0.0093	0.0003	0.0122	0.0071	0.0001
σ_3^2	0.3452	0.3208	0.2999	0.3418	0.3215	0.2642	0.3502	0.3313	0.2817
	0.0172	0.0123	0.0002	0.0227	0.0139	0.0018	0.0253	0.0207	0.0009
σ_4^2	0.3953	0.4094	0.3891	0.3920	0.4080	0.4013	0.4067	0.4005	0.4011
	0.0036	0.0032	0.0004	0.0017	0.0018	0.0000	0.0012	0.0012	0.0001

Table-4.1-4.3 shows the behaviour of ordinary least square (OLS) and Bayes estimators of parameter with varying values of time series at different autoregressive coefficient. It can be easily seen that MSEs of all estimators decrease as the sample size of series increases. The difference between the estimated value and the true value in OLS is large, which explains that average bias is maximum as compared to Bayes estimator. The Bayes estimator under both loss functions performs better because of additional information given about the parameter. A better estimated value for autoregressive coefficient, mean term and error variance is obtained by ELF as compared to SELF due to less MSE for low value of ρ. For a high value of ρ, Bayes estimator obtained under different loss function is equally applicable to estimate the parameters since both the estimators show more or less same magnitudes for their MSE. An increase in the number of breaks points, MSE also increases because of the length of the segment is small and take less observation to estimate the parameters.

After estimation of structural break parameters, testing of the unit root is considered. We can calculate posterior odds ratio values with different values of ρ and varying size of the series, which are reported in Table-4.4. The table shows

that as the value of ρ increases, POR reduces to zero and this rejects the null hypothesis, i.e. unit root hypothesis. Thus, the model which contains multiple breaks in mean and variance have a stationary model for this simulated series.

Table-4.4. Posterior odds ratio with varying ρ and T

T	ρ=0.90	ρ=0.92	ρ=0.94	ρ=0.96	ρ=0.98
50	0.0750	0.0416	0.0245	0.0098	0.0037
60	0.0645	0.0412	0.0291	0.0118	0.0044
70	0.0723	0.0447	0.0316	0.0149	0.0060
80	0.0647	0.0441	0.0338	0.0159	0.0067

5. Real Data Analysis

To provide a practical application of our model and verify the result obtained by simulation study, we apply our proposed work to a real data set. We use agricultural production and productivity of various crops of food grains data set consisting of 60 years' time series of Rice (R), Wheat (W) and Coarse Cereals (CC) variables for the annually book of "*Handbook of Statistics on Indian Economy*" from 1954-55 to 2014-15. The source of food grains data set is taken through Ministry of Agriculture & Farmers Welfare, Government of India by Reserve Bank of India and the book was published by Data Management and Dissemination Division (DMDD), Department of Statistics and Information Management (DSIM), Reserve Bank of India (RBI). This book provides statistical data on a wide range of economic and financial indicators related to national income variable, output, prices, money, banking, financial markets, etc. To determine the number of break points and their positions in food grains data set, we can use "*strucchange*" package developed by Zeileis *et al.* (2002) in R-language. The command "*breakpoints*" is considered to suggest the number of break points and identify their location in respective individual series. The results are reported in Table-5.1, which is given below:

Table-5.1: Structural break point present in different cereals

Break Point	Rice	Wheat	Coarse Cereals
T_1	12	12	12
T_2	22	22	33
T_3	33	30	51
T_4	42	39	
T_5	51	51	

From the above table one can observe that each series could not contain equal number of break points and their positions also differ from one series to another because of finding the break points individually. After applying the procedure we observe that each series includes two similar break points 12 and 51, which is near and far from the series. The remaining break points (22, 30, 33, 39, 42) mainly occur in between these points. For analysis, we make different

combinations of break points containing various numbers of breaks. There are seven break points established; 127 combinations included single as well as multiple breaks. To identify a suitable model for this data set by using log-likelihood function, Akaike information criterion (AIC) and Bayes information criterion (BIC) define how many break points and their location is present in the model. For examination, consider only starting eight combinations, which have minimum AIC and BIC values, as shown in the Table-5.2.

Table-5.2. Break point detection

Break point	POR	ρ	AIC	BIC	Log L
(22,42,51)	0.00878	0.9803	1433.0263	1474.3162	-703.5131
(22,30,51)	0.01617	0.9778	1454.7726	1496.0625	-714.3863
(22,30,42,51)	0.00362	0.9732	1460.8775	1511.6959	-714.4387
(22,42)	0.06997	0.9808	1471.8446	1513.6061	-725.9223
(12,30,51)	0.01872	0.9749	1477.9163	1519.2062	-725.9581
(12,30,42,51)	0.00745	0.9693	1487.6784	1538.4968	-727.8392
22	0.51915	0.9881	1524.1465	1546.3796	-755.0733
(22,30,42)	0.01195	0.9738	1524.5874	1565.8773	-749.2937

By using information criterion, Table-5.2 gives appropriate conclusion about the number of break points and their positions to obtain a suitable model for this data set. The table shows that data follow a PAR(1) model having three break points (22, 42, 51) because of minimum AIC and BIC. Considering higher number of break points, i.e. 5, 6 or 7 in the series, no combination occurs in the last eight observations because the increase in the break point is inconvenient to partition the series in small segments. The maximum number of break points is 4 in this table, which can be considered after 3 break points. When the break point position occurs at only single point, i.e. 22, then AIC and BIC have large value as compared to two break points (22, 42). If we think about these two break points, AIC and BIC values may or may not be larger than three or four break points at different locations. If positions occur mostly at 22 and 51, statistic values is minimum in our combination so that mean ${}^{7}C_{i}$, I = 1 to 7. Minimum difference between break points may increase the statistic values, which directly reject these points of the model. An increase in the number of break points, posterior odds ratio value tends to zero, which concludes our model, which contains multiple breaks in mean and variance, is a better model compared to no-break model, i.e. PAR(1) model. The table also shows that data series is a stationary series for any combination of break point considered in the model. As a break point increases, the value of POR tends to zero, which concludes the model is stationary and no unit root is present in the model to make this as a difference stationary.

Once we acquire the number of break points and their positions in the proposed model, use this to estimate the parameters of the model for the data set using Gibbs procedure. The results are summarized in the Table-5.3. The table gives an appropriate value of the estimate parameter for the food grains data set in different types of estimation technique.

Table-5.3. Estimates Value Using Real Data Set

Parameter	OLS	SELF	ELF
ρ	0.7786	0.7788	0.7788
μ_{11}	42.6226	41.6685	41.6810
μ_{12}	21.3355	20.8544	20.8631
μ_{13}	26.8882	26.2820	26.2926
μ_{21}	72.9170	71.2050	71.2258
μ_{22}	57.7628	56.4043	56.4220
μ_{23}	31.0778	30.3408	30.3538
μ_{31}	94.2339	91.1963	91.2235
μ_{32}	77.5137	75.0114	75.0358
μ_{33}	34.8718	33.7356	33.7521
μ_{41}	109.4301	105.7798	105.8104
μ_{42}	95.1015	91.9243	91.9550
μ_{43}	46.2301	44.6730	44.6951
σ_1^2	12.9941	12.9820	12.9262
σ_2^2	25.5067	25.4830	25.3736
σ_3^2	51.1556	51.1081	50.8886
σ_4^2	25.9380	25.9139	25.8027

6. Conclusions

This paper deals with multiple structural breaks, which are present in mean and error variance in panel AR (1) model. Bayesian framework is used for estimating and testing the unit root hypothesis. Bayesian estimator gives better estimated value of the parameter as compared to OLS estimator in simulation as well as in real data. Testing of the hypothesis gives appropriate conclusion about the simulated series, which is stationary, and this is also verified by the real data set at each combination of break points. Break point identification is also done in real data set by using information criterion. This model may be extended to panel AR (p) model with similar types of breaks as well as to VAR model.

Acknowledgement

The second author gratefully acknowledges the financial assistance from UGC, India under MRP Scheme (Grant No.42-43/2013).

REFERENCES

ALTISSIMO, F., CORRADI, V., (2003). Strong rules for detecting the number of breaks in a time series. Journal of Econometrics, 117 (2), pp. 207–244.

AUE, A., HORVÁTH, L., (2013). Structural breaks in time series. Journal of Time Series Analysis, 34 (1), pp. 1–16.

BAI, J., PERRON, P., (1998). Estimating and testing linear models with multiple structural changes. Econometrica, 66 (1), pp. 47–78.

BAI, J., PERRON, P., (2003). Computation and analysis of multiple structural change models. Journal of Applied Econometrics, 18 (1), pp. 1–22.

EO, Y., MORLEY, J., (2015). Likelihood-based confidence sets for the timing of structural breaks. Quantitative Economics, 6, pp. 463–497.

EO, Y., (2012). Bayesian Inference about the Types of Structural Breaks When There are Many Breaks. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2011825

GEMAN, S., GEMAN, D., (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, pp. 721–741.

GEWEKE, J., JIANG, Y., (2011). Inference and prediction in a multiple-structural-break model. Journal of Econometrics, 163 (2), pp. 172–185.

JIN, B., SHI, X., WU, Y., (2013). A novel and fast methodology for simultaneous multiple structural break estimation and variable selection for nonstationary time series models. Statistics and Computing, 23 (2), pp. 1–11.

LI, D., QIAN, J., SU, L., (2016). Panel data models with interactive fixed effects and multiple structural breaks. Journal of the American Statistical Association, 111 (516), pp. 1804–1819.

LI, X., (2004). A Quasi-Bayesian Analysis of Structural Breaks in China's Output and Productivity Series. International Journal of Business and Economics, 3 (1), pp. 57–65.

LIU, D., LI, R., WANG, Z., (2011). Testing for structural breaks in panel varying coefficient models: with an application to OECD health expenditure. Empirical Economics, 40 (1), pp. 95–118.

MELIGKOTSIDOU, L., TZAVALIS, E., VRONTOS, I. D., (2017). On Bayesian Analysis and Unit Root Testing for Autoregressive Models in the Presence of Multiple Structural Breaks. Econometrics and Statistics, 4, pp.70–90.

PHILLIPS, P.C., (1991) To criticize the critics: An objective Bayesian analysis of stochastic trends. Journal of Applied Econometrics, 6, pp.333–364.

PREUSS, P., PUCHSTEIN, R., DETTE, H., (2015). Detection of multiple structural breaks in multivariate time series. Journal of the American Statistical Association, 110 (510), pp. 654–668.

SCHOTMAN, P. C., VAN DIJK, H. K., (1991). A Bayesian routes to unit roots. Journal of Applied Econometrics, 6, pp. 387–401.

SUGITA, K., (2006). Bayesian Analysis of Dynamic Multivariate Models with Multiple Structural Breaks, No 2006-14, Discussion Papers, Graduate School of Economics, Hitotsubashi University, http://EconPapers.repec.org/RePEc:hit:econdp:2006-14.

WANG, J., ZIVOT, E., (2000). A Bayesian time series model of multiple structural changes in level, trend, and variance. Journal of Business & Economic Statistics, 18, pp. 374–386.

ZEILEIS, F., LEISCH, K., H., KLEIBER, C., (2002). Strucchange: An R package for testing for structural change in linear regression models. Journal of Statistical Software, 7, pp. 1–38.

APPENDIX

In this appendix, we have derived the posterior probability with the help of likelihood function and prior distribution, which are given below:

(A.1) For alternative hypothesis $H_1 : \rho \in S, \mu_{i1} \neq \mu_{i2}, \sigma_1^2 \neq \sigma_2^2$, expression of posterior probability can be derived with the help of likelihood function (3) and prior distribution given (4), which is given as

$$P(y \mid H_1)$$

$$= \int_{l}^{1} \int_{R_{B+1}^{+}} \int_{R_{n(B+1)}} L(\rho, \mu, \sigma \mid y)\, p(\mu) p(\sigma) p(\rho) d\mu d\sigma d\rho$$

$$= \int_{l}^{1} \int_{R_{B+1}^{+}} \int_{R_{n(B+1)}} (2\pi)^{-\frac{nT}{2}} \prod_{j=1}^{B+1} \left(\sigma_j^{-n(T_j - T_{j-1})} \right) \exp\left[-\frac{1}{2} \sum_{j=1}^{B+1} \left\{ \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \left(y_{it} - \rho y_{i,t-1} - (1-\rho)\mu_{ij} \right)^2 \right\} \right]$$

$$\frac{d_1^{c_1}}{\Gamma c_1} \frac{d_2^{c_2}}{\Gamma c_2} \cdots \frac{d_{B+1}^{c_{B+1}}}{\Gamma c_{B+1}} \left(\sigma_1^2 \right)^{-c_1 - 1} \left(\sigma_2^2 \right)^{-c_{21} - 1} \cdots \left(\sigma_{B+1}^2 \right)^{-c_{B+1} - 1} \exp\left[-\frac{d_1}{\sigma_1^2} \cdots - \frac{d_{B+1}}{\sigma_{B+1}^2} \right] (2\pi)^{-\frac{n(B+1)}{2}}$$

$$\sigma_1^{-n} \cdots \sigma_{B+1}^{-n} \exp\left[-\frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (\mu_{i1} - \gamma_{i1})^2 \cdots - \frac{1}{2\sigma_{B+1}^2} \sum_{i=1}^{n} (\mu_{i,B+1} - \gamma_{i,B+1})^2 \right] d\mu d\sigma d\rho$$

$$= \int_{l}^{1} \int_{R_{B+1}^{+}} \int_{R_{n(B+1)}} \frac{(2\pi)^{-\frac{n(T+B+1)}{2}}}{1-l} \prod_{j=1}^{B+1} \left(\frac{d_j^{c_j}}{\Gamma c_j} \left(\sigma_j^2 \right)^{-\left[\frac{n(T_j - T_{j-1}+1)}{2} + c_j + 1 \right]} \right) \exp\left[-\frac{1}{2} \sum_{j=1}^{B+1} \left\{ \frac{1}{\sigma_j^2} \left(\sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \left(y_{it} - \rho y_{i,t-1} \right)^2 \right.\right.\right.$$

$$- 2(1-\rho) \sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \left(y_{it} - \rho y_{i,t-1} \right) \mu_{ij} + (1-\rho)^2 \left(T_j - T_{j-1} \right) \sum_{i=1}^{n} \mu_{ij}^2 + \sum_{i=1}^{n} \mu_{ij}^2 - 2 \sum_{i=1}^{n} \gamma_{ij} \mu_{ij}$$

$$\left.\left.\left. + \sum_{i=1}^{n} \gamma_{ij}^2 + 2d_j \right) \right\} \right] d\mu d\sigma d\rho$$

$$= \int_{l}^{1} \int_{R_{B+1}^{+}} \int_{R_{n(B+1)}} \frac{(2\pi)^{-\frac{n(T+B+1)}{2}}}{1-l} \prod_{j=1}^{B+1} \left(\frac{d_j^{c_j}}{\Gamma c_j} \left(\sigma_j^2 \right)^{-\left[\frac{n(T_j - T_{j-1}+1)}{2} + c_j + 1 \right]} \right) \exp\left[-\frac{1}{2} \sum_{j=1}^{B+1} \left\{ \frac{1}{\sigma_j^2} \left(\sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \left(y_{it} - \rho y_{i,t-1} \right)^2 \right.\right.\right.$$

$$+ 2d_j + \sum_{i=1}^{n} \gamma_{ij}^2 + \left((1-\rho)^2 \left(T_j - T_{j-1} \right) + 1 \right) \sum_{i=1}^{n} \mu_{ij}^2$$

$$\left.\left.\left. - 2 \sum_{i=1}^{n} \left((1-\rho) \sum_{t=T_{j-1}}^{T_j} \left(y_{it} - \rho y_{i,t-1} \right) + \gamma_{ij} \right) \mu_{ij} \right) \right\} \right] d\mu d\sigma d\rho$$

Let us consider

$$A_j = (1-\rho)^2\left(T_j - T_{j-1}\right)+1$$

$$B_{ij} = (1-\rho)\sum_{t=T_{j-1}}^{T_j}\left(y_{it} - \rho y_{i,t-1}\right)+\gamma_{ij}$$

Then we may write

$$P(y \mid H_1)$$

$$= \int_l^1 \int_{R_{B+1}^+} \int_{R_{n(B+1)}} \frac{(2\pi)^{-\frac{n(T+B+1)}{2}}}{1-l} \prod_{j=1}^{B+1}\left(\frac{d_j^{c_j}}{\Gamma c_j}\left(\sigma_j^2\right)^{-\left[\frac{n\left(T_j - T_{j-1}+1\right)}{2}+c_j+1\right]}\right) \exp\left[-\frac{1}{2}\sum_{j=1}^{B+1}\left\{\frac{1}{\sigma_j^2}\left(\sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-\rho y_{i,t-1}\right)^2 + \sum_{i=1}^{n}\gamma_{ij}^2 + 2d_j + A_j\sum_{i=1}^{n}\left(\mu_{ij} - \frac{B_{ij}}{A_j}\right)^2 - \sum_{i=1}^{n}\frac{B_{ij}^2}{A_j}\right)\right\}\right] d\mu\, d\sigma\, d\rho$$

$$= \int_l^1 \int_{R_{B+1}^+} \frac{(2\pi)^{-\frac{nT}{2}}}{1-l} \prod_{j=1}^{B+1}\left(\frac{d_j^{c_j}\left(\sigma_j^2\right)^{-\left[\frac{n\left(T_j - T_{j-1}\right)}{2}+c_j+1\right]}}{\left(A_j\right)^{\frac{n}{2}}\Gamma c_j}\right) \exp\left[-\frac{1}{2}\sum_{j=1}^{B+1}\left\{\frac{1}{\sigma_j^2}\left(\sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-\rho y_{i,t-1}\right)^2 + \sum_{i=1}^{n}\gamma_{ij}^2 + 2d_j - \sum_{i=1}^{n}\frac{B_{ij}^2}{A_j}\right)\right\}\right] d\sigma\, d\rho$$

$$= \int_l^1 \int_{R_{B+1}^+} \frac{(2\pi)^{-\frac{nT}{2}}}{1-l} \prod_{j=1}^{B+1}\left(\frac{d_j^{c_j}\left(\sigma_j^2\right)^{-\left[\frac{n\left(T_j - T_{j-1}\right)}{2}+c_j+1\right]}}{\left(A_j\right)^{\frac{n}{2}}\Gamma c_j}\right) \exp\left[-\sum_{j=1}^{B+1}\frac{C_{ij}}{\sigma_j^2}\right] d\sigma\, d\rho$$

$$= \int_l^1 \frac{(2\pi)^{-\frac{nT}{2}}}{1-l} \prod_{j=1}^{B+1} \frac{d_j^{c_j}\Gamma\left(\frac{n\left(T_j - T_{j-1}\right)}{2}+c_j\right)}{\Gamma c_j \left(A_j\right)^{\frac{n}{2}} C_j^{\frac{n\left(T_j - T_{j-1}\right)}{2}+c_j}} d\rho$$

where

$$C_{ij} = d_j + \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{t=T_{j-1}}^{T_j}\left(y_{it}-\rho y_{i,t-1}\right)^2 + \sum_{i=1}^{n}\gamma_{ij}^2 - \sum_{i=1}^{n}\frac{B_{ij}^2}{A_j}\right)$$

(A.2) Under unit root hypothesis $H_0 : \rho = 1, \mu_{i1} \neq \mu_{i2}, \sigma_1^2 \neq \sigma_2^2$, the joint likelihood function is given as

$$L(\sigma \mid y) = \prod_{j=1}^{B+1} L(\sigma_j \mid y) = \prod_{j=1}^{B+1} \left((2\pi)^{-\frac{n(T_j - T_{j-1})}{2}} \sigma_j^{-n(T_j - T_{j-1})} \exp\left[-\frac{1}{2\sigma_j^2} \sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \varepsilon_{it}^2 \right] \right)$$

$$= (2\pi)^{-\frac{nT}{2}} \prod_{j=1}^{B+1} \left(\sigma_j^{-n(T_j - T_{j-1})} \right) \exp\left[-\frac{1}{2} \sum_{j=1}^{B+1} \left\{ \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \Delta y_{it}^2 \right\} \right]$$

By using similar mathematical manipulations as above, we get the posterior probability

$$P(y \mid H_0)$$

$$= \int_{R_{B+1}^+} L(\sigma \mid y) p(\sigma) d\sigma$$

$$= \int_{R_{B+1}^+} (2\pi)^{-\frac{nT}{2}} \prod_{j=1}^{B+1} \left(\sigma_j^{-n(T_j - T_{j-1})} \right) \exp\left[-\frac{1}{2} \sum_{j=1}^{B+1} \left\{ \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \Delta y_{it}^2 \right\} \right] \frac{d_1^{c_1}}{\Gamma c_1} \frac{d_2^{c_2}}{\Gamma c_2} \cdots \frac{d_{B+1}^{c_{B+1}}}{\Gamma c_{B+1}}$$

$$\left(\sigma_1^2\right)^{-c_1 - 1} \left(\sigma_2^2\right)^{-c_{21} - 1} \cdots \left(\sigma_{B+1}^2\right)^{-c_{B+1} - 1} \exp\left[-\frac{d_1}{\sigma_1^2} - \frac{d_2}{\sigma_2^2} \cdots - \frac{d_{B+1}}{\sigma_{B+1}^2} \right] d\sigma$$

$$= \int_{R_{B+1}^+} (2\pi)^{-\frac{nT}{2}} \prod_{j=1}^{B+1} \left(\frac{d_j^{c_j}}{\Gamma c_j} \left(\sigma_j^2\right)^{-\left[\frac{n(T_j - T_{j-1})}{2} + c_j + 1 \right]} \right) \exp\left[-\sum_{j=1}^{B+1} \left\{ \frac{1}{\sigma_j^2} \left(\frac{1}{2} \sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \Delta y_{it}^2 + d_j \right) \right\} \right] d\sigma$$

$$= (2\pi)^{-\frac{nT}{2}} \prod_{j=1}^{B+1} \frac{d_j^{c_j} \Gamma\left(\frac{n(T_j - T_{j-1})}{2} + c_j \right)}{\Gamma c_j \left[\frac{1}{2} \sum_{i=1}^{n} \sum_{t=T_{j-1}}^{T_j} \Delta y_{it}^2 + d_j \right]^{\frac{n(T_j - T_{j-1})}{2} + c_j}}$$

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 25–44, DOI 10.21307/stattrans-2018-002

IMPROVED ROTATION PATTERNS USING TWO AUXILIARY VARIABLES IN SUCCESSIVE SAMPLING

Jaishree Prabha Karna[1], **Dilip Chandra Nath**[2]

ABSTRACT

The present paper emphasizes the role of two auxiliary variables on both the occasions to improve the precision of estimates at the current (second) occasion in two-occasion successive sampling. Information on two auxiliary variables, which are positively correlated with the study variable, has been used with the aid of exponential type structures and an efficient estimation procedure of population mean on the current (second) occasion has been suggested. The behaviour of the proposed estimator has been studied and compared with the sample mean estimator, when there is no matching from the previous occasion and natural successive sampling estimator, which is a linear combination of the means of the matched and unmatched portions of the sample at the current (second) occasion. Optimal replacement strategy is also discussed. The concluding remarks are discussed justifying utility of the proposed sampling scheme. The results have been well supported analytically as well as empirically by using real life data.

Key words: exponential type estimators, bias, mean squared error, optimum replacement strategy.
Mathematics subject classification: 62D05

1. Introduction

Repeated surveys over years or seasons or months are commonly used on many occasions for estimating same characteristics at different points of time. The information collected on previous occasion can be used to study the change or the total value over occasion for the character and also in addition to study the average value for the most recent occasion. There are several possibilities: (i) the same sample may be used on each occasion (ii) a new sample may be drawn on each occasion or (iii) a part of the sample may be retained while the remainder of the sample may be drawn afresh. Intuition suggests that for estimating changes from one occasion to the next, it may be best to retain the same sample on each occasion, while for estimating the mean on each occasion it may be advised to draw a fresh sample on each occasion. If it is desired to estimate the population mean on each occasion and also the change from one occasion to the next, it is

[1] Department of Statistics, Gauhati University, Guwahati – 781014, India.
E-mail: jaishree.prabha@gmail.com
[2] Department of Statistics, Gauhati University, Guwahati – 781014, India.

always better to retain a part of the sample and draw the remainder of the sample afresh.

In successive (rotation) sampling, it is more advantageous to utilize the entire information collected in the previous investigations (occasions), to cite one may refer the papers by Jessen (1942), Patterson (1950), Rao and Graham (1964), Gupta (1979), Das (1982) and Chaturvedi and Tripathi (1983). Sen (1971) has also used this technique successfully with the utilization of information on two auxiliary variables, which was readily available on previous occasion, and proposed the estimators of population mean on the current occasion in two-occasion successive sampling. Sen (1972, 1973) further extended his work for several auxiliary variables. In many situations, information on an auxiliary variable may be readily available on the first as well as on the second occasion. For instance, to study the case of public health and welfare of a state or country, several factors are available that can be treated as auxiliary variables, such as the number of beds in different hospitals may be known, number of doctors and supporting staffs may be available, the amount of funds available for medicine etc. may be known. Likewise, there is a wealth of information available, which if used efficiently can improve the precision of estimates. Utilizing the auxiliary information on both the occasions Feng and Zou (1997), Biradar and Singh (2001), Singh (2005) and Singh and Karna (2009 a, b) proposed ratio and regression type estimators for estimating the population mean on the current (second) occasion in two-occasion successive sampling. More recently, the contributions of Ralte and Das (2015), Singh and Pal (2015), Karna and Nath (2016) and Beevi and Chandran (2017) established beneficial results by using the auxiliary variables on both the occasions.

Exponential type estimators support increasing the precision of the estimates of population parameters such as mean, median, total, etc. The exponential ratio and product type estimators in the estimation of finite population mean was introduced by Bahl and Tuteja (1991), when variable of interest and auxiliary variable is negatively or positively correlated. Further, the works of Upadhyaya et al. (2011), Yadav and Cadilar (2013), Singh and Pal (2017) examine the advantageous property of the exponential type estimators.

In line with the preceding works, we propose more an effective and relevant estimator using exponential type estimators for population mean at the current occasion in two-occasion successive sampling. Properties of the proposed estimator and optimum replacement policy have been discussed. Empirical support have been given to validate the theoretical results.

2. Formulation of the estimator Δ

Let $U = (U_1, U_2, \ldots, U_N)$ be the finite population of N units, which has been sampled over two occasions. The character under study is denoted by x (y) on the first (second) occasion respectively. Assume that the information on two auxiliary variables z_1 and z_2, whose population means $\bar{Z}_1$ and $\bar{Z}_2$ are known, is available on the current (second) occasion and is closely related (positively correlated) to y on the second occasion. Let a simple random sample (without replacement) of size n be selected on the first occasion. A random sub-sample of

m = nλ units is retained (matched) for its use on the second occasion, while a fresh simple random sample (without replacement) of u = (n-m) = nμ units is drawn on the second occasion from the entire population so that the sample size on the second occasion is also n. λ and μ (λ + μ = 1) are the fractions of matched and fresh samples at the second (current) occasion. Further, we consider the following notations throughout this work:

$\overline{X}, \overline{Y}, \overline{Z}_1, \overline{Z}_2$: population means of the variables x, y, z_1 and z_2 respectively;

$\overline{x}_n, \overline{x}_m, \overline{y}_u, \overline{y}_m, \overline{z}_{1u}, \overline{z}_{2u}, \overline{z}_{1n}, \overline{z}_{2n}$: sample means of the respective variables based on the sample sizes shown in suffices;

b_{yx}: sample regression coefficient of the variable y on x;

β_{yx} : population regression coefficient of the variable y on x;

$\rho_{yx}, \rho_{yz1}, \rho_{yz2}, \rho_{xz1}, \rho_{xz2}, \rho_{z1z2}$: correlation coefficients between the variables shown in suffices;

$S_X^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i-\overline{X})^2$: population mean square of the variable x;

$S_X^2, S_Y^2, S_{Z1}^2, S_{Z2}^2$: population mean squares of the variables x, y, z_1 and z_2 respectively.

Utilizing information on two auxiliary variables, two different estimators of the population mean $\overline{Y}$ on the current (second) occasion may be considered. Motivated with the work of Bahl and Tuteja (1991), the estimator based on the fresh sample of size u drawn on the second occasion is defined by

$$\Delta_u = \overline{y}_u \exp\left(\frac{\overline{Z}_1-\overline{z}_{1u}}{\overline{Z}_1+\overline{z}_{1u}}\right) \exp\left(\frac{\overline{Z}_2-\overline{z}_{2u}}{\overline{Z}_2+\overline{z}_{2u}}\right). \quad (1)$$

The second estimator Δ_m is also a modified chain ratio type estimator based on the sample of size m, common with both the occasions, and is defined as

$$\Delta_m = \left[\overline{y}_m + b_{yx}(m)\left(\overline{x}_n-\overline{x}_m\right)\right] \exp\left(\frac{\overline{Z}_1-\overline{z}_{1n}}{\overline{Z}_1+\overline{z}_{1n}}\right) \exp\left(\frac{\overline{Z}_2-\overline{z}_{2n}}{\overline{Z}_2+\overline{z}_{2n}}\right). \quad (2)$$

Now, considering the convex linear combination of Δ_u and Δ_m , we define the final estimator of population mean $\overline{Y}$ on the current occasion as

$$\Delta = \varphi\Delta_u + (1-\varphi)\Delta_m \quad (3)$$

where φ is an unknown constant to be determined so as to minimize mean square error of the estimator Δ.

For estimating the mean on each occasion the estimator Δ_u is suitable, which implies that more belief on Δ_u could be shown by choosing φ as 1 (or close to 1), while for estimating the change from one occasion to the next, the estimator Δ_m could be more useful so φ might be chosen as 0 (or close to 0). For asserting both the problems simultaneously, a suitable (optimum) choice of φ is required.

3. Properties of the proposed estimator Δ

3.1 Bias and mean square error of estimator Δ

Since, Δ_u and Δ_m are ratio and chain-type regression in ratio estimators respectively, they are biased for population mean $\bar{Y}$ .Therefore, the resulting estimators Δ defined in (3) is also biased estimator of $\bar{Y}$. To obtain bias B (.) and mean square errors M (.) of Δ up to the first order of approximations we consider the following transformations:

$\bar{y}_u = (1+e_1)\bar{Y}$, $\bar{y}_m = (1+e_2)\bar{Y}$, $\bar{x}_m = (1+e_3)\bar{X}$, $\bar{x}_n = (1+e_4)\bar{X}$,
$\bar{z}_{1u} = (1+e_5)\bar{Z}_1$, $\bar{z}_{1n} = (1+e_6)\bar{Z}_1$, $\bar{z}_{2u} = (1+e_7)\bar{Z}_2$, $\bar{z}_{2n} = (1+e_8)\bar{Z}_2$,
$s_{yx}(m) = (1+e_9)S_{yx}$, $s_x^2(m) = (1+e_{10})S_x^2$; such that $E(e_k) = 0$ and
$|e_k| < 1 \ \forall \ k = 1, 2, 3, ..., 10.$

Relative variances and covariances are derived as

$$E(e_1^2) = \left(\frac{1}{u} - \frac{1}{N}\right)C_y^2, \quad E(e_2^2) = \left(\frac{1}{m} - \frac{1}{N}\right)C_y^2,$$

$$E(e_3^2) = \left(\frac{1}{m} - \frac{1}{N}\right)C_x^2, \quad E(e_4^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_x^2,$$

$$E(e_5^2) = \left(\frac{1}{u} - \frac{1}{N}\right)C_{z1}^2, \quad E(e_6^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_{z1}^2,$$

$$E(e_7^2) = \left(\frac{1}{u} - \frac{1}{N}\right)C_{z2}^2, \quad E(e_8^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_{z2}^2,$$

$$E(e_3e_4) = \left(\frac{1}{n} - \frac{1}{N}\right)C_x^2, \quad E(e_1e_5) = \left(\frac{1}{u} - \frac{1}{N}\right)\rho_{yz1}C_yC_{z1},$$

$$E(e_1e_7) = \left(\frac{1}{u} - \frac{1}{N}\right)\rho_{yz2}C_yC_{z2}, \quad E(e_5e_7) = \left(\frac{1}{u} - \frac{1}{N}\right)\rho_{z1z2}C_{z1}C_{z2},$$

$$E(e_2e_6) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{yz1}C_yC_{z1}, \quad E(e_2e_8) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{yz2}C_yC_{z2},$$

$$E(e_6e_8) = \left(\frac{1}{n}-\frac{1}{N}\right)\rho_{z1z2}C_{z1}C_{z2}, \quad E(e_2e_3) = \left(\frac{1}{m}-\frac{1}{N}\right)\rho_{yx}C_yC_x,$$

$$E(e_2e_4) = \left(\frac{1}{n}-\frac{1}{N}\right)\rho_{yx}C_yC_x, \quad E(e_3e_6) = \left(\frac{1}{n}-\frac{1}{N}\right)\rho_{xz1}C_xC_{z1},$$

$$E(e_4e_6) = \left(\frac{1}{n}-\frac{1}{N}\right)\rho_{xz1}C_xC_{z1}, \quad E(e_3e_8) = \left(\frac{1}{n}-\frac{1}{N}\right)\rho_{xz2}C_xC_{z2},$$

$$E(e_4e_8) = \left(\frac{1}{n}-\frac{1}{N}\right)\rho_{xz2}C_xC_{z2}, \quad E(e_3e_9) = \left(\frac{1}{m}-\frac{1}{N}\right)\frac{\alpha_{2100}}{\bar{X}S_{yx}},$$

$$E(e_4e_9) = \left(\frac{1}{n}-\frac{1}{N}\right)\frac{\alpha_{2100}}{\bar{X}S_{yx}}, \quad E(e_3e_{10}) = \left(\frac{1}{m}-\frac{1}{N}\right)\frac{\alpha_{3000}}{\bar{X}S_x^2},$$

$$E(e_4e_{10}) = \left(\frac{1}{n}-\frac{1}{N}\right)\frac{\alpha_{3000}}{\bar{X}S_x^2}.$$

where $\alpha_{qrst} = E\left[\left(x-\bar{X}\right)^q\left(y-\bar{Y}\right)^r\left(z_1-\bar{Z}_1\right)^s\left(z_2-\bar{Z}_2\right)^t\right]$; ((q, r, s, t) ≥ 0 are integers)

Under the above transformations Δ_u and Δ_m take the following forms:

$$\Delta_u = (1+e_1)\bar{Y}\exp\left(\frac{-e_5}{2+e_5}\right)\exp\left(\frac{-e_7}{2+e_7}\right) \tag{4}$$

$$\Delta_m = \left[(1+e_2)\bar{Y}+(1+e_9)(e_4-e_3)(1+e_{10})^{-1}\beta_{yx}\bar{X}\right]\exp\left(\frac{-e_6}{2+e_6}\right)\exp\left(\frac{-e_8}{2+e_8}\right). \tag{5}$$

Subsequently, we have the following theorems:

Theorem 3.1: Bias of the estimator Δ to the first order of approximations is obtained as

$$B(T) = \varphi\, B(\Delta_u) + (1-\varphi)\, B(\Delta_m) \tag{6}$$

where

$$B(\Delta_u) = \frac{1}{\bar{Y}}\left(\frac{1}{u}-\frac{1}{N}\right)\left(\frac{3}{4}-\frac{\rho_{yz2}}{2}-\frac{\rho_{yz1}}{2}+\frac{\rho_{z1z2}}{4}\right)S_y^2 \tag{7}$$

$$B(\Delta_m) = \frac{1}{\bar{Y}}\left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{3}{4}-\frac{\rho_{yz2}}{2}-\frac{\rho_{yz1}}{2}+\frac{\rho_{z1z2}}{4}\right)+\left(\frac{1}{m}-\frac{1}{n}\right)\left(\frac{\alpha_{2100}}{S_x^2}+\frac{\alpha_{3000}S_{yx}}{S_x^4}\right)$$
$$-\frac{1}{2}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{\rho_{xz2}S_{yx}}{\bar{X}} \qquad (8)$$

Proof: The bias of the estimator Δ is given by

$$B(\Delta) = E\left[\Delta-\bar{Y}\right] = \varphi E(\Delta_u-\bar{Y}) + (1-\varphi)\,E(\Delta_m-\bar{Y})$$

$$= \varphi B(\Delta_u)+(1-\varphi)B(\Delta_m) \qquad (9)$$

where $B(\Delta_u) = E\left[\Delta_u-\bar{Y}\right]$ and $B(\Delta_m) = E\left[\Delta_m-\bar{Y}\right]$.

The bias of Δ_u and Δ_m is derived as follows:

$$B(\Delta_u) = E\left[\Delta_u-\bar{Y}\right] .$$

Substituting the expression of Δ_u from from equation (4), we get

$$B(\Delta_u) = E\left[(1+e_1)\bar{Y}\exp\left(\frac{-e_5}{2+e_5}\right)\exp\left(\frac{-e_7}{2+e_7}\right)-\bar{Y}\right] .$$

Now, expanding the right hand side of the above expression, taking expectations and retaining the terms up to the first order of approximations, we have

$$B(\Delta_u) = \frac{1}{\bar{Y}}\left(\frac{1}{u}-\frac{1}{N}\right)\left(\frac{3}{4}-\frac{\rho_{yz2}}{2}-\frac{\rho_{yz1}}{2}+\frac{\rho_{z1z2}}{4}\right). \qquad (10)$$

Similarly

$$B(\Delta_m) = E\left[\Delta_m-\bar{Y}\right]$$
$$= E\left[\left\{(1+e_2)\bar{Y}+(1+e_9)(e_4-e_3)(1+e_{10})^{-1}\beta_{yx}\bar{X}\right\}\exp\left(\frac{-e_6}{2+e_6}\right)\exp\left(\frac{-e_8}{2+e_8}\right)-\bar{Y}\right]$$
$$= \frac{1}{\bar{Y}}\left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{3}{4}-\frac{\rho_{yz2}}{2}-\frac{\rho_{yz1}}{2}+\frac{\rho_{z1z2}}{4}\right)+\left(\frac{1}{m}-\frac{1}{n}\right)\left(\frac{\alpha_{2100}}{S_x^2}+\frac{\alpha_{3000}S_{yx}}{S_x^4}\right)$$
$$-\frac{1}{2}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{\rho_{xz2}S_{yx}}{\bar{X}}. \qquad (11)$$

Substituting the values of $B(\Delta_u)$ and $B(\Delta_m)$ from equations (10) and (11) in the equation (9) we get the bias of estimator Δ as shown in equation (6).

Theorem 3.2: Mean square error of the estimator Δ to the first order of approximations is obtained as

$$M(\Delta) = \varphi^2 M(\Delta_u) + (1-\varphi)^2 M(\Delta_m) + 2\varphi(1-\varphi)\mathrm{Cov}(\Delta_u, \Delta_m) \tag{12}$$

where

$$M(\Delta_u) = E(\Delta_u - \bar{Y})^2 = \left(\frac{1}{u} - \frac{1}{N}\right) A_1 S_y^2 \tag{13}$$

$$M(\Delta_m) = E(\Delta_m - \bar{Y})^2 = \left[\left(\frac{1}{m}\right) A_2 + \left(\frac{1}{n}\right) A_3 - \left(\frac{1}{N}\right) A_4\right] S_y^2 \tag{14}$$

and $$\mathrm{Cov}(\Delta_u, \Delta_m) = E\left[(\Delta_u - \bar{Y})(\Delta_m - \bar{Y})\right] = -\frac{A_1 S_y^2}{N} \tag{15}$$

where $A_1 = \frac{3}{2} + \frac{1}{2}\left\{\rho_{z1z2} - 2(\rho_{yz2} + \rho_{yz1})\right\}$, $A_2 = 1 - \rho_{yx}^2$,

$A_3 = \frac{1}{2}\left\{\rho_{z1z2} - 2(\rho_{yz2} + \rho_{yz1})\right\} + \rho_{yx}^2$ and $A_4 = \frac{1}{2}\left\{\rho_{z1z2} - 2(\rho_{yz2} + \rho_{yz1})\right\} + 1$.

Proof: It is obvious that mean square error of the estimator Δ is given by

$$M(\Delta) = E\left[\Delta - \bar{Y}\right]^2 = E\left[\varphi(\Delta_u - \bar{Y}) + (1-\varphi)(\Delta_m - \bar{Y})\right]^2$$

$$= \varphi^2 M(\Delta_u) + (1-\varphi)^2 M(\Delta_m) + 2\varphi(1-\varphi)\mathrm{Cov}(\Delta_u, \Delta_m) \tag{16}$$

where $M(\Delta_u) = E\left[\Delta_u - \bar{Y}\right]^2$, $M(\Delta_m) = E\left[\Delta_m - \bar{Y}\right]^2$ and

$\mathrm{Cov}(\Delta_u, \Delta_m) = E\left[(\Delta_u - \bar{Y})(\Delta_m - \bar{Y})\right]$.

The mean square errors of Δ_u and Δ_m are derived as follows:

$$M(\Delta_u) = E\left[\Delta_u - \bar{Y}\right]^2 .$$

Substituting the expression of Δ_u from equation (4), we get

$$M(\Delta_u) = E\left[(1+e_1)\bar{Y}\exp\left(\frac{-e_5}{2+e_5}\right)\exp\left(\frac{-e_7}{2+e_7}\right) - \bar{Y}\right]^2 .$$

Now, expanding the right-hand side of above expression, taking expectations and retaining the terms up to the first order of approximations, we have

$$M(\Delta_u) = E\left(\Delta_u - \bar{Y}\right)^2 = \left(\frac{1}{u} - \frac{1}{N}\right)\left[\frac{3}{2} + \frac{1}{2}\left\{\rho_{z1z2} - 2\left(\rho_{yz2} + \rho_{yz1}\right)\right\}\right] S_y^2 . \qquad (17)$$

Similarly

$$\begin{aligned} M\left(\Delta_m\right) &= E\left[\Delta_m - \bar{Y}\right]^2 \\ &= E\left[\left\{(1+e_2)\bar{Y} + (1+e_9)(e_4 - e_3)(1+e_{10})^{-1}\beta_{yx}\bar{X}\right\}\exp\left(\frac{-e_6}{2+e_6}\right)\exp\left(\frac{-e_8}{2+e_8}\right) - \bar{Y}\right]^2 \\ &= \left[\left(\frac{1}{m}\right)\left(1-\rho_{yx}^2\right) + \left(\frac{1}{n}\right)\left\{\frac{1}{2}\left\{\rho_{z1z2} - 2\left(\rho_{yz2} + \rho_{yz1}\right)\right\} + \rho_{yx}^2\right\} - \left(\frac{1}{N}\right)\left\{\frac{1}{2}\left\{\rho_{z1z2} - 2\left(\rho_{yz2} + \rho_{yz1}\right)\right\} + 1\right\}\right] S_y^2 \end{aligned} \qquad (18)$$

and

$$Cov\left(\Delta_u, \Delta_m\right) = E\left[\left(\Delta_u - \bar{Y}\right)\left(\Delta_m - \bar{Y}\right)\right] = -\left[\frac{3}{2} + \frac{1}{2}\left\{\rho_{z1z2} - 2\left(\rho_{yz2} + \rho_{yz1}\right)\right\}\right]\frac{S_y^2}{N} . \qquad (19)$$

Substituting the values of $M(\Delta_u)$, $M(\Delta_m)$ and $Cov(\Delta_u, \Delta_m)$ from equations (17), (18) and (19) in the equation (16) we get mean square error of estimator Δ as shown in equation (12).

Remark 1: Since x and y are the same study variable over two occasions and z_1 and z_2 are auxiliary variables positively correlated with x and y, therefore, considering the stable behaviour of coefficient of variations [Reddy (1978)] the expressions of bias mean square errors in equations (6) and (12) are derived under the assumption that the coefficients of variation of x, y and z_1 and z_2 are approximately equal, i.e. $C_x = C_y = C_{z1} = C_{z2}$.

3.2. Minimum mean square error of Δ

The mean square error of the estimator Δ in equation (12) is the function of unknown constant φ, therefore, it is minimized with respect to φ and subsequently the optimum value of φ is obtained as

$$\varphi_{opt} = \frac{M(\Delta_m) - Cov\left(\Delta_u, \Delta_m\right)}{M(\Delta_u) + M(\Delta_m) - 2Cov\left(\Delta_u, \Delta_m\right)} . \qquad (20)$$

Now, substituting the value of φ_{opt} in equation (12), we get the optimum mean square error of Δ as

$$M(\Delta)_{opt} = \frac{M(\Delta_u).M(\Delta_m) - \{Cov(\Delta_u, \Delta_m)\}^2}{M(\Delta_u) + M(\Delta_m) - 2Cov(\Delta_u, \Delta_m)}. \quad (21)$$

Further, substituting the values from equations (17) - (19) in equations (20) and (21), we get the simplified value of $M(\Delta)_{opt}$ as:

$$M(\Delta)_{opt} = \left[\frac{P_1 + \mu P_2 + \mu^2 P_3}{A_1 + \mu P_4 + \mu^2 P_5}\right]\frac{S_y^2}{n} \quad (22)$$

where $P_1 = A_1(A_2 + A_3 - fA_4)$, $P_2 = A_1\{f^2(A_4 - A_1) + f(A_4 - A_3 - A_2) - A_3\}$,
$P_3 = A_1\{f^2(A_1 - A_4) + fA_3\}$, $P_4 = f(A_1 - A_4) + (A_3 + A_2 - A_1)$,
$P_5 = \{f(A_4 - A_1) - A_3\}$.

4. Optimum replacement policy

To determine the optimum value of μ (fraction of sample to be drawn afresh on the current occasion) so that population mean $\bar{Y}$ may be estimated with maximum precision and at the lowest cost, we minimize the mean square error of Δ given in equation (22) with respect to μ , which results in a quadratic equation in μ. The quadratic equation in μ and respective solutions of μ say $\hat{\mu}$ are given below:

$$Q_1\mu^2 + 2\mu Q_2 + Q_3 = 0 \quad (23)$$

$$\hat{\mu} = \frac{-Q_2 \pm \sqrt{Q_2^2 - Q_1 Q_3}}{Q_1} \quad (24)$$

where $Q_1 = P_3P_4 - P_2P_5$, $Q_2 = A_1P_3 - P_1P_5$, $Q_3 = A_1P_2 - P_1P_4$.

From equation (24), it is clear that the real values of $\hat{\mu}$ exist if the quantity under square root is greater than or equal to zero. For any combinations of ρ_{yx} and ρ_{yz} , which satisfy the condition of real solutions, two real values of $\hat{\mu}$ are possible. Hence, while choosing the value of $\hat{\mu}$, it should be remembered that $0 \le \hat{\mu} \le 1$, all others value of $\hat{\mu}$ are inadmissible. If both the values are admissible the lowest one will be the best choice because it reduces the cost of the survey. Substituting the admissible value of $\hat{\mu}$ say μ^* from equation (24) into equation

(22), we have the optimum value of mean square error of the estimator Δ, which is shown below

$$M(T)_{opt^*} = \left[\frac{P_1+\mu^* P_2+\mu^{*2} P_3}{A_1+\mu^* P_4+\mu^{*2} P_5} \right] \frac{S_y^2}{n}. \quad (25)$$

5. Efficiency comparison

For comparing the efficiencies of the proposed estimator we have considered two estimators: (i) sample mean estimator $\bar{y}_n$, when there is no matching and (ii) natural successive sampling estimator $\hat{\bar{Y}}=\varphi^*\bar{y}_u+(1-\varphi^*)\bar{y}'_m$, when no auxiliary information is used at any occasion, where $\bar{y}'_m=\bar{y}_m+\beta_{yx}(\bar{x}_n-\bar{x}_m)$.

Clearly different estimators proposed in successive (rotation) sampling have their own assumptions and limitations; therefore, practically it is not feasible to compare the proposed estimators with the other estimators available in the survey literature. Hence, the efficiency comparisons have been made with the sample mean estimator and the natural successive sampling estimator. The percent relative efficiencies of the estimator Δ with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$, have been obtained for different choices of correlations. Since $\bar{y}_n$ and $\hat{\bar{Y}}$ are unbiased estimators of $\bar{Y}$, therefore, following Sukhatme et al. (1984) the variance of $\bar{y}_n$ and optimum variance of $\hat{\bar{Y}}$ are given by

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 \quad (26)$$

$$V\left(\hat{\bar{Y}}\right)_{opt^*} = \left[1+\sqrt{1-\rho_{yx}^2} \right] \frac{S_y^2}{2n} - \frac{S_y^2}{N}. \quad (27)$$

For N = 5000, n = 500 and different choices of ρ_{z1z2}, ρ_{yz1}, ρ_{yz2} and ρ_{yx} Tables 1-3 give the optimum values of μ and percent relative efficiencies E_1 and E_2 of Δ with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ respectively, where

$$E_1 = \frac{V(\bar{y}_n)}{M(\Delta)_{opt^*}} \times 100 \text{ and } E_2 = \frac{V\left(\hat{\bar{Y}}\right)_{opt^*}}{M(\Delta)_{opt^*}} \times 100.$$

Remark 2: To compare the performance of the estimators Δ with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$, we introduce assumptions $\rho_{xz1} = \rho_{yz1}$, $\rho_{xz2} = \rho_{yz2}$, which are intuitive

assumption, considered, for example, by Cochran (1977) and Feng and Zou (1997).

Table 1. Optimum values of μ and percent relative efficiencies of the estimator Δ with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ for f = 0.1.

ρ_{z1z2} ↓	ρ_{yz1} ↓	ρ_{yz2} →	0.5						
		ρ_{yx} →	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.3	0.5	μ^*	0.8015	0.8421	0.9058	*	*	*	*
		E_1	148.61	150.60	152.73	-	-	-	-
		E_2	144.80	143.61	141.36	-	-	-	-
	0.7	μ^*	0.7032	0.7319	0.7746	0.8385	0.9393	*	*
		E_1	200.68	204.61	209.74	215.80	221.31	-	-
		E_2	195.54	195.12	194.13	191.82	186.16	-	-
	0.9	μ^*	0.6099	0.6329	0.6663	0.7143	0.7846	0.8929	*
		E_1	319.16	327.21	338.28	352.94	371.49	392.07	-
		E_2	310.99	312.03	313.10	313.72	312.50	304.95	-
0.5	0.5	μ^*	0.8625	0.9143	*	*	*	*	*
		E_1	131.45	132.63	-	-	-	-	-
		E_2	128.09	126.48	-	-	-	-	-
	0.7	μ^*	0.7500	0.7834	0.8342	0.9135	*	*	*
		E_1	170.64	173.50	177.02	180.54	-	-	-
		E_2	166.27	165.45	163.84	160.48	-	-	-
	0.9	μ^*	0.6578	0.6832	0.7204	0.7747	0.8565	0.9891	*
		E_1	245.03	250.50	257.86	267.20	277.92	285.66	-
		E_2	238.76	238.88	238.67	237.51	233.78	222.18	-
0.7	0.5	μ^*	0.9402	*	*	*	*	*	*
		E_1	117.38	-	-	-	-	-	-
		E_2	114.38	-	-	-	-	-	-
	0.7	μ^*	0.8015	0.8421	0.9058	*	*	*	*
		E_1	148.61	150.60	152.73	-	-	-	-
		E_2	144.80	143.61	141.36	-	-	-	-
	0.9	μ^*	0.7032	0.7319	0.7746	0.8385	0.9393	*	*
		E_1	200.68	204.61	209.74	215.80	221.31	-	-
		E_2	195.54	195.12	194.13	191.82	186.16	-	-
0.9	0.5	μ^*	*	*	*	*	*	*	*
		E_1	-	-	-	-	-	-	-
		E_2	-	-	-	-	-	-	-
	0.7	μ^*	0.8625	0.9143	*	*	*	*	*
		E_1	131.45	132.63	-	-	-	-	-
		E_2	128.09	126.48	-	-	-	-	-
	0.9	μ^*	0.7500	0.7834	0.8342	0.9135	*	*	*
		E_1	170.64	173.50	177.02	180.54	-	-	-
		E_2	166.27	165.45	163.84	160.48	-	-	-

Table 2. Optimum values of μ and percent relative efficiencies of the estimator Δ with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ for f = 0.1.

ρ_{z1z2} ↓	ρ_{yz1} ↓	ρ_{yz2} →	0.7						
		ρ_{yx} →	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.3	0.5	μ^*	0.7032	0.7319	0.7746	0.8385	0.9393	*	*
		E_1	200.68	204.61	209.74	215.80	221.31	-	-
		E_2	195.54	195.12	194.13	191.82	186.16	-	-
	0.7	μ^*	0.6099	0.6329	0.6663	0.7143	0.7846	0.8929	*
		E_1	319.16	327.21	338.28	352.94	371.49	392.07	-
		E_2	310.99	312.03	313.10	313.72	312.50	304.95	-
	0.9	μ^*	0.4682	0.4879	0.5162	0.5566	0.6150	0.7028	0.8432
		E_1	1134.4	1175.1	1233.1	1314.4	1429.2	1594.2	1831.7
		E_2	1105.3	1120.6	1141.3	1168.4	1202.2	1239.9	1257.7
0.5	0.5	μ^*	0.7500	0.7834	0.8342	0.9135	*	*	*
		E_1	170.64	173.50	177.02	180.54	-	-	-
		E_2	166.27	165.45	163.84	160.48	-	-	-
	0.7	μ^*	0.6578	0.6832	0.7204	0.7747	0.8565	0.9891	*
		E_1	245.03	250.50	257.86	267.20	277.92	285.66	-
		E_2	238.76	238.88	238.67	237.51	233.78	222.18	-
	0.9	μ^*	0.5533	0.5745	0.6052	0.6488	0.7118	0.8066	0.9568
		E_1	475.73	489.52	508.81	535.09	570.38	616.27	663.09
		E_2	463.56	466.81	470.94	475.64	479.80	479.32	455.28
0.7	0.5	μ^*	0.8015	0.8421	0.9058	*	*	*	*
		E_1	148.61	150.60	152.73	-	-	-	-
		E_2	144.80	143.61	141.36	-	-	-	-
	0.7	μ^*	0.7032	0.7319	0.7746	0.8385	0.9393	*	*
		E_1	200.68	204.61	209.74	215.80	221.31	-	-
		E_2	195.54	195.12	194.13	191.82	186.16	-	-
	0.9	μ^*	0.6099	0.6329	0.6663	0.7143	0.7846	0.8929	*
		E_1	319.16	327.21	338.28	352.94	371.49	392.07	-
		E_2	310.99	312.03	313.10	313.72	312.50	304.95	-
0.9	0.5	μ^*	0.8625	0.9143	*	*	*	*	*
		E_1	131.45	132.63	-	-	-	-	-
		E_2	128.09	126.48	-	-	-	-	-
	0.7	μ^*	0.7500	0.7834	0.8342	0.9135	*	*	*
		E_1	170.64	173.50	177.02	180.54	-	-	-
		E_2	166.27	165.45	163.84	160.48	-	-	-
	0.9	μ^*	0.6578	0.6832	0.7204	0.7747	0.8565	0.9891	*
		E_1	245.03	250.50	257.86	267.20	277.92	285.66	-
		E_2	238.76	238.88	238.67	237.51	233.78	222.18	-

Table 3. Optimum values of μ and percent relative efficiencies of the estimator Δ with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ for f = 0.1.

ρ_{z1z2} ↓	ρ_{yz1} ↓	ρ_{yz2} →	0.9						
		ρ_{yx} →	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.3	0.5	μ^*	0.6099	0.6329	0.6663	0.7143	0.7846	0.8929	*
		E_1	319.16	327.21	338.28	352.94	371.49	392.07	-
		E_2	310.99	312.03	313.10	313.72	312.50	304.95	-
	0.7	μ^*	**0.4682**	0.4879	0.5162	0.5566	0.6150	0.7028	0.8432
		E_1	1134.4	1175.1	1233.1	1314.4	1429.2	1594.2	**1831.7**
		E_2	1105.3	1120.6	1141.3	1168.4	1202.2	1239.9	**1257.7**
	0.9	μ^*	*	*	*	*	*	*	*
		E_1	-	-	-	-	-	-	-
		E_2	-	-	-	-	-	-	-
0.5	0.5	μ^*	0.6578	0.6832	0.7204	0.7747	0.8565	0.9891	*
		E_1	245.03	250.50	257.86	267.20	277.92	285.66	-
		E_2	238.76	238.88	238.67	237.51	233.78	222.18	-
	0.7	μ^*	0.5533	0.5745	0.6052	0.6488	0.7118	0.8066	0.9568
		E_1	475.73	489.52	508.81	535.09	570.38	616.27	663.09
		E_2	463.56	466.81	470.94	475.64	479.80	479.32	455.28
	0.9	μ^*	*	*	*	*	*	*	*
		E_1	-	-	-	-	-	-	-
		E_2	-	-	-	-	-	-	-
0.7	0.5	μ^*	0.7032	0.7319	0.7746	0.8385	0.9393	*	*
		E_1	200.68	204.61	209.74	215.80	221.31	-	-
		E_2	195.54	195.12	194.13	191.82	186.16	-	-
	0.7	μ^*	0.6099	0.6329	0.6663	0.7143	0.7846	0.8929	*
		E_1	319.16	327.21	338.28	352.94	371.49	392.07	-
		E_2	310.99	312.03	313.10	313.72	312.50	304.95	-
	0.9	μ^*	0.4682	0.4879	0.5162	0.5566	0.6150	0.7028	0.8432
		E_1	1134.4	1175.1	1233.1	1314.4	1429.2	1594.2	1831.7
		E_2	1105.3	1120.6	1141.3	1168.4	1202.2	1239.9	1257.7
0.9	0.5	μ^*	0.7500	0.7834	0.8342	0.9135	*	*	*
		E_1	170.64	173.50	177.02	180.54	-	-	-
		E_2	166.27	165.45	163.84	160.48	-	-	-
	0.7	μ^*	0.6578	0.6832	0.7204	0.7747	0.8565	0.9891	*
		E_1	245.03	250.50	257.86	267.20	277.92	285.66	-
		E_2	238.76	238.88	238.67	237.51	233.78	222.18	-
	0.9	μ^*	0.5533	0.5745	0.6052	0.6488	0.7118	0.8066	0.9568
		E_1	475.73	489.52	508.81	535.09	570.38	616.27	663.09
		E_2	463.56	466.81	470.94	475.64	479.80	479.32	455.28

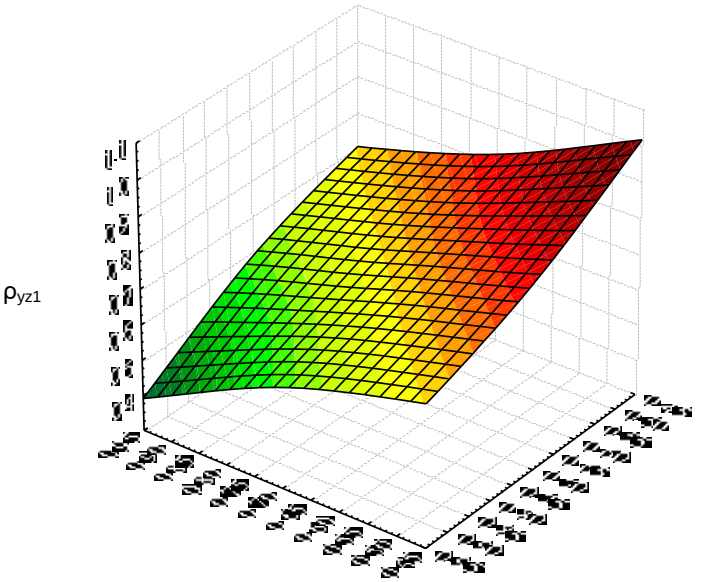


Figure 1. 3D Surface Plot of μ against ρ_{yz1} and ρ_{yz2} for ρ_{yx} = 0.3 and ρ_{z1z2} = 0.7

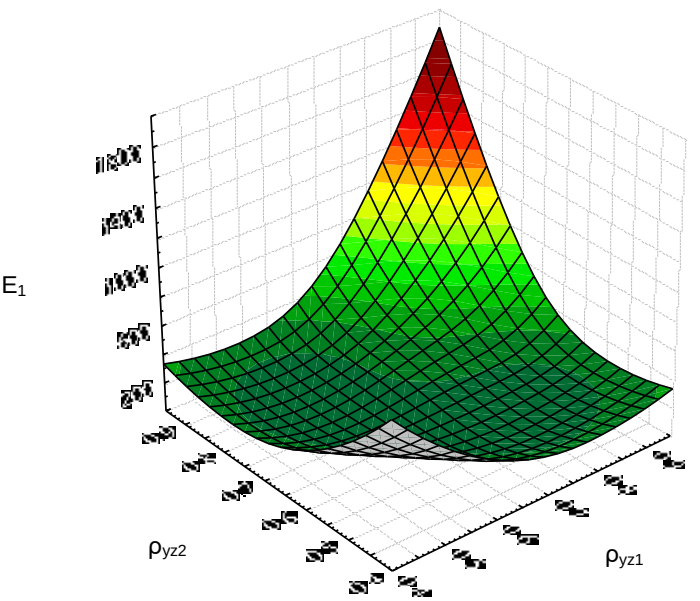


Figure 2. 3D Surface Plot of E_1 against ρ_{yz1} and ρ_{yz2} for ρ_{yx} = 0.5 and ρ_{z1z2} = 0.7

Remark 3:"*" in the Tables 1-3 indicates that μ^* do not exist for the corresponding combinations of correlations.

The following results may be extracted from Tables 1-3:

(a) For the fixed value of ρ_{yx} and ρ_{z1z2}, the value of μ decreases but E_1 and E_2 increases with the increasing values of ρ_{yz1} and ρ_{yz2}. This phenomenon is expected as availability of highly correlated auxiliary variables results in reducing the cost of survey and enhancing the precision of estimates.

(b) For the fixed value of ρ_{z1z2}, ρ_{yz1} and ρ_{yz2}, the value μ, E_1 and E_2 increases with the increasing values of ρ_{yx}. This behaviour indicates that when the study characters over the occasions are highly correlated, a larger fresh sample at the current occasion is required but the efficiency of estimates increases.

(c) For fixed values of ρ_{yz1}, ρ_{yz2} and ρ_{yx}, the values of μ are increasing whereas decreasing trends are observed in the efficiencies E_1 and E_2.

(d) The lowest value of μ is observed as 0.4682, which indicates that the fraction of fresh sample to be drawn at the current occasion is as low as 46% of the total sample size.

(e) Percent relative gain in efficiencies E_1 and E_2 seems to be appreciably high under optimal conditions. Highest values of E_1 and E_2 are observed as 1831.7 and 1257.7 respectively.

(f) A close perusal of Figure 1 suggests that for the fixed value of ρ_{yx} and ρ_{z1z2}, the value of μ decreases with the increasing values of ρ_{yz1} and ρ_{yz2}. This behaviour supports the fact that if highly correlated auxiliary variables are available, the less fresh sample is required to be drawn at the current occasion, which subsequently reduces the cost of survey.

(h) It can be clearly seen from Figure 2 that for the fixed value of ρ_{yx} and ρ_{z1z2}, E_1 increases with the increasing values of ρ_{yz1} and ρ_{yz2}. This phenomenon justifies the use of two auxiliary variables, which results in increasing the precision of estimates.

6. Illustrations using real life data

To illustrate the comparison of relative efficiency of the proposed estimator with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$, using real life approach, the data from the Census of India (2001) and (2011) were taken into account. We define the variables as

x (y): the total number of workers in villages in the district of Ranchi, India in 2001 (2011)

z_1: the total number of literate people in villages in the district of Ranchi, India.

z_2: the total number of females in the district of Ranchi, India.

Using successive sampling strategy defined in Section 2 for the above data set we have taken n (sample drawn at first occasion) = 70, m (matched portion of the sample over two occasions) = 35 and u (fresh sample drawn at second

occasion) = 35. The values of the different estimators computed from the sample along with their corresponding mean square errors and efficiency of the proposed estimator Δ with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ have been shown in the Table 4.

Table 4. Relative Efficiency (%) of estimator Δ with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ using real life data

Estimators	Estimates	MSE	% Efficiency
Δ	481	341.86	100
$\bar{y}_n$	465	1926.13	563.42
$\hat{\bar{Y}}$	422	1360.48	397.96

The above table validates the conclusions observed from the Tables 1, 2 and 3 that the proposed estimator Δ is more efficient than $\bar{y}_n$ and $\hat{\bar{Y}}$ with maximum gain in efficiency occurring while comparing it with mean per unit estimator, which is the expected phenomenon.

7. Conclusion

In the context of the preceding interpretations, it may be concluded that the use of two auxiliary variables for the estimation of population mean at the current occasion in two-occasion successive sampling is highly appreciable, as demonstrated through empirical results. It is also observed and justified through this study that the use of exponential type estimator is highly rewarding in terms of precision. Hence, the proposed estimator Δ may be recommended for its practical use by survey practitioners.

Acknowledgement

Authors are thankful to the honourable referees and University Grants Commission, New Delhi (PDFWM-2014-15-GE-ASS-28217) for providing necessary infrastructure to carry out the present work.

REFERENCES

BEEVI, N. T., CHANDRAN, C., (2017). Efficient family of ratio-type estimators for mean estimation in successive sampling using auxiliary information on both occasions. Statistics In Transition – New Series, 18 (2), pp. 227–246.

BIRADAR, R. S., SINGH, H. P., (2001). Successive sampling using auxiliary information on both occasions. Cal. Statist. Assoc. Bull. 51, pp. 243–251.

CENSUS OF INDIA, (2001). www.cesusindiagov.in.

CENSUS OF INDIA, (2011). www.cesusindiagov.in.

CHATURVEDI, D. K., TRIPATHI, T. P., (1983). Estimation of population ratio on two occasions using multivariate auxiliary information. Journal of Indian Statistical Association 21, pp. 113–120.

COCHRAN, W. G., (1977). Sampling Techniques, Wiley Eastern Limited, New Delhi, III Edition.

DAS, A. K., (1982). Estimation of population ratio on two occasions, Jour Ind. Soc. Agr. Statist. 34, pp. 1–9.

FENG, S., ZOU, G., (1997). Sample rotation method with auxiliary variable. Communications in Statistics-Theory and Methods, 26 (6), pp. 1497–1509.

GUPTA, P. C., (1979). Sampling on two successive occasions. Jour. Statist. Res. 13, pp. 7–16.

JESSEN, R. J., (1942). Statistical Investigation of a Sample Survey for obtaining farm facts, Iowa Agricultural Experiment Station Research Bulletin No. 304, Ames, Iowa, USA, pp. 1–104.

KARNA, J. P., NATH, D. C., (2016). Rotation sampling scheme using transformed auxiliary variable. Journal of Statistics & Management Systems, 19 (6), pp. 739–754.

PATTERSON, H. D., (1950). Sampling on successive occasions with partial replacement of units, Journal of the Royal Statistical Society, 12, pp. 241–255.

RALTE, Z., DAS, G., (2015). Ratio-to-regression estimator in successive sampling using one auxiliary variable. Statistics in Transition – new series, 16 (2), pp. 183–202.

RAO, J. N. K., GRAHAM, J. E., (1964). Rotation designs for sampling on repeated occasions, Journal of the American Statistical Association, 59, pp. 492–509.

REDDY, V. N., (1978). A study on the use of prior knowledge on certain population parameters. Sankhya, 40, C, pp. 29–37.

SEN, A. R., (1971). Successive sampling with two auxiliary variables, Sankhya, 33, Series B, pp. 371–378.

SEN, A. R., (1972). Successive sampling with p $(p \geq 1)$ auxiliary variables, Ann. Math. Statist., 43, pp. 2031–2034.

SEN, A. R., (1973). Some theory of sampling on successive occasions, Australian Journal of Statistics, 15, pp. 105–110.

SINGH, G. N., (2005). On the use of chain-type ratio estimator in successive sampling, Statistics in Transition, 7, pp. 21–26.

SINGH, G. N., KARNA, J. P., (2009a). Estimation of Population mean on the current in two-occasion successive sampling. Metron, 67(1), 2009, pp. 87–103.

SINGH, G. N., KARNA, J. P., (2009b). Search of effective rotation patterns in presence of auxiliary information in successive sampling over two occasions. Statistics in Transition-new series, 10, pp. 209, 59–73.

SINGH, H. P., PAL, S. K., (2015). An efficient effective rotation pattern in successive sampling over two occasions. Communication in Statistics: Theory and Methods, 45 (17), pp. 5017–5027.

SINGH, H. P., PAL, S. K., (2017). A class of exponential type estimators of a general parameter. Communication in Statistics – Theory and Methods, 46 (8), pp. 3957–3984.

SUKHATME, P. V., SUKHATME, B. V., SUKHATME, S., ASOK, C., (1984). Sampling theory of surveys with applications. Iowa State University Press, Ames, Iowa (USA) and Indian Society of Agricultural Statistics, New Delhi (India),III Revised Edition.

UPADHYAYA, L. N., SINGH, H. P., CHATTERJEE, S., YADAV, R., (2011). Improved ratio and product exponential type estimators. J. Stat. Theo. Pract. 5 (2), pp. 285–302.

YADAV, S. K., KADILAR, C., (2013). Improved exponential type ratio estimator of population variance. Revis. Colum. de Estadist. 36 (1), pp. 145–152.

APPENDIX

Bias of estimator Δ_u

$$B(\Delta_u) = E\left[\Delta_u - \bar{Y}\right]$$

$$= E\left[\bar{y}_u \exp\left(\frac{\bar{Z}_1 - \bar{z}_{1u}}{\bar{Z}_1 + \bar{z}_{1u}}\right) \exp\left(\frac{\bar{Z}_2 - \bar{z}_{2u}}{\bar{Z}_2 + \bar{z}_{2u}}\right) - \bar{Y}\right]$$

$$= E\left[(1+e_1)\bar{Y}\exp\left(\frac{-e_5}{2+e_5}\right)\exp\left(\frac{-e_7}{2+e_7}\right) - \bar{Y}\right], \text{ from eq. (4)}$$

$$= \bar{Y}E\left[e_1\exp\left(\frac{-e_5}{2+e_5}\right)\exp\left(\frac{-e_7}{2+e_7}\right)\right]$$

$$= \bar{Y}E\left[e_1\exp\left\{-\frac{e_5}{2}\left(1+\frac{e_5}{2}\right)^{-1}\right\}\exp\left\{-\frac{e_7}{2}\left(1+\frac{e_7}{2}\right)^{-1}\right\}\right]$$

$$= \bar{Y}E\left[e_1\left\{1-\frac{e_5}{2}\left(1+\frac{e_5}{2}\right)^{-1}+\frac{1}{2}.\frac{e_5^2}{4}\left(1+\frac{e_5}{2}\right)^{-2}+\ldots\right\}\left\{1-\frac{e_7}{2}\left(1+\frac{e_7}{2}\right)^{-1}+\frac{1}{2}.\frac{e_7^2}{4}\left(1+\frac{e_7}{2}\right)^{-2}+\ldots\right\}\right]$$

Retaining the terms up to the first order of approximations, we get

$$B(\Delta_u) = \bar{Y}E\left[e_1\left\{1-\frac{e_5}{2}+\frac{3e_5^2}{8}\right\}\left\{1-\frac{e_7}{2}+\frac{3e_7^2}{8}\right\}\right]$$

$$= \bar{Y}E\left[e_1-\frac{e_5}{2}-\frac{e_7}{2}+\frac{3e_5^2}{8}+\frac{3e_7^2}{8}-\frac{e_1e_7}{2}-\frac{e_1e_5}{2}+\frac{e_5e_7}{4}\right]$$

$$= \bar{Y}E\left[e_1-\frac{e_5}{2}-\frac{e_7}{2}+\frac{3e_5^2}{8}+\frac{3e_7^2}{8}-\frac{e_1e_7}{2}-\frac{e_1e_5}{2}+\frac{e_5e_7}{4}\right]$$

Taking expectations as discussed in Section 3.1 and using Note 1, we have

$$B(\Delta_u) = \frac{1}{\bar{Y}}\left(\frac{1}{u}-\frac{1}{N}\right)\left(\frac{3}{4}-\frac{\rho_{yz2}}{2}-\frac{\rho_{yz1}}{2}+\frac{\rho_{z1z2}}{4}\right)S_y^2$$

Bias of estimator Δ_m

$$B(\Delta_m) = E\left[\Delta_m - \bar{Y}\right]$$

$$= E\left[\left[(1+e_2)\bar{Y} + (1+e_9)(e_4 - e_3)(1+e_{10})^{-1}\beta_{yx}\bar{X}\right]\exp\left(\frac{-e_6}{2+e_6}\right)\exp\left(\frac{-e_8}{2+e_8}\right) - \bar{Y}\right], \text{ from eq (5)}$$

Expanding the exponentials as in the case of bias of estimator Δ_u and retaining the terms up to first order of approximations, we get

$$B(\Delta_m) = E\left[\left\{(1+e_2)\bar{Y} + (e_4 - e_3 + e_4e_9 - e_3e_9 - e_4e_{10} + e_3e_{10})\beta_{yx}\bar{X}\right\}\left\{1 - \frac{e_8}{2} + \frac{3}{8}e_8^2 - \frac{e_6}{2} + \frac{e_6e_8}{4} + \frac{3}{8}e_6^2\right\} - \bar{Y}\right]$$

$$= E\left[\left(\frac{3}{8}e_8^2 + \frac{3}{8}e_6^2 - \frac{e_2e_8}{2} - \frac{e_2e_6}{2} + \frac{e_6e_8}{4}\right)\bar{Y} + \left(e_4e_9 - e_3e_9 - e_4e_{10} + e_3e_{10} - e_4e_8 - \frac{e_4e_6}{2} + \frac{e_3e_8}{2} + \frac{e_3e_6}{2}\right)\beta_{yx}\bar{X}\right]$$

Now, taking expectations as discussed in Section 3.1 and using Note 1, we have

$$B(\Delta_m) = \frac{1}{\bar{Y}}\left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{3}{4} - \frac{\rho_{yz2}}{2} - \frac{\rho_{yz1}}{2} + \frac{\rho_{z1z2}}{4}\right) + \left(\frac{1}{m} - \frac{1}{n}\right)\left(\frac{\alpha_{2100}}{S_x^2} + \frac{\alpha_{3000}S_{yx}}{S_x^4}\right)$$

$$- \frac{1}{2}\left(\frac{1}{n} - \frac{1}{N}\right)\frac{\rho_{xz2}S_{yx}}{\bar{X}}$$

- The mean square errors of the estimators Δ_u and Δ_m have been obtained in the similar manner as the bias of the estimators.
- The variance of the estimator natural successive sampling estimator $\hat{\bar{Y}}$ has been discussed in detail in Sukhatme et al. (1984), pages 256-260.

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 45–60, DOI 10.21307/stattrans-2018-003

JOINT RESPONSE PROPENSITY AND CALIBRATION METHOD

Seppo Laaksonen[1], **Auli Hämäläinen**[2]

ABSTRACT

This paper examines the chain of weights, beginning with the basic sampling weights for the respondents. These were then converted to reweights to reduce the bias due to missing quantities. If micro auxiliary variables are available for a gross sample, we suggest taking advantage first of the response propensity weights, and then of the calibrated weights with macro (aggregate) auxiliary variables. We also examined the calibration methodology that starts from the basic weights. Simulated data based on a real survey were used for comparison. The sampling design used was stratified simple random sampling, but the same methodology works for multi-stage sampling as well. Eight indicators were examined and estimated. We found differences in the performance of the reweighting methods. However, the main conclusion was that the response propensity weights are the best starting weights for calibration, since the auxiliary variables can be more completely exploited in this case. We also tested problems of calibration methods, since some weights may lead to unacceptable weights, such as below 1 or even negative.

Key words: reweighting, simulation study, macro vs. micro auxiliary variables, case of negative and other implausible weights

1. Introduction

Nonresponse and coverage problems are common in surveys. Both problems are increasing rather than declining. Unless the fieldwork is successful or special data collection modes are found and used, post-survey adjustments are the only option to try for improving the data quality. In this study, we concentrated on weighting adjustments. Reweighting is useless without appropriate auxiliary data. That is, we cannot do much without these variables. We tested two types of auxiliary variables: (i) aggregate, or macro, and (ii) micro.

Both types of auxiliary variables require that their values be available both for the respondents and for the non-respondents, and hopefully for ineligibles as well. In multi-stage designs the micro variables are often more difficult to get since the first stage is an area or an address, but macro variables still can be created. In the case of element-based sampling such as stratified simple random

[1] University of Helsinki. Finland. E-mail: Seppo.Laaksonen@Helsinki.Fi.
[2] University of Helsinki. Finland.

sampling, there are principle reasons favouring to try to get as good and many micro auxiliary variables as possible to be used. On the other hand, there are many alternatives for macro variables. For instance, if a primary sampling unit (PSU) is an area cluster, it is possible to get various types of aggregate figures at this level. For instance, Laaksonen et al. (2015) could not get education at the micro level, but they had access to real grid data that gave the opportunity to construct the proportion of highly educated people at each grid. This is not, of course, as good as an education code with several categories at the micro level, but improved their weighting to some extent. In general, some types of calibration margins can always be used as macro auxiliary variables. This feature is a good reason to take advantage of calibration methods.

Brick (2013) presented an overview to weighting adjustments in the case of unit nonresponse. His earliest citations were from 1940's. He identified three major themes for nonresponse. Statistical adjustment of the survey weights to adjust for survey nonresponse is his third theme, while retaining the design-based mode of inference. He presented, at a general level, the following weighting methodologies: response propensity weighting, response homogeneity group weighting or weighting class methodology (see also Little and Rubin, 2002), propensity stratification (Valliant et al., 2013) and calibration estimation following Deville and Särndal (1992). Brick mentions also that post-stratification as a basic calibration estimator has been used for decades (Holt and Smith, 1979; Smith, 1991).

Post-stratification was augmented by Deville and Särndal (1992), leading to a more general approach so that several margins can be used to benchmark the reweights as precisely as the recent known population figures imply. This is the initial approach to calibration and can be used if the certain population margins are available. The quality of these margins should be as good as possible to succeed well in calibration. They are the types of benchmarking figures so that these 'estimates' will be automatically like 'true' values. This quality is not guaranteed for other estimates or for proper survey estimates, but it is expected that their bias will be reduced to some extent. The reason is that the benchmark margins correct often for frame and nonresponse errors. And more generally, it follows that an appropriate calibration method could possibly be an "ending method", after possible other weighting methods based on micro auxiliary variables. We follow this strategy.

Särndal and Deville worked later together with Sautory (1993) leading to a SAS macro Calmar. After that, the prosperity of the calibration methodology was ready to begin. Later, a second version of the SAS macro, Calmar 2, was published and became publicly available, providing new options for calibration (le Guennec and Sautory, 2005), including five distance functions. The macro gives opportunity to insert the margins of two levels (e.g. households and household members), but we do not examine this feature in this paper.

The distance function is used to minimize the change between the starting weights and the new calibrated weights while the benchmark margins are satisfied. The often applied distance function is linear, but this might be problematic since some weights can be negative or below one and this is not acceptable since the weight of every respondent should be at least one. Calmar 2 fortunately has four other functions, and two of them never yield to implausible weights; that is, raking ratio and sinus hyperbolicus, respectively. Two of these

other methods include the bound option that may help in getting correct weights, but it is not clear which bounds to use. It is also possible that the algorithm does not work with inappropriate bounds. In general, the use of bounds is usually subjective, and the target is to get acceptable weights, but it is not guaranteed that they are reducing the bias in estimates. Valliant et al. (2013) say that these can be moved to the boundary, but we do not agree with this statement since it is even more subjective. We thus do not recommend subjective strategies in weighting or other survey methodologies.

Calibration methodology flourished extensively in the 2000's, and various specifications were developed (Kott, 2006; Kott & Chang, 2008, 2010; Lumley et al., 2011; Särndal, 2007). However, linear calibration was the common method. The often used form of it was the generalized linear regression estimation method GREG (Estevao and Särndal, 2006; Särndal, 2007; Henry and Valliant, 2015; Valliant et al., 2013).

Another approach to reweighting is to exploit micro level auxiliary variables as well as possible. The basic ideas of this methodology are mainly from the 1980's (Little 1986). Little and Rubin (2002) use the term "propensity weighting" that we also use, but adding the word "response", which was used by Brick (2013). The model behind the response propensity (RP) weighting is usually logistic regression, but probit regression (Laaksonen and Heiskanen, 2014) and other link functions can be applied as well. First applications of RP weighting were done at the group level, often called response homogeneity groups, adjustment cells or weighting classes (Valliant et al., 2013; Brick, 2013; Little and Rubin, 2002; Ekholm and Laaksonen, 1991). The group methods are still much used (Haziza and Lesage, 2016).

Laaksonen (2007) applied the RP weighting technique so that he first estimates a logistic regression model for predicting the response propensities to the individual respondents. In the second stage, he divides the basic sampling weights with these propensities to get the preliminary weights. Finally, he benchmarks these weights to correspond to the known population of the explicit strata. This was done since the sampling design was the stratified random sampling and these population figures were available in the beginning, before the fieldwork. The success of this methodology depends on the richness of the micro-level auxiliary variables. Macro variables can be used, and they are usually predicting the missing quantities as well. The benchmarking in this study was ensured at the stratum level, or at the post-stratum level, if applied after post-stratification (Laaksonen et al., 2015).

This study combined both approaches, that is, calibration methods and response propensity weighting, which were not mentioned in Brick (2013) or in the textbook of Valliant et al. (2013). Our approach consisted of the three steps. First, the basic sampling weights were computed using the sampling design of the survey and assuming that the response mechanism was ignorable within strata, but not between strata. Then the response propensity weights were constructed, and finally, these weights were used as the starting weights in the calibration. The strategy gave more benchmarks than the initial RP weighting does. It took advantage of both micro and macro auxiliary variables that were available. We compared the different methods with each other using a simulated data set that

was based on a modification of a real data set, but is less complex than the initial data set (Laaksonen and Heiskanen, 2014).

This artificial population was extracted from the data set of the about 3000 respondents and then copied enough times to get the universe of about 180 thousand. The missing indicators remained in the data, but some randomness was added in survey variables to get a more realistic file. This new universe gave the opportunity to see how well each method works in nearly real-life situations. We also compared variations of calibration methods that were invented in Calmar 2.

In Section 2, we present both the calibration methods of Calmar 2, our response propensity weighting method and the principles of the joint response propensity and the calibration method. Section 3 goes on to describe our simulated data. The following section summaries our empirical results, which clearly show that our main method was best on average, and never worst, even though it did not lead to essential improvement in all cases. Section 5 continues analysing the calibration weights of Calmar 2 with a partially new sample, which illustrates variation between distance functions. This analysis also exposed that both linear and logistic distance functions may bear unacceptable weights, and such weights cannot be used in practice without manipulation. The final section provides our conclusions.

The bias of mean estimates only is considered in the empirical part. We follow Brick and Jones (2008) in this sense. They said that variability can be measured reasonably well, whereas bias is difficult to measure due to nonresponse.

2. Calibration, response propensity weighting and their joint method

Successful calibration methods were developed in the early 1990's, when the Deville and Särndal article (1992) was published. The methods were further developed when the first Calmar SAS macro was coded by the French statistical office INSEE in 1993 (Deville et al., 1993; Sautory, 2003; Willenberg, 2009). The new version, Calmar 2, published in 2003, offered new resources for performing calibrations and implements the generalized calibration method of handling non-response.

The theory of calibration estimators takes advantage of a distance function between the starting weight and the new calibrated weight, $G(w_k/d_k) = G(x_k) = G(x)$. This distance should be minimized while the desired calibration margins are satisfied. These margins are vectors of the macro auxiliary variables given by the user. The calibrated weights yield these same "estimates", thus concerning aggregates of auxiliary variables. It does not ensure anything about the accuracy of survey estimates. Some improvement is expected if the margins correct for the frame errors, for example. The results are expected to be less biased if the macro auxiliary variables and the survey variables are correlated.

Calmar 2 is a SAS macro into which a user can choose the three types of options:

(i) the initial or starting weight, d_k, which will be calibrated (we have two alternatives for this weight),

(ii) the calibration margins that are expected to be as true population totals as possible,

(iii) the calibration methods with alternative distance functions.

Calmar uses Lagrange multipliers in minimizing the distance function $x=d_k/w_k$, in which w_k is the calibrated weight. The original version of Calmar offered four calibration methods and later one more was offered (Le Guennec and Sautory, 2005; Mc Cormack, 2006), corresponding to different distance functions. This number is more than in most other software packages that usually mention the two first methods (Brick and Jones, 2008; Valliant et al., 2013). On the other hand, many other distance functions are mentioned in theoretical papers (Plikusas and Pumputis, 2010). The Calmar 2 methods are characterized by the form of function as follows:

- the *linear* method (the formula is $G(x) = \frac{1}{2}(x-1)^2$): the calibrated estimator is the generalized regression estimator,
- the *exponential* method (the formula is $G(x) = x\log \mathrm{x} - \mathrm{x} + 1$): this is also called the *raking ratio method*,
- the *logistic* method (the formula is $G(x) = xlog\frac{x}{1-x}$): this method provides lower limits *L* and upper limits *U* on the weight ratios *x* (bounds),
- the *truncated linear* method, in which the distance function is linear, but the bounds are included as in logistic method, and
- the sinus hyperbolicus method, which (the formula $G(x) = \frac{1}{2\alpha}\int_1^x \sinh\left[\alpha\left(t-\frac{1}{t}\right)\right]dt, \quad \alpha > 0 \quad implemented\ in\ CALMAR\ 2)$ does not give negative or other implausible weights.

Implausible weights are not ensured in the cases of the linear and the logistic methods without any bounds. However, it should be noted that a user should make the selection of these limits, maybe subjectively. We do not recommend using the bounds for this reason.

There are other tools to calibrate weights, but Calmar 2 is, to our knowledge, one of the best ones and used extensively in Europe. It includes several methods as well. Lumley's (2013) *R* package has been used much as well. For instance, the first nonresponse weights of the European Social Survey (2014) were produced using this *R* package. These are called post-stratified weights even though they are like the raking-ratio weights of Calmar 2.

For this paper, we obtained results with all five methods, but we present the detailed simulation results only with the linear method. This is due to fairly similar results obtained with all five methods, and hence these minor differences are not interesting. The reason for the minor differences was our ordinary sampling design (Section 4). On the other hand, we continued with a specific sampling design in Section 5, which illustrates better problems obtained with linear and logistic calibration. Moreover, this design and its sample also illustrate the role of bounds. We used similar bounds in all empirical analysis so that they are symmetric, relatively. This means that the upper bound was equal to 5 and the

lower bound was its inverse, that is 0.2=1/5**.** The range of these was ordinary. Using limits, the algorithm failed to converge in one case (see Section 5).

The strategy for creating 'response propensity weights' was as follows (Laaksonen and Heiskanen, 2014):

(i) We obtained the gross sample design weights that are the inverses of the inclusion probabilities. These inclusion probabilities varied by strata but were constant within each stratum.

(ii) We assumed that the response mechanism within each stratum is ignorable (assuming that the response mechanism is random within strata but not random between strata), and hence computed the basic weights analogously to the weights (i). These are available only for the respondents k, and symbolised by $w_k = \frac{N_h}{r_h}$, in which N_h refers to the target population, r to the respondents and h to four strata.

(iii) Next, we took those basic weights and divided them by the estimated response probabilities (called also response propensities) of each respondent obtained from the logit (probit link gives quite similar results) model, and symbolised by p_k. It is good to concentrate on the model building if several auxiliary variables are available, e.g. interactions can be tried.

(iv) Before going forward, it is good to check that the probabilities p_k are realistic, that is they are not too small (let say below 0.05), for instance. All probabilities were, of course, below 1, and hence all weights were plausible.

(v) Since the sum of the weights (iii) did not match the known population statistics by strata h, they should be calibrated so that the sums are equal to the sums of the basic weights in each stratum. This was done by multiplying the weights (iii) by the ratio $q_h = \frac{\sum_h w_k}{\sum_h w_k / p_k}$. This is one option for the response propensity modelling weighting, called "Pure" in Table 2.

(vi) It is good also to check these weights against basic statistics, such as the mean, the maximum, the minimum and the coefficient of variation. This was done for the first sample and as soon as the weights gave plausible results, the next repetitions were performed in the same way.

The joint response propensity and calibration (JRPC) weighting means that we used the response propensity weights as the starting weights in calibration, whereas these are the basic weights in pure calibration. This study is focused on the joint method, but we compared all other weights in the same framework. JRPC weighting is a two-stage calibration method. It is possible to perform it in one stage as well. Kott & Chang (2008, 2010) present such a method, but it is not as straightforward as our solution to work with both methods. The methodology by Haziza & Lesage (2016) follows the similar strategy that combined response propensity weighting and linear calibration. Their simulation application is

'fictional' (not from a real case) and hence it is difficult to see how well comparable it is with our framework. On the other hand, it is not possible to see details of their response propensity model that are of high importance in practice.

We combined both methods, but each stage can be considered separately. We think that the two-stage strategy is rational. It is not even necessary to continue to calibrate if 'a client' is happy with the RP weighting. Much depends on the availability of good calibration margins. It is, however, good to remember that the calibration as an ending method is useful if the calibration margins are correct. These margins are often known well by users (e.g. distribution by gender or age group) and if they are not correct, survey estimates may not be trusted, either, even though these are not necessarily related to each other.

3. Data and simulation principles

The data for simulations were created from the 2010 Finnish Security Survey (Laaksonen and Heiskanen, 2014) so that its three independent data sets from the respondents (face-to-face, phone and web) were first pooled together. Then this data set was extended from about 3,000 respondents to the artificial target population data set with 180,000 people. The extension was rather straightforward, but minor randomness was added to income values, among others, to avoid same values. As far as the absence of the target population was concerned, we followed as well as possible the initial unit nonresponse, and hence the response rate of our data was about equal and about 49%.

We expended much effort in our initial study to gather as many auxiliary variables as possible. So, we had the same chance to use these in our simulations as well. Since the simulation sample was essentially smaller than in the initial survey, we had to apply a bit less demanding model. However, our final response propensity model consisted of the following explanatory variables (number of categories in parenthesis): interaction of gender (2) and age group (5), education level (6), stratum (4), partnership (2), children or not at home (2), unemployed or not (2), mother tongue (3), number of rooms of house (4), if living in the municipality born or not (2). These were not very significant in all samples, but a good point in response propensity-based adjustments is that insignificancy does not violate the adjusted weights, but its impact on an estimate is less remarkable in such a case. Thus, it may not improve the estimates in all simulation samples.

The absence itself was very randomized, but it had a similar feature as in the initial survey. There was thus an absence indicator for each target population unit (Brick and Jones, 2008). When drawing samples from this population, absence varied in each sample correspondingly. This added uncertainty to the simulations. The response of each sample followed the Bernoulli scheme, that is, the number and distribution of the respondents (and by strata) varied randomly to some extent.

There are the types of variables that do not exactly correspond to those of the appendix of the paper by Laaksonen and Heiskanen (2014). The violence variables were based on 8 to 10 binary questions as to whether a respondent has met that violence problem at least once. Then, the prevalence indicator was estimated. The income variable was continuous, and it was expected to be

explained better than the other seven using the auxiliary variables available that were the same as in the published paper.

Table 1 shows the averages of all indicators in the entire population. In simulations, we try to get the estimates that are as close to these values as possible. It is not easy for many reasons. One special reason is that there can be a rather small number of observations for some indicators. This is indicated in the right column. This absence is not due to nonresponse, it is mainly due to the topic itself. For example, if a person never had a partner, the answer is empty. The reason for some absence is not exactly known, such as violence by stranger recently or harassment ever. On the other hand, we do not need to know absences well, we only need to calculate the estimates and compare these to those true target population figures. When interpreting the results, it is however good to keep in mind that some estimates are computed from a small sample size.

Table 1. Major statistics for 15-79 years old population. 179,985 persons in the simulations. The response rate below 99% means that the question did not concern them

Indicator in simulations	Average in population	Response rate, %
Income (yearly)	44905€	100
Worry (about crime)	28%	100
Harassment recently	43%	99
Harassment ever	74%	24
Violence by stranger recently	33%	24
Violence by stranger ever	87%	41
Violence by partner	16%	74
Violence by ex-partner	30%	45

The simulation strategy, naturally, followed the survey principles:

(i) Four explicit strata by four large regions were formed.
(ii) A simple random sample with a disproportional allocation was drawn from each stratum, and altogether equalled 2,000 individuals. The disproportionality was moderate (the maximum 3 times as big as the minimum). This simple design meant that paying attention to side effects due to complex sample design was not needed.
(iii) Basic sample weights were computed for the respondents as usual, dividing the target population sizes by the number of the respondents (assuming ignorable unit non-response).
(iv) The calibrated weights were computed using Calmar 2 from the basic weights. The margin variables were four strata, two genders and five age groups. These variables are quite easily available in many countries. In principle, we could add more margins, but this was not realistic since it is possible in a few cases in practice only, and would require more resources. We had more margins in our specific study (Section 5).
(v) Response propensity (RP) weights were respectively calculated.
(vi) The similar calibration as in (iv) was performed taking the RP weights as the starting weights.
(vii) The mean estimates were calculated using all weights.

(viii) The procedure from (ii) to (vii) was repeated 150 times and the output data set obtained.
(ix) The results between simulated results and true values were compared.

4. Summary of the simulation results

We drew 150 samples from this simulation population using stratified simple random sampling, which is the most common design in countries with a population register frame. This number of simulations is not big. One reason is that the whole Calmar procedure was fairly demanding, and took time while the outputs were not easy to handle further. The second reason is that the estimates were found to be stable, even after 70 simulations. Two indicators, however, did not become very stable. These were “violence due to stranger ever” and ”violence due to partner”. These results should be interpreted with caution. If the difference is minor between the two methods compared, it should not be taken seriously. This point is not critical to our simulation study, but it is general in surveys with enough complex estimates. The estimates thus are not always ideal even using good weights.

The relative bias from the true value (=(estimate-true value))/true value) is the most illustrative way to compare results since it is not needed to look at the indicator values themselves (their averages are in Table 1). It is common in other studies such as Brick and Jones (2008). The comparisons are presented in Table 2.

Table 2. Results for the relative bias from the true value (%) with basic weights and RP weights, and both continued with linear calibration. The term “Pure” means without calibration. The most biased estimates are in red, the best ones are bolded. The order of the indicators is by the success of the joint RP and linear calibration (last column). The standard error of simulations is small (from 0.1% to 0.5%)

Indicator	Basic weight		Response propensity	
	Pure	Calibration	Pure	Calibration
Violence by ex-partner	**0.60**	-2.72	-0.83	-1.21
Harassment ever	-1.36	-2.22	-0.53	**-0.42**
Worry	0.76	-0.84	0.16	**-0.02**
Violence by stranger recently	-1.03	**-0.10**	0.12	0.15
Harassment recently	6.91	0.55	0.62	**0.32**
Income	2.06	1.79	0.39	**0.33**
Violence by partner	7.24	**4.22**	4.68	4.52
Violence by stranger ever	6.50	**2.39**	4.86	5.27
Average success ranking by four methods (1=best, 2 =second best, 3=third best, 4=worst)	3.38	2.50	2,25	**1.88**

The last row of Table 2 shows an overall ranking of the methods. If the ranking was interpreted straightforwardly, we see that the best method was joint response propensity and calibration, with pure response propensity being the second and pure basic weighting the worst. There are exceptions, nevertheless. Pure basic weighting was best for "violence by ex-partner", for which any method does not work well. Pure calibration worked well for another difficult indicator, "violence by partner". This indicator seemed to be most difficult to estimate well with any method. We cannot explain why the joint method was as bad with this method, but better than obtained by the basic weights.

Another hard indicator to estimate correctly was "violence by stranger ever". The auxiliary variables were not appropriate to predict absence if any method was not reasonably good. Fortunately, we see that this succeeds well with some indicators, the best being worry, then income, violence by stranger recently and then harassment recently. In these cases, basic weights did not lead to reliable estimates. These weights were created without other auxiliary variables except region that is used in sampling design. It was well understood that the bias in income can be reduced using micro auxiliary variables available in our study. It was interesting that the similar reduction was found in worry as well. Our joint method even gave slightly better results than pure RP weighting.

In general, almost all weights with adjustments improved the estimates to some extent, but they were either upward or downward biased. Secondly, it seems that even sophisticated weights did not always improve the accuracy substantially. A good point is that they did not deteriorate them, either, although the improvement was minor. It is good to keep in mind that the calibration as the ending method was good if the margins were "true population totals". Respective estimates, such as income aggregates, were obviously more reliable as well.

5. Testing possibility to get inappropriate weights

We tested five weights, although our simulation results in Table 2 concerned only linear calibration. The estimates by the different calibration weights are approximately equal. This is due to our sampling design, in which the sample allocation into strata was fairly proportional (Section 3) and the response mechanism was same as earlier. Our auxiliary variables in calibration were also of good quality. For these reasons, all of our weights were correct, at least in the sense that their values were above one.

However, it was realized that some weights may lead to implausible weights that are below one or even below zero. We did not find references with empirical examples that examined this problem, although it is well known (the problem was mentioned in Deville and Särndal, 1992). Hence, we wanted to test this awkward opportunity with our simulation data. The linear weights are most well-known problematic weights. The implausible weights may be avoided using the bounds given for the Calmar 2 macro. These bounds are the relative limits of the calibrated weights compared to the starting weight, thus a lower, and the upper bound, respectively. The bounds naturally are given subjectively. After some attempts, we decided to choose the following limits: LOW=0.2, UP=5. The range is rather ordinary, but we failed in all experiments to get any result. These limits are possible to give for the two distance functions, both for linear and logistic, but

we do not recommend using such bounds. Raking ratio and sinus hyperbolicus never give implausible weights.

Our special experiment included the modification for our basic simulations. One is the sample size that was reduced with 50 percent (from 2000 to 1000) but the same absence indicator was used as in simulations. The sample allocation was made more disproportional. Interestingly, the minimum gross sample design weights were close to each other, and decreased from 36 to 31. It is good to recognize that a small weight can more easily remove below 1 unless very strict limits in bounds are used. On the other hand, a strict limit may mean that the algorithm fails to converge.

The second difference is that we created more calibration margins. Moreover, we tested two types of auxiliary margins, those derived from the target population frame and those derived from one sample. The latter is often used in practice, since true margins are not always possible to get. It has been supposed that it does not matter much if the sample data do not fit well to the real-life population. For example, the European Social Survey (2014) used margins for post-stratified weights that were derived from the Eurostat labour force survey although its quality is not complete. This data source was the best available and, hence, it was used.

We tested the three types of auxiliary margins presented in Table 3, all based on both the real population and the sample data.

We knew in advance that it was possible to get implausible weights either with linear calibration or logistic calibration, and hence we did not take care of other calibration methods, although they were used. Table 4 gives the summary of linear calibration. The weights were calculated similarly as in simulations.

Table 3. The margins used in the specific examination, obtained from the simulation data (true values) or from one sample data

(i)	The same as in our simulations, thus gender, age group and region but using 13 age groups (instead of 5 in the simulation).
(ii)	Adding first education with five categories
(iii)	Adding then marital status with four categories.

As Table 3 shows, all margins required more than in the first simulations since the number of age groups was essentially larger. This did not damage the weights when starting from basic weights, or even when the margins were not ideal, that is, when using sample-based margins. The negative weights were not met while the three margins were correct and the weights were response propensity-based. Instead, starting from the response propensity weights and from sample margins, negative weights were received. This problem worsened when the number of calibration margin variables was increased.

Table 4. Amount of invalid weights in linear calibration. Margins, see Table 3.

Starting weights	Negative weights, per cent	
	Population margins	Sample margins
Basic weights, margins (i)	0.0	0.0
RP Adjusted weights, margins (i)	0.0	7.8

Basic weights, margins (ii)	6.2	6.6
RP Adjusted weights, margins (ii)	3.7	9.7
Basic weights, margins (iii)	7.9	13.0
RP Adjusted weights, margins (iii)	7.8	11.0

Table 5 summaries the results of incorrect weights using logistic calibration. This method only can give weights below 1, but not negative weights. These problems were less dramatic than in the case of linear weighting, but however they may occur. The special case is that the Calmar 2 algorithm could not get any weights if the margins were sample-based and the calibration was started from RP weights. We found that this is more common if the bounds are stricter than we used.

Table 5. Amount of invalid weights in logistic calibration with bounds. N.A. means that the calibration algorithm did not converge

Starting weights	Weights below 1, per cent	
	Population margins	Sample margins
Basic weights, margins (i)	0.0	0.0
RP Adjusted weights, margins (i)	0.0	0.0
Basic weights, margins (ii)	0.0	0.0
RP Adjusted weights, margins (ii)	1.4	1.6
Basic weights, margins (iii)	0.0	0.0
RP Adjusted weights, margins (iii)	3.7	N.A.

What to do if the weights are below one? Our recommendation is not to increase these weights subjectively above one, but to change the calibration strategy. There are several options to do so. The best one is to use another distance function, but it is possible also to collapse calibration margins. Naturally, it is good to use as good margins as possible, since sample-based margins might be inconsistent. This topic should be further examined.

6. Conclusion

This study compared four weighting methods so that one group of these methods was either the basic weighting or response propensity weighting. The second group was calibrated so that either the basic weights or the response propensity weights were used as the starting weights for calibration. The calibration of the first applications was following the linear distance function that works technically in the first group when the number of the calibration margin variables is three and they are true values. In this case when the number of calibration categories is moderate, and they are true values, the impact of the distance function was not big, that is, the estimates were about equal. This pattern does always give plausible weights, but that is not the case if more calibration margin categories are added.

If the calibration margins are not true population values, there is a danger that implausible weights can be obtained. This was tested in this paper in the second

part, using distance functions other than linear. When the calibration margins were drawn from the sample, some implausible weights were found. Our data set was not simple, as is often the case in real-life. Our purpose was not to get the ideal estimates only, but those that are realistic. Our eight indicators were used in exercises and hence well illustrated the situation in survey practice. Two of these eight indicators were found to be difficult to estimate well due to the lack of good macro and micro auxiliary variables. Fortunately, we found that the reweights for the rest of other indicators substantially reduced the bias. It should be noted that the data environment was demanding and indicators concerning crimes due to violence even more so. Their prevalence was hard to examine in surveys, in general, due to their sensitivity, and it was expected that the weights would not help much. The drawback is the reality that the outcome depends on the respondents, since the weights can only use such data; if certain groups are represented to a small extent among the respondents, the weights cannot help.

We examined easier variables, such as income and worry due to crime, and the reweights helped much more. We cannot make any straightforward conclusion about the weights applied, however our study shows that the combination of the response propensity weighting and calibration is the best of all four methods. This takes advantage both of micro and macro auxiliary variables. The first ones were especially used in response propensity weighting and the second ones in the calibration performed from these first weights. This two-stage strategy is not often used, but we recommend it. It is definitely better than the often used pure calibration with linear distance function. The calibration is good to be used as the ending method since it ensures that estimates derived from the known population margins are correct.

We conducted tests with linear calibration and found that it works correctly if the variation of the starting weights is moderate, the number of margins is not big and the margins are accurate. In other cases, linear calibration may lead to negative weights. It is also possible that logistic calibration may give the weights below one; if the bounds used are strict, the algorithm of the method does not always converge. Further investigations with weighting adjustments are needed. Special attention could be paid to get good predicting auxiliary variables with a high quality, both at micro and macro level.

REFERENCES

BRICK, J. M., (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review, Journal of Official Statistics, 29, 3, pp. 329–353.

BRICK, J. M., JONES, M. E., (2008). Propensity to respond and Nonresponse Bias, METRON – International Journal of Statistics, LXVI, 1, pp. 51–73.

DEVILLE, J-C., SÄRNDAL, C-E., (1992). Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, pp. 376–382.

DEVILLE, J-C., SÄRNDAL, C-E., SAUTORY, O., (1993). Generalized Raking Procedures in Survey Sampling, Journal of the American Statistical Association, pp. 1013–1020.

EKHOLM, A., LAAKSONEN, S., (1991). Weighting via Response Modelling in the Finnish Household Budget Survey, Journal of Official Statistics. 7.2, pp. 325–337.

ESS _ European Social Survey, (2014). Documentation of ESS Post-Stratification Weights.

http://www.europeansocialsurvey.org/docs/methodology/ESS_post_stratification_weights_documentation.pdf.

ESTEVAO, V. M., SÄRNDAL, C-E., (2006). Survey Estimates by Calibration on Complex Auxiliary Information, International Statistical Review 74, 2, pp. 127–147.

HAZIZA, D., LESAGE, E. (2016). A Discussion of Weighting Procedures for Unit Nonresponse, Journal of Official Statistics 32, 1, pp. 129–145, http://dx.doi.org/10.1515/JOS-2016-0006.

HENRY, K. A., VALLIANT, R., (2015). A design effect measure for calibration weighting in single stage samples, Survey Methodology 41, 2, pp. 315–331.

HOLT, D., SMITH, T. M. F., (1979). Post-Stratification. Journal of the Royal Statistical Society, Series A (General). Vol. 142, pp. 33–46.

KOTT, P., (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. Survey Methodology 32, 2, pp. 133–142.

KOTT, P. CHANG, T., (2008). Can calibration be used for 'nonignorable' nonresponse, Proceedings of Joint Statistical Meeting. Section of Survey Research, pp. 251–260.

KOTT, P., CHANG, T., (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse, Journal of the American Statistical Association, pp. 105–491.

LUMLEY, T., (2013). Survey: Analysis of Complex Survey Samples. R package version 3.29, URL http://CRAN.R-project.org/package=survey.

LUMLEY, T., SHAW, P., DAI, J. Y., (2011). Connections between Survey Calibration Estimators and Semiparametric Models for Incomplete Data. International Statistical Review 79, 2, pp. 200–220, DOI: 10.1111/j.1751-5823.2011.00138.x.

LAAKSONEN, S., (2007). Weighting for Two-Phase Surveyed Data. Survey Methodology, December Vol. 33, No. 2, pp. 121–130, Statistics Canada.

LAAKSONEN, S., (2015). Sampling Design Data File, Survey Statistician 72, pp 61–66.

LAAKSONEN, S., HEISKANEN, M., (2014). Comparison of Three modes for a Crime Victimization Survey, Journal of Survey Statistics and Methodology, 2 (4), pp. 459–483, DOI 10.1093/jssam/smu018.

LAAKSONEN, S., KEMPPAINEN, T., STJERNBERG, M., KORTTEINEN, M., VAATTOVAARA, M., LÖNNQVIST, H., (2015). Tackling City-Regional Dynamics in a Survey Using Grid Sampling. Survey. Research Methods by the European Survey Research Association, Vol 9, No 1, pp. 55–65, www.surveyresearch.org.

Le GUENNEC. J., SAUTORY. O., (2005). CALMAR 2: Une Nouvelle Version de la Macro Calmar de Redressment D'Échantillion Par Calage, http://vserver-insee.nexen.net/jms2005/site/files/documents/2005/327_1-JMS2002_SESSION1_LE-GUENNEC-SAUTORY_CALMAR-2_ACTES.PDF.

LITTLE. R. J. A., (1986). Survey Nonresponse Adjustments for Estimates of Means, International Statistical Review, 54, pp. 139–157.

LITTLE, R. J. A., RUBIN, D. B., (2002). Statistical Analysis with Missing Data, 2nd Edition. Wiley.

LUNDQUIST, P., SÄRNDAL, C-E., (2013). Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey. Journal of Official Statistics 29, 4, pp. 557–582, DOI: org/10.2478/jos-2013-0040.

LUNDSTRÖM, S., SÄRNDAL, C-E., (1999). Calibration as a Standard Method for Treatment of Nonresponse, Journal of Official Statistics, 15, 2, pp. 305–327.

McCORMACK, K., (2006). The calibration software CALMAR – What is it? Central Statistics Office Ireland, http://vesselinov.com/CalmarEngDoc.pdf.

ROBINS, J. M., ROTNITZKY, A., ZHAO, L.-P., (1994). Estimation of regression coefficients when some regressors are not always observed, Journal of the American Statistical Association 89, pp. 846–866.

PLIKUSAS, A., PUMPUTIS, D., (2010). Estimation of the finite population covariance using calibration. Nonlinear Analysis: Modelling and Control, Vol. 15, 3, pp. 325–340.

SAUTORY, O., (2003). CALMAR 2: A New Version of the Calmar Calibration Adjustment Program, Proceedings of Statistics Canada's Symposium: Challenges in Survey Taking for the Next Decade.

SÄRNDAL, C-E., (2007). Calibration Approach in Survey Theory and Practice, Survey Methodology, 33, No. 2, pp. 99–119.

SMITH, T. M. F., (1991). Post-stratification, The Statistician 40, pp. 315–323.

VALLIANT, R., DEVER, J. A., KREUTER, F., (2013). Practical Tools for Designing and Weighting Survey Samples, Statistics for Social and Behavioral Sciences. Springer.

WITTENBERG, M., (2009). Sample Survey Calibration: An Information-theoretic perspective, A Southern Africa Labour and Development Research Unit Working Paper Number 41, Cape Town: SALDRU, University of Cape Town.

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 61–74, DOI 10.21307/stattrans-2018-004

A NEW AND UNIFIED APPROACH IN GENERALIZING THE LINDLEY'S DISTRIBUTION WITH APPLICATIONS

Lahsen Bouchahed[1], **Halim Zeghdoudi**[2]

ABSTRACT

This paper proposes a new family of continuous distributions with one extra shape parameter called the generalized Zeghdoudi distributions (GZD). We investigate the shapes of the density and hazard rate function. We derive explicit expressions for some of its mathematical quantities. Various statistical properties like stochastic ordering, moment method, maximum likelihood estimation, entropies and limiting distribution of extreme order statistics are established. We prove the flexibility of the new family by means of applications to several real data sets.

Key words: Lindley distribution, exponential distribution, Gamma distribution, stochastic ordering, maximum-likelihood estimation.

1. Introduction

Statistical models are very useful in describing and predicting real-world phenomena. Numerous extended distributions have been extensively used over the last decades for modelling data in several areas. Recent developments focus on defining new families that extend well-known distributions and at the same time provide great flexibility in modelling data in practice. Thus, many lifetime distributions for modeling lifetime data such as Lindley, exponential, Gamma, Weibull, Lognormal, Akash, Shanker, Sujatha, Amarendra distributions have been proposed in the statistical literature.

The probability density function of Lindley distribution is given by

$$f_1(x,\theta) = \frac{\theta^2}{\theta+1}(1+x)\exp(-\theta x);\ x>0,\ \theta>0$$

It has been generalized many times by researchers including Zakerzadeh and Dolati (2009), Bakouch et al. (2012), Shanker et al. (2013), Elbatal et al. (2013), Ghitany et al. (2013), Suigh et al. (2014), Abouamoh et al. (2015).

Shanker used a procedure based on certain mixtures of the gamma and exponential distributions to obtain three new distributions:
the Akash distribution, see (Shanker $2015a$), whose probability density function is given by

$$f_2(x,\theta) = \frac{\theta^3\left(1+x^2\right)}{\theta^2+1}\exp(-\theta x);\ x>0,\ \theta>0$$

[1]LaPS laboratory, Badji-Mokhtar University, Box 12, Annaba, 23000,ALGERIA. E-mail: lbouchahed@hotmail.com
[2]LaPS laboratory, Badji-Mokhtar University, Box 12, Annaba, 23000,ALGERIA. E-mail: halim.zeghdoudi@univ-annaba.dz

the Aradhana distribution, see (Shanker $2016a$), whose probability density function is given by

$$f_2(x,\theta) = \frac{\theta^3 (1+x)^2}{\theta^2+2\theta+2} \exp(-\theta x);\ x>0,\ \theta>0$$

the Sujatha distribution, see (Shanker $2016b$), whose probability density function is given by

$$f_3(x,\theta) = \frac{\theta^3 \left(1+x+x^2\right)}{\theta^2+\theta+2} \exp(-\theta x);\ x>0,\ \theta>0$$

the Amarendra distribution, see (Shanker $2016c$), whose probability density function is given by

$$f_4(x,\theta) = \frac{\theta^4 \left(1+x+x^2+x^3\right)}{\theta^3+\theta^2+2\theta+6} \exp(-\theta x);\ x>0,\ \theta>0.$$

the Devya distribution, see (Shanker $2016d$), whose probability density function is given by

$$f_5(x,\theta) = \frac{\theta^5 \left(1+x+x^2+x^3+x^4\right)}{\theta^4+\theta^3+2\theta^2+6\theta+24} \exp(-\theta x);\ x>0,\ \theta>0.$$

the Shambhu distribution, see (Shanker $2016e$), whose probability density function is given by

$$f_6(x,\theta) = \frac{\theta^6 \left(1+x+x^2+x^3+x^4+x^5\right)}{\theta^5+\theta^4+2\theta^3+6\theta^2+24\theta+120} \exp(-\theta x);\ x>0,\ \theta>0.$$

In this paper we introduce a new one-parameter family of continuous distributions by considering a polynomial exponential family, which contains Akash, Sujatha, Amarendra, Devya and Shambhu distributions as particular cases.

The rest of the paper is outlined as follows. Sections (2-10) are devoted to present various statistical properties like stochastic ordering, moment method, maximum likelihood estimation, entropies, and limiting distribution of extreme order statistics are established. In section 11 we present the new distribution called Zeghdoudi distribution, and in section 12 we give numerical examples.

We derive explicit expressions for some of its mathematical quantities. Various statistical properties like stochastic ordering, moment method, maximum likelihood estimation, entropies and limiting distribution of extreme order statistics are established. We prove the flexibility of the new family by means of applications to several real data sets.

Basic Theory

Suppose that X is random variable taking values in $]0,\infty)$, and that the distribution of X depends on an unspecified parameter θ taking values in $]0,\infty)$. So, the distribution of X might be absolutely continuous or discrete. In both cases, let f_θ be the probability distribution function with respect to the Lebesgue measure or to the counting measure on a countable set including discontinuity jumps of f_θ.

The distribution of X is a one-parameter polynomial exponential family and the probability density function can be written as

$$f_{GZD}(x;\theta) = h(\theta)\, p(x) \exp(-\theta x), \qquad \theta, x > 0$$

where $h(\theta)$ is real-valued functions on $]0,\infty)$, and where $p(x) = \sum_{k=0}^{n} a_k x^k$ and $a_n \neq 0$, is polynomial functions on $]0,\infty)$. Moreover, k is a positive integer and a_k is positive real number with $a_n \neq 0$. We can check immediately:

1) It is non-negative for $x > 0$;

2) $P[a < x < b] = \int_a^b f_\theta(x)\, dx$;

3) $\int_0^\infty f_\theta(x)\, dx = 1$ for $h(\theta) = \dfrac{1}{\sum_{k=0}^{n} a_k \frac{k!}{\theta^{k+1}}}$.

Now, the probability density function of Generalized Zeghdoudi Distribution X is:

$$f_{GZD}(x;\theta) = \frac{\sum_{k=0}^{n} a_k x^k \exp(-\theta x)}{\sum_{k=0}^{n} a_k \frac{k!}{\theta^{k+1}}}, \qquad \theta, x > 0 \tag{1}$$

Examples and Special Cases

Many of the special distributions studied in this work are general exponential families, at least with respect to some of their parameters. On the other hand, most commonly, a parametric family fails to be a general exponential family because the support set depends on the parameter.

2. General one-parameter distribution and some properties

In this section, we give the general one-parameter distribution and study its properties.

The first and second derivatives of $f_{GZD,\theta}(x)$

$$\frac{d}{dx} f_{GZD}(x;\theta) = \frac{\left[(a_1 - \theta a_0) + \ldots + (n a_n - \theta a_{n-1}) x^{n-1} + a_n x^n\right] \exp(-\theta x)}{\sum_{k=0}^{n} a_k \frac{k!}{\theta^{k+1}}} = 0$$

gives $x_1, x_{2,\ldots,} x_n$ solutions.

We can find easily the cumulative distribution function (c.d.f) of the general one-parameter distribution:

$$F_{GZD}(x) = 1 - \frac{\sum_{k=0}^{n} \frac{a_k \Gamma(k+1, x\theta)}{\theta^{k+1}}}{\sum_{k=0}^{n} a_k \frac{k!}{\theta^{k+1}}}; x, \theta > 0 \tag{2}$$

2.1. Survival and hazard rate function

Let

$$S_{GZD}(x) = 1 - F_{GZD}(x) = \frac{\sum_{k=0}^{n} \frac{a_k \Gamma(k+1, x\theta)}{\theta^{k+1}}}{\sum_{k=0}^{n} a_k \frac{k!}{\theta^{k+1}}}; x, \theta > 0$$

and

$$h_{GZD}(x) = \frac{f_{GZD}(x)}{1 - F_{GZD}(x)} = \frac{\sum_{k=0}^{n} a_k x^k \exp(-\theta x)}{\sum_{k=0}^{n} \frac{a_k \Gamma(k+1, x\theta)}{\theta^{k+1}}}$$

be the survival and hazard rate function, respectively.

Proposition1. *Let $h_\theta(x)$ be the hazard rate function of X. Then $h_\theta(x)$ is increasing for $\sum_0^m (k+1)(m-2k)a_{m-k}a_{k+1} \geq 0,\ m = 0, ..., 2n-1$*

Proof. According to Glaser (1980) and from the density function (2) we have

$$\rho(x) = -\frac{f'_{GZD}(x;\theta)}{f_{GZD}(x;\theta)} = -\frac{\sum_{k=1}^{n} k a_k x^{k-1}}{\sum_{k=0}^{n} a_k x^k} + \theta$$

After simple computations we obtain

$$\rho'(x) = \frac{\sum_{m=0}^{2n-1} \sum_{k=0}^{m} (k+1)(m-2k) a_{m-k} a_{k+1} x^{m-1}}{\left(\sum_{k=0}^{n} a_k x^k\right)^2}.$$

Which implies that $h_\theta(x)$ is increasing *for* $\sum_{k=0}^{m} (k+1)(m-2k)a_{m-k}a_{k+1} \geq 0,\ m = 0, ..., 2n-1$

3. Moments and related measures

The kth moment about the origin of the GZD is :

$$\mathbb{E}(X^i) = \frac{\sum_{k=0}^{n} \frac{a_k}{\theta^{k+i+1}} (k+i)!}{\sum_{k=0}^{n} a_k \frac{k!}{\theta^{k+1}}}, i = 1, 2, ...$$

Remark 2 The kth moment about the origin of the Lindley distribution is

$$\mathbb{E}(X^i) = \frac{i!(\theta + i + 1)}{\theta^i (\theta + 1)}$$

Corollary 1. *Let $X \sim GZD(\theta)$, the mean of X is:*

$$\mathbb{E}(X) = \frac{\sum_{k=0}^{n} \frac{a_k}{\theta^{k+2}} (k+1)!}{\sum_{k=0}^{n} a_k \frac{k!}{\theta^{k+1}}}.$$

Theorem 1. *Let $X \sim GZD(\theta)$, me =median(X) and $\mu = E(X)$. Then $me < \mu$*

Proof. According to the increasingness of $F(x)$ for all x and θ,

$$F_{GZD}(me) = \frac{1}{2}$$

and

$$F_{GZD}(\mu) = 1 - h(\theta) \sum_{k=0}^{n} \frac{a_k \Gamma\left(k+1, \theta h(\theta) \sum_{k=0}^{n} \frac{a_k}{\theta^{k+2}} (k+1)!\right)}{\theta^{k+1}}$$

Note that $\frac{1}{2} < F(\mu) < 1$. It is easy to check that $F(me) < F(\mu)$. To this end, we have $me < \mu$.■.
The coefficients of variation γ , skewness and kurtosis of the GZD have been obtained as

$$\begin{aligned} \gamma &= \frac{\sqrt{Var(X)}}{\mathbb{E}(X)} \\ skewness &= \frac{\mathbb{E}(X^3)}{(Var(X))^{\frac{3}{2}}} \\ kurtosis &= \frac{\mathbb{E}(X^4)}{(Var(X))^2}. \end{aligned}$$

4. Estimation of parameter

Let $X_1, ..., X_n$ be a random sample of GZD. The ln-likelihood function, $\ln l(x_i;\theta)$ is given by :

$$\ln l(x_i;\theta) = n \ln h(\theta) + \sum_{i=1}^{n} \ln \left(\sum_{k=0}^{m} a_k x_i^k \right) - \theta \sum_{i=1}^{n} x_i.$$

The derivative of $\ln l(x_i;\theta)$ with respect to θ is :

$$\frac{d \ln l(x_i;\theta)}{d\theta} = \frac{n\dot{h}(\theta)}{h(\theta)} - \sum_{i=1}^{n} x_i.$$

From the Zeghdoudi distribution (2), the method of moments (MoM) and the maximum likelihood (ML) estimators of the parameter θ are the same and it can be obtained by solving the following non-linear equation

$$\frac{\dot{h}(\theta)}{h(\theta)} - \bar{x} = 0, \text{ where } \dot{h}(\theta) = \frac{dh(\theta)}{d\theta} \tag{4}$$

The equation to be solved is

$$\sum_{k=0}^{m} \frac{a_k k!}{\theta^k} ((k+1) - \bar{x}\theta) = 0 \tag{5}$$

Note that we can solve the equation (5) exactly for $m \leq 4$ and for $m \geq 5$ the equation (5) is to be solved numerically.

Special cases

For $\mathbf{m = 0}$, we have $\hat{\theta}_{MV} = \frac{1}{\bar{x}}$

For $\mathbf{m = 1}$, we have $\hat{\theta}_{MV} = \frac{1}{2xa_0}\left(a_0 - xa_1 + \sqrt{x^2a_1^2 + a_0^2 + 6xa_0a_1}\right)$

For $\mathbf{m = 2}$, $\hat{\theta}_{MV}$ is one of the two solutions :

$$\left\{ \frac{\left(-a_1 + \sqrt{x^2a_2^2 + a_1^2 - 6a_0a_2 + 4xa_1a_2} + xa_2\right)}{a_0 - xa_1}, -\frac{\left(a_1 + \sqrt{x^2a_2^2 + a_1^2 - 6a_0a_2 + 4xa_1a_2} - xa_2\right)}{a_0 - xa_1} \right\}$$

For $\mathbf{m = 3}$ and $\mathbf{m = 4}$, we can solve exactly equation (5) using methods such as Cardan and Ferrari method

For $\mathbf{m \geq 5}$, according to Galois theorem, there is no general method to solve exactly equation (5).

5. Stochastic orders

Definition 1. *Consider two random variables X and Y. Then X is said to be smaller than Y in the: a) Stochastic order ($X \prec_s Y$), if $F_X(t) \geq F_Y(t)$, $\forall t$.*
b) Convex order ($X \leq_{cx} Y$), if for all convex functions ϕ and provided expectation exist, $E[\phi(X)] \leq E[\phi(Y)]$.
c) Hazard rate order ($X \prec_{hr} Y$), if $h_X(t) \geq h_Y(t)$, $\forall t$.
d) Likelihood ratio order ($X \prec_{lr} Y$), if $\frac{f_X(t)}{f_Y(t)}$ is decreasing in t.

Remark 3 *Likelihood ratio order⇒Hazard rate order⇒Stochastic order.*
If $E[X] = E[Y]$, then Convex order⇔Stochastic order.

Theorem 4. *Let $X_i \sim GZD(\theta_i), i = 1,2$ be two random variables. If $\theta_1 \geq \theta_2$, then $X_1 \prec_{lr} X_2, X_1 \prec_{hr} X_2, X_1 \prec_s X_2$ and $X_1 \leq_{cx} X_2$.*

Proof. We have

$$\frac{f_{X_1}(t)}{f_{X_2}(t)} = \frac{\sum_{k=0}^{n} a_k \frac{k!}{\theta_2^{k+1}}}{\sum_{k=0}^{n} a_k \frac{k!}{\theta_1^{k+1}}} e^{-(\theta_1 - \theta_2)t}.$$

For simplification, we use $\ln\left(\frac{f_{X_1}(t)}{f_{X_2}(t)}\right)$. Now, we can find

$$\frac{d}{dt} \ln\left(\frac{f_{X_1}(t)}{f_{X_2}(t)}\right) = -(\theta_1 - \theta_2)$$

To this end, if $\theta_1 \geq \theta_2$, we have $\frac{d}{dt}\ln\left(\frac{f_{X_1}(t)}{f_{X_2}(t)}\right) \leq 0$. This means that $X_1 \prec_{lr} X_2$. Also, according to *Remark 3* the theorem is proved.

6. Mean Deviations

These are two mean deviation: about the mean and about the median, defined as $MD_1 = \int_0^\infty |x-\mu| f(x)\,dx$ and $MD_2 = \int_0^\infty |x-me| f(x)\,dx$ respectively, where $\mu = E(X)$

and me =Median(X). The measures MD_1 and MD_2 can be computed using the following simplified formulas

$$MD_1 = 2\mu F(\mu) - 2\int_0^\mu xf(x)\,dx$$

$$MD_2 = \mu - 2\int_0^{me} xf(x)\,dx$$

7. Extreme domain of attraction

As to the extreme value stability, the cdf F_{GZD} is in the Gumbel extreme value domain of attraction, that is, there exist two sequences $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ of real numbers such that for any $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to+\infty} F_{GZD}(a_n x + b_n)^n = \exp(-\exp(-x)).$$

This follows from Formula 1.2.4 in theorem 1.2.1 (De Haan and Ferreira (2006)) since we have

$$\begin{aligned}\lim_{t\to\infty}\frac{1-F_{GZD}(t+xf(t))}{1-F_{GZD}(t)} &= \lim_{t\to\infty}\frac{f_{GZD}(t+xf(t))}{f_{GZD}(t)} \\ &= \lim_{t\to\infty}\frac{\sum_{k=0}^n a_k\,(xf(t)+t)^k\, e^{(-\theta(xf(t)+t))}}{\sum_{k=0}^n a_k t^k e^{(-\theta t)}} = \exp(-x),\end{aligned}$$

(such formula is called Γ-variation). Then, F_{GZD} lies in the Gumbel extreme domain of attraction. In his case, $f(t) = \frac{1}{\theta}$. So, for (as in the invoked theorem) $a_n = f(F_{GZD}^{-1}(1-1/n)) = \frac{1}{\theta}$ and $b_n = F_{GZD}^{-1}(1-1/n)$, we have

$$\lim_{n\to+\infty} F_{GZD}(a_n x + b_n)^n = \exp(-\exp(-x))$$

8. Estimation of the Stress-Strength Parameter

The stress-strength parameter (R) plays an important role in the reliability analysis as it measures the system performance. Moreover, R provides the probability of a system failure, the system fails whenever the applied stress is greater than its

strength, i.e. $R = P(X > Y)$. Here $X \sim GZD(\theta_1)$ denotes the strength of a system subject to stress Y, and $Y \sim GZD(\theta_2)$, X and Y are independent of each other. In our case, the stress-strength parameter R is given by

$$\begin{aligned} R &= P(X > Y) = \int_0^{\infty} S_X(y) f_Y(y) dy \\ &= \frac{\int_0^{\infty} \sum_{k=0}^{n} \frac{a_k \Gamma(k+1, y\theta_1)}{\theta_1^{k+1}} \sum_{k=0}^{n} a_k y^k \exp(-\theta_2 y) dy}{\left(\sum_{k=0}^{n} a_k \frac{k!}{\theta_1^{k+1}}\right)\left(\sum_{k=0}^{n} a_k \frac{k!}{\theta_2^{k+1}}\right)} \end{aligned}$$

9. Lorenz curve

The Lorenz curve is often used to characterize income and wealth distributions. The Lorenz curve for a positive random variable X is defined as the graph of the ratio

$$L(F(x)) = \frac{E(X|X \leq x)F(x)}{E(X)}$$

against $F(x)$ with the properties $L(p) \leq p, L(0) = 0$ and $L(1) = 1$. If X represents annual income, $L(p)$ is the proportion of total income that accrues to individuals having the $100\ p\%$ lowest incomes. If all individuals earn the same income then $L(p) = p$ for all p. The area between the line $L(p) = p$ and the Lorenz curve may be regarded as a measure of inequality of income, or more generally, of the variability of X. For the exponential distribution, it is well known that the Lorenz curve is given by

$$L(p) = p\{p + (1-p)log(1-p)\}.$$

For the GZ distribution in (3),

$$E(X|X \leq x)F_{GZD}(x) = \frac{\sum_{k=0}^{n} \frac{a_k}{\theta^{k+2}} (k+1)!}{\sum_{k=0}^{n} a_k \frac{k!}{\theta^{k+1}}} \left(1 - \frac{\sum_{k=0}^{n} \frac{a_k \Gamma(k+1, x\theta)}{\theta^{k+1}}}{\sum_{k=0}^{n} a_k \frac{k!}{\theta^{k+1}}}\right)$$

10. Entropies

It is well known that entropy and information can be considered as measures of uncertainty of probability distribution. However, there are many relationships established on the basis of the properties of entropy.

An entropy of a random variable X is a measure of variation of the uncertainty. Rényi entropy is defined by

$$J(\gamma)=\frac{1}{1-\gamma}log\left\{\int f^{\gamma}(x)\,dx\right\}$$

where $\gamma>0$ and $\gamma\neq 1$. For the GZ distribution in (1), note that, for γ integer we have

$$\begin{aligned}\int f_{GZD}^{\gamma}(x)\,dx &= \frac{\int\left(\sum_{k=0}^{n}a_k x^k\right)^{\gamma}\exp\left(-\theta\gamma x\right)dx}{\left(\sum_{k=0}^{n}a_k\frac{k!}{\theta^{k+1}}\right)^{\gamma}}\\ &= \frac{\sum_{k=0}^{n}b_k(\gamma)\int x^{k\gamma}\exp\left(-\theta\gamma x\right)dx}{\left(\sum_{k=0}^{n}a_k\frac{k!}{\theta^{k+1}}\right)^{\gamma}}\end{aligned}$$

where, $\int x^{k\gamma}\exp\left(-\theta\gamma x\right)dx=-\frac{1}{\theta\gamma(\theta\gamma)^{k\gamma}}\Gamma\left(k\gamma+1,x\theta\gamma\right)$ and $b_k(\gamma)$ in function the a_k and γ.

Now, the Rényi entropy as given by

$$J(\gamma)=\frac{1}{1-\gamma}\log\left[\frac{\frac{(k\gamma)!}{\theta\gamma(\theta\gamma)^{k\gamma}}\sum_{k=0}^{n}b_k(\gamma)}{\left(\sum_{k=0}^{n}a_k\frac{k!}{\theta^{k+1}}\right)^{\gamma}}\right]$$

if $\gamma\to\infty$, we find Shannon entropy.

11. Zeghdoudi distribution(ZD) and immediate properties

In this section we will introduce a new distribution called Zeghdoudi distribution(ZD) (see Zeghdoudi and Messadia 2018), which is a member of the new family.

The density function of $X\rightsquigarrow$ Zeghdoudi distribution is given by:

$$f_{ZD}(x;\theta)=\begin{cases}\frac{\theta^3x(1+x)e^{-\theta x}}{\theta+2} & x,\theta>0\\ 0, & \text{otherwise.}\end{cases}$$

We note that ZD distribution is a member of the new family where $n=2,a_0=0,a_1=a_2=1$ using formula (1). Therefore, the mode of ZD is given by

$$\text{mode}(X)=\frac{-\theta+2+\sqrt{\theta^2+4}}{2\theta}\quad\text{for }\theta>0$$

We can find easily the cumulative distribution function(c.d.f) of the ZD :

$$F_{ZD}(x)=1-\left(\frac{\theta^2x^2+\theta(\theta+2)x+\theta+2}{\theta+2}\right)e^{-\theta x};x>0,\theta>0$$

From the Zeghdoudi distribution, the method of moments (MoM) and the maximum likelihood (ML) estimators of the parameter θ are the same and can be obtained by solving the following non-linear equation

$$\frac{3}{\theta} - \frac{1}{\theta+2} - \bar{x} = 0$$

$$\hat{\theta}_M = \hat{\theta}_{ML} = \left\{ \frac{1}{\bar{x}} \left(-\bar{x} + \sqrt{4\bar{x} + \bar{x}^2 + 1} + 1 \right) \right\}$$

We can show that the estimator of θ is positively biased.

12. Simulation and Goodness of Fit

12.1. Simulation study

We can see that the equation $F(x) = u$, where u is an observation from the uniform distribution on $(0,1)$, cannot be solved explicitly in x (cannot use **lambert W function in the case** $k \geq 2$), the inversion method for generating random data from the GZ distribution fails. However, we can use the fact that the GZ distribution is a mixture of gamma (k,θ) and gamma $(k+1,\theta)$ distributions:

$$f_{GZD}(x;\theta) = p(\theta)\ gamma(k,\theta) + (1-p(\theta))\, gamma(k+1,\theta), 0 < p(\theta) < 1$$

In this subsection, we investigate the behaviour of the ML estimators for a finite sample size (n). A simulation study consisting of the following steps is being carried out $N = 10000$ times for selected values of $(\theta;n)$, where $\theta = 0.01, 0.1, 0.5, 1, 3, 10$ and $n = 10, 30, 50$.

- Generate $U_i \sim Uniform(0,1)$, $i = 1,...n$.
- Generate $Y_i \sim Gamma(k,\theta)$, $i = 1,...n$.
- Generate $Z_i \sim Gamma(k+1,\theta)$, $i = 1,...n$.
- If $U_i \leq p(\theta)$, then set $X_i = Y_i$, otherwise, set $X_i = Y_i$, $i = 1,...n$,

$$average\ bais(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{\theta}_i - \theta\right)$$

and the average square error.

$$MSE(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{\theta}_i - \theta\right)^2, i = 1,...N.$$

12.2. Applications

Example 1

Tables 1 and 2 represent the Data of survival times (in months) of (94,91) guinea individus infected with Ebola virus.

Table 1 Chi-Square Values for Lindley and Zeghdoudi distributions

Survival time $m = 3.17$	Obs freq	LD $\hat{\theta} = 0.522$	ZD $\hat{\theta} = 0.852$
$[0,2[$	32	38.217	33.252
$[2,4[$	35	28.16	32.366
$[4,6[$	17	15.089	17.799
$[6,8[$	7	7.33	7. 133
$[8,10]$	3	3.152	2.418
Total	94	94	94
χ^2	-	2.9244	**0.40020**

Table 2 Chi-Square Values for Lindley and Zeghdoudi distributions

Survival time $m = 3$	Obs freq	LD $\hat{\theta} = 0.548$	ZD $\hat{\theta} = 0.896$
$[0,2[$	35	39.100	34.56
$[2,4[$	32	27.390	31.497
$[4,6[$	16	13.92	16.206
$[6,8[$	6	6.247 5	6.069 7
$[8,10]$	2	2.618 9	1.921 8
Total	91	91	91
χ^2	-	1.616 5	**0.01995**

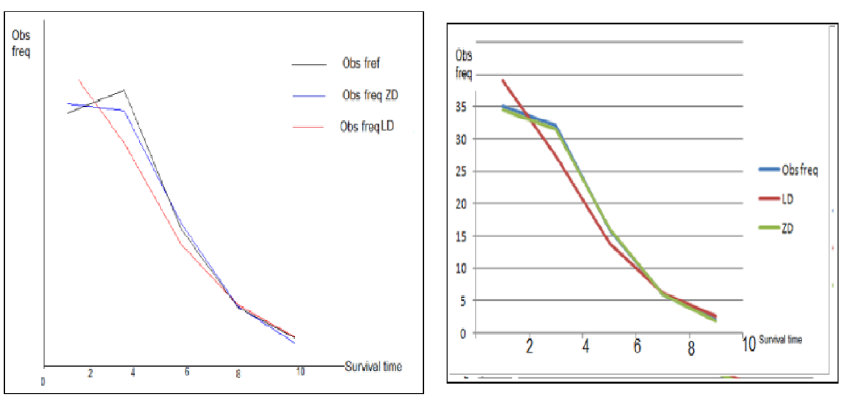


Figure 1: Plots of the density function of LD and ZD

Example 2

Now, we compare one-parameter (Aradhana, Akash, Shanker, Amarendra, Devya, Shambhu) distributions see Shanker (2015a, 2015b, 2016a, 2016b, 2016c, 2016d, 2016e) and two-parameter (Gamma, Weibull, Lognormal) distributions with Zeghdoudi distribution (see tables 3 and 4).

Table 3 Comparison of ZD with one-parameter distributions

Data	Distribution	θ	log-likelihood	Kolmogrov-Smirnov
data 1	ZD	1.365	**-52.4**	**0.292**
$n = 20$	Aradhana	1.123	-56.4	0.302
$m = 1.9$	Akash	1.157	-59.5	0.320
$s = 0.704$	Shanker	0.804	-59.7	0.315
	Amarendra	1.481	-55.64	0.286
	Devya	1.842	-54.50	0.268
	Shambhu	2.215	-53.9	0.254

Data	Distribution	θ	log-likelihood	Kolmogrov-Smirnov
data 2	ZD	0.095	**-239.7**	**0.251**
$n = 25$	Aradhana	0.094	-242.2	0.274
$m = 30.811$	Akash	0.097	-240.7	0.266
$s = 7.253$	Shanker	0.063	-252.3	0.326
	Amarendra	1.481	-233.41	0.225
	Devya	1.842	-227.68	0.193
	Shambhu	2.215	-223.40	0.167

Table 4 Comparison of ZD with two-parameter distributions

Data	Distribution	β	θ	log-likelihood	Kolmogrov-Smirnov
data 3	ZD	—	0.107	**-58.67**	**0.081**
$n = 15$	Gamma	1.442	0.052	-64.197	0.102
$m = 27.54$	Weibull	1.306	0.034	-64.026	0.450
$s = 20.06$	Lognormal	1.061	2.931	-65.626	0.163
data 4	ZD	—	0.0167	**-142.32**	**0.108**
$n = 25$	Gamma	1.794	0.010	-152.371	0.135
$m = 178.32$	Weibull	1.414	0.005	-152.440	0.697
$s = 131.09$	Lognormal	0.891	4.880	-154.092	0.155

According to tables 1, 2, 3, 4 and figures 1, 2, we can observe that Zeghdoudi distribution provide smallest -LL and K-S values as compared to one-parameter (Aradhana, Akash, Shanker, Amarendra, Devya, Shambhu) distributions, and two-parameter (Gamma, Weibull, Lognormal), and hence best fits the data among all the models considered.

13. Conclusions

In this work we propose a one-parameter family GZD. Several properties have been discussed: moments, cumulants, characteristic function, failure rate function,

stochastic ordering, the maximum likelihood method and the method of moments. The LD does not provide enough flexibility for analyzing and modelling different types of lifetime data and survival analysis. But GZD is flexible, simple and easy to handle. Two real data sets are analyzed using the new distribution and are compared with five immediate sub-models mentioned above in addition to other distributions (Lindley, Exponential, Gamma, Weibull, Lognormal distributions). The results of the comparisons confirm the goodness of fit of GZ distribution. We hope our new distribution family might attract wider sets of applications in lifetime data reliability analysis and actuarial sciences. For future studies, we can go more generally by using $p(x;\theta)$. Also, we can explain the derivation of posterior distributions for the GZ distribution under Linex loss functions and squared error using non-informative and informative priors (the extension of Jeffreys and Inverted Gamma priors) respectively.

Annexe

Data set 1: represents the lifetime data relating to relief times (in minutes) of 20 patients receiving an analgesic and reported by Gross and Clark (1975, P. 105). The data are as follows: 1.1, 1.4, 1.3, 1.7, 1.9, 1.8, 1.6, 2.2, 1.7, 2.7, 4.1, 1.8, 1.5, 1.2, 1.4, 3.0, 1.7, 2.3, 1.6, 2.0

Data set 2: is the strength data of glass of the aircraft window reported by Fuller et al. (1994): 18.83, 20.80, 21.657, 23.03, 23.23, 24.05, 24.321, 25.50, 25.52, 25.80, 26.69, 26.77, 26.78, 27.05, 27.67, 29.90, 31.11, 33.20, 33.73, 33.76, 33.89, 34.76, 35.75, 35.91, 36.98, 37.08, 37.09, 39.58, 44.045, 45.29, 45.381

Acknowledgment

The authors acknowledge Editor-in-Chief, Pr. Patryk Barszcz, of this journal for the constant encouragement to finalize the paper. Further, the authors acknowledge profound thanks to anonymous referees for giving critical comments, which have immensely improved the presentation of the paper. Also, this work was financed by European Mathematical Society.

REFERENCES

ASGHARZADEH, A., BAKOUCH, H, S., ESMAEILI, L., (2013). Pareto Poisson-Lindley distribution and its application, Journal of Applied Statistics, pp. 1–18.

FULLER, E.J., FRIEMAN, S., QUINN, J., QUINN, G., CARTER, W. Fracture mechanics approach to the design of glass aircraft windows: A case study, SPIE Proc 2286, pp. 419–430 (1994).

GHITANY, M, E., AL-MUTAIRI, D, K., NADARAJAH, S., (2008a). Zero-truncated Poisson-Lindley distribution and its application, Math. Comput. Simulation, 79, pp. 279–287.

GHITANY, M, E., ATIEH, B., NADARAJAH, S., (2008b). Lindley distribution and its applications. Mathematics and Computers in Simulation, 78, pp. 493–506.

GROSS, A,J., CLARK, V, A.,(1975). Survival Distributions: Reliability Applications in the Biometrical Sciences, John Wiley, New York.

LAURENS DE HAAN, FERREIRA, A., (2006). Extreme value theory: An introduction, Springer.

LAWLESS, J, F., (2003). Statistical models and methods for lifetime data, Wiley, New York.

LEADBETTER,M,R., LINDGREN, G., ROOTZén, H., (1987). Extremes and Related Properties of Random Sequences and Processes, Springer Verlag, New York.

LINDLEY, D, V., (1958). Fiducial distributions and Bayes' theorem, Journal of the Royal Society, series B, 20, pp. 102–107.

SANKARAN, M., (1970). The discrete Poisson-Lindley distribution, Biometrics, 26, pp. 145–149.

SHANKER, R., SHARMA, S., SHANKER, R., (2013). A two-parameter Lindley distribution for modeling waiting and survival times data. Applied Mathematics, Vol. 4, pp. 363–368.

SHANKER, R., (2015 a). Akash distribution and Its Applications. International Journal of Probability and Statistics, 4 (3), pp. 65–75.

SHANKER, R., (2015 b). Shanker distribution and Its Applications, International Journal of Statistics and Applications, 5 (6), pp. 338–348.

SHANKER, R., (2016 a). Aradhana distribution and Its Applications, International Journal of Statistics and Applications, 6 (1), pp. 23–34.

SHANKER, R., (2016 b). Sujatha distribution and Its Applications, Statistics in Transition new series, 17 (3), pp. 391–410.

SHANKER, R., (2016 c). Amarendra distribution and Its Applications, American Journal of Mathematics and Statistics, 6 (1), pp. 44–56.

SHANKER, R., (2016 d). Devya distribution and Its Applications, International Journal of Statistics and Applications, 6 (4), pp. 189–202.

SHANKER, R., (2016 e). Shambhu Distribution and Its Applications, International Journal of Probability and Statistics, 5 (2), pp. 48–63.

ZAKERZADAH, H., DOLATI, A., (2010). Generalized Lindley distribution, J. Math. Ext, 3(2), pp. 13–25.

ZEGHDOUDI, H., NEDJAR, S., (2016a). Gamma Lindley distribution and its application, Journal of Applied Probability and Statistics, Vol. 11 (1), 129–138.

ZEGHDOUDI, H., NEDJAR, S., (2016b). On gamma Lindley distribution: Properties and Simulations, Journal of Computational and Applied Mathematics, Vol 298, pp 167–174.

ZEGHDOUDI, H., NEDJAR, S., (2016c). A pseudo Lindley distribution and its application, Journal Afrika Statistika, Vol. 11 (1), pp. 923–932.

ZEGHDOUDI, H., MESSADIA,H., (2018). Zeghdoudi Distribution and its Applications. International Journal of Computing Science and Mathematics, In Press.

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 75–85, DOI 10.21307/stattrans-2018-005

CANONICAL CORRELATION ANALYSIS IN THE CASE OF MULTIVARIATE REPEATED MEASURES DATA

Mirosław Krzyśko[1], **Wojciech Łukaszonek**[2], **Waldemar Wołyński**[3]

ABSTRACT

In this paper, we present, in the real example, canonical variables applicable in the case of multivariate repeated measures data under the following assumptions: (1) multivariate normality for the vector of observations and (2) Kronecker product structure of the positive definite covariance matrix. These variables are especially useful when the number of observations is not large enough to estimate the covariance matrix, and thus the traditional canonical variables fail. Computational schemes for maximum likelihood estimates of required parameters are also given.

Key words: canonical correlation analysis, repeated measures data (doubly multivariate data), Kronecker product covariance structure, maximum likelihood estimates.

1. Introduction

Suppose that we have a sample of n objects characterized by $(p+q)$-variables $X_1,\ldots,X_p,X_{p+1},\ldots,X_{p+q}$ measured in T different time-points or physical conditions. Such data are often referred to in the statistical literature as multivariate repeated data or doubly multivariate data. Analysis of such data is complicated by the existence of correlation among the measurements of different variables as well as correlation among measurements taken at different time points. If we take observations on $(p+q)$-variables at T time-points, then these observations can be represented as $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p,\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q}$, where $\boldsymbol{x}_i$ are T-vectors. Let $\mathrm{Cov}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ be the covariance between the T-vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. When we choose $\mathrm{Cov}(\boldsymbol{x}_i,\boldsymbol{x}_j) = (\sigma_{ij}\boldsymbol{V})$, $i,j = 1\ldots,p,p+1,\ldots,p+q$, and assume normality, then the distribution of the $(p+q)$-random vectors is

$$\mathrm{vec}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p,\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q}) \sim N_{(p+q)T}(\boldsymbol{\mu},\boldsymbol{\Omega}),$$

where $\boldsymbol{\mu} = \mathrm{vec}(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_{p+q})$.

The covariance matrix $\boldsymbol{\Omega}$ is positive definite. Its estimate $\hat{\boldsymbol{\Omega}}$ is positive definite with probability 1 if and only if $n > (p+q)T$ (see, e.g., Giri (1996), p. 93).

[1]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. Interfaculty Institute of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland. E-mail: mkrzysko@amu.edu.pl

[2]Interfaculty Institute of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland. E-mail: w.lukaszonek@g.pl

[3]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. E-mail: wolynski@amu.edu.pl

Hence, estimation of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ will require a very large sample, which may not always be feasible. Hence, we assume $\boldsymbol{\Omega}$ to be of the form:

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{V},$$

where $\boldsymbol{\Sigma}$ is a $(p+q) \times (p+q)$ positive definite matrix and $\boldsymbol{V}$ is $T \times T$ positive definite matrix and $\boldsymbol{\Sigma} \otimes \boldsymbol{V}$ is the Kronecker product of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$. In this case the estimates of the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are positive definite with probability 1 if and only if $n > \max(p+q, T)$.

The matrix $\boldsymbol{\Sigma}$ represents the covariance between all $(p+q)$-variables on a given object and for a given time-point. Likewise, $\boldsymbol{V}$ represents the covariance between repeated measures on a given object and for a given variable. The above covariance structure is subject to an implicit assumption that for all variables the correlation structure between repeated measures remains the same, and that covariance between variables does not depend on time and remains constant for all time-points.

Classification rules in the case of multivariate repeated measures data under the assumption of multivariate normality for classes and with compound symmetric correlation structure on the matrix $\boldsymbol{V}$ were given by Roy and Khattree (2005). Next, Roy and Khattree (2008) gave the solution of this problem for the case when the correlation matrix $\boldsymbol{V}$ has the first order autoregressive [AR(1)] structure. Srivastava and Naik (2008), and McCollum (2010) describe the structure of canonical correlation and canonical variables (Hotelling 1936) based on the variables $X_1, \ldots, X_p$ and $X_{p+1}, \ldots, X_{p+q}$ observed at T time-points. Srivastava et al. (2008) found the form of maximum likelihood estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ and using these estimates gave the test of the hypothesis that the covariance matrix $\boldsymbol{\Omega}$ has the form $\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{V}$ against the alternative that the covariance matrix is not of Kronecker product structure, when $n > (p+q)T$. Krzyśko and Skorzybut (2009) and Krzyśko et al. (2011) proposed some new classification rules applicable in the case when no structures whatsoever are imposed on $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ except that they are positive definite. Deręgowski and Krzyśko (2009) constructed principal components for this model. Application of principal component analysis for the different types of genotypes of blackcurrants is described in the paper by Krzyśko et al. (2010), while the use of principal components to analyse a data set obtained from the experiments with varieties of winter rye is described in the paper Krzyśko et al. (2014).

The aim of this paper is to examine the relationship between the characteristics of higher education and the quality of life and human capital characteristics observed in 2002-2014 in each of the 16 Polish provinces. A particular model of canonical analysis, described in Srivastava and Naik (2008), will be used as a research tool.

In Section 2 we characterize our data set. In Section 3 canonical analysis for doubly multivariate data, based on results of Srivastava and Naik (2008), and McCollum (2010), is presented in the case when no structures whatsoever are imposed on $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ except that they are positive definite. In Section 4 we present the computational schemes for maximum likelihood estimates of unknown parameters.

In Section 5 the analysis results are presented.

2. Characteristics of the data set

The data used are taken from the Local Data Bank (https://bdl.stat.gov.pl). Local Data Bank is Poland's largest database containing data with respect to economy, households, innovation, public finance, society, demographics and the environment. The analysis relates to 16 Polish provinces $(n = 16)$. On the graphs, the provinces are denoted by numbers as given in Table 1.

Table 1. Designations of provinces

1	Dolnośląskie	9	Podkarpackie
2	Kujawsko-Pomorskie	10	Podlaskie
3	Lubelskie	11	Pomorskie
4	Lubuskie	12	Śląskie
5	Łódzkie	13	Świętokrzyskie
6	Małopolskie	14	Warmińsko-Mazurskie
7	Mazowieckie	15	Wielkopolskie
8	Opolskie	16	Zachodniopomorskie

The analysed data cover a period of 13 years, from 2002 to 2014 $(T = 13)$. Each province was characterized by two vectors of features:

$\boldsymbol{X}_1 = (X_1, \ldots, X_7)'$ characteristics of higher education $(p = 7)$

X_1 – The number of universities per 1 million inhabitants,
X_2 – The number of students per 1000 inhabitants,
X_3 – The number of university graduates per 1000 inhabitants,
X_4 – The number of academic teachers per 1000 inhabitants,
X_5 – The number of professors per 100 000 inhabitants,
X_6 – The number of post-graduate students per 10 000 inhabitants,
X_7 – The number of doctoral students per 10 000 inhabitants,

and

$\boldsymbol{X}_2 = (X_8, \ldots, X_{15})'$ quality of life and human capital characteristics $(q = 8)$

X_8 – Infant mortality rate per 1000 live births,
X_9 – Incidence of pulmonary tuberculosis per 100 000 inhabitants,
X_{10} – GDP per capita,
X_{11} – The Registered Unemployment Rate,
X_{12} – The percentage of inhabitants working in industry,
X_{13} – The percentage of inhabitants with university education,
X_{14} – The percentage of people learning and further training at the age of 25-69,
X_{15} – Employed in Research & Development per 1000 inhabitants.

3. Canonical analysis for doubly multivariate data

Let $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p,\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q}) = (\boldsymbol{X}_1,\boldsymbol{X}_2)$, where $\boldsymbol{X}_1 = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p)$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q})$ and let

$$\begin{aligned}\mathrm{Var}(\mathrm{vec}(\boldsymbol{X})) &= \mathrm{Var}\begin{pmatrix}\mathrm{vec}(\boldsymbol{X}_1)\\ \mathrm{vec}(\boldsymbol{X}_2)\end{pmatrix} = \boldsymbol{\Omega} = \boldsymbol{\Sigma}\otimes\boldsymbol{V}\\ &= \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\otimes\boldsymbol{V} = \begin{bmatrix}\boldsymbol{\Sigma}_{11}\otimes\boldsymbol{V} & \boldsymbol{\Sigma}_{12}\otimes\boldsymbol{V}\\ \boldsymbol{\Sigma}_{21}\otimes\boldsymbol{V} & \boldsymbol{\Sigma}_{22}\otimes\boldsymbol{V}\end{bmatrix},\end{aligned}$$

where $\boldsymbol{\Sigma}_{11}$ is $p\times p$ matrix.

In this model, the basis of canonical analysis are the eigenvalues and the eigenvectors of the matrices $\boldsymbol{A} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\otimes\boldsymbol{I}_T$ and $\boldsymbol{B} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\otimes\boldsymbol{I}_T$ (Srivastava and Naik, 2008).

One of the main reasons for the use of the Kronecker product is a simple relationship between the eigenvalues and the eigenvectors $\boldsymbol{A}$ and $\boldsymbol{I}_T$ and $\boldsymbol{A}\otimes\boldsymbol{I}_T$ (see, e.g., Lancaster and Tismenetsky (1985), p. 412; Ortega (1987), p. 237). If $\alpha_1,\ldots,\alpha_p$ are the eigenvalues of $\boldsymbol{A}$ and $\beta_1,\ldots,\beta_T$ are the eigenvalues of $\boldsymbol{I}_T$, then eigenvalues of $\boldsymbol{A}\otimes\boldsymbol{I}_T$ are the pT numbers $\alpha_r\beta_s$, $r = 1,\ldots,p$, $s = 1,\ldots,T$. If $\boldsymbol{u}$ is an eigenvector of $\boldsymbol{A}$ corresponding to the eigenvalues α, and $\boldsymbol{w}$ is an eigenvector of $\boldsymbol{I}_T$ corresponding to the eigenvalues β, then an eigenvector $\boldsymbol{\gamma}$ of $\boldsymbol{A}\otimes\boldsymbol{I}_T$ associated with $\alpha\beta$ is $\boldsymbol{\gamma} = \boldsymbol{u}\otimes\boldsymbol{w} = (\boldsymbol{u}_1\boldsymbol{w}',\boldsymbol{u}_2\boldsymbol{w}',\ldots,\boldsymbol{u}_p\boldsymbol{w}')'$.

Eigenvalues of matrix $\boldsymbol{I}_T$ are identical and equal to one. The corresponding eigenvectors are of the form:

$$\boldsymbol{w}_j = (0,\ldots,0,1,0,\ldots,0)',$$

where the only nonzero element is 1 in the $(T+1-j)$th position, $j = 1,2,\ldots,T$. Hence, the nonzero eigenvalues ρ_i^2 of matrix $\boldsymbol{A}\otimes\boldsymbol{I}_T$ are equal

$$\underbrace{\alpha_1,\ldots,\alpha_1}_{T\text{ times}},\underbrace{\alpha_2,\ldots,\alpha_2}_{T\text{ times}},\ldots,\underbrace{\alpha_k,\ldots,\alpha_k}_{T\text{ times}},$$

where $k = \mathrm{rank}(\boldsymbol{\Sigma}_{12}) \le \min(p,q)$.

The corresponding eigenvectors are of the form:

$$\boldsymbol{\gamma}_{ij} = \boldsymbol{u}_i\otimes\boldsymbol{w}_j,\ i = 1,\ldots,k,\ j = 1,\ldots,T.$$

The lth element of the vector $\boldsymbol{\gamma}_{ij}$ has the form:

$$\gamma_{ij,l} = \begin{cases} u_{il}, & \text{if } l = i(T+1-j),\\ 0, & \text{if } l \ne i(T+1-j),\end{cases}$$

$i = 1,\ldots,k$, $j = 1,\ldots,T$, $l = 1,\ldots,p$.

Nonzero values $\rho_1 \ge \rho_2 \ge \cdots \ge \rho_{kT}$, which are positive square roots of $\rho_1^2 \ge \rho_2^2 \ge$

$\cdots \geq \rho_{kT}^2$, are called the canonical correlations. The variables

$$U_{ij} = \boldsymbol{\gamma}_{ij}' \mathrm{vec}(\boldsymbol{X}_1),\ i = 1,\ldots,k,\ j = 1,\ldots,T,$$

are called the canonical variables in the space $\boldsymbol{X}_1$.

Similarly, if $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_{kT}^2$ are non-zero eigenvalues of the matrix $\boldsymbol{B}$ and $\boldsymbol{\psi}_{ij}$, $i = 1,\ldots,k,\ j = 1,\ldots,T$ are the corresponding eigenvectors, then the variables

$$V_{ij} = \boldsymbol{\psi}_{ij}' \mathrm{vec}(\boldsymbol{X}_2),\ i = 1,\ldots,k,\ j = 1,\ldots,T,$$

are called the canonical variables in the space $\boldsymbol{X}_2$.

In practice, the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are replaced by their estimators.

4. Maximum likelihood estimators of $\boldsymbol{\mu}$, Σ, and $\boldsymbol{V}$

We will use a different estimation method from the method used in Srivastava and Naik (2008), namely we will select the estimators obtained in Srivastava et al. (2008). For estimating the unknown parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$ we require n observations on the $T \times (p+q)$-matrix $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p,\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q})$. These n observation matrices will be denoted by

$$\boldsymbol{X}_j = (\boldsymbol{x}_{1j},\ldots,\boldsymbol{x}_{pj},\boldsymbol{x}_{p+1,j},\ldots,\boldsymbol{x}_{p+q,j}),\ j = 1,\ldots,n.$$

Let

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j,\ \boldsymbol{X}_{j,c} = \boldsymbol{X}_j - \bar{\boldsymbol{X}},\ j = 1,\ldots,n.$$

We consider a model, denoted by I, described as follows: all observations $\boldsymbol{X}_j$ are independent and $\mathrm{vec}(\boldsymbol{X}_j) \sim N_{(p+q)T}(\boldsymbol{\mu},\boldsymbol{\Sigma}\otimes\boldsymbol{V})$, where $\boldsymbol{\Sigma}$ is $(p+q)\times(p+q)$ positive definite covariance matrix and $\boldsymbol{V}$ is $T \times T$ positive definite covariance matrix, $j = 1,\ldots,n$, $n > \max(p+q,T)$. The maximum likelihood estimation equations are of the form (Srivastava et al. 2008):

$$\hat{\boldsymbol{\mu}} = \mathrm{vec}(\bar{\boldsymbol{X}}) \tag{1}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{nT}\sum_{j=1}^{n} \boldsymbol{X}_{j,c}' \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}_{j,c}, \tag{2}$$

$$\hat{\boldsymbol{V}} = \frac{1}{n(p+q)}\sum_{j=1}^{n} \boldsymbol{X}_{j,c} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}_{j,c}'. \tag{3}$$

In this case no explicit maximum likelihood estimates (MLEs) of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are available. The MLEs of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are obtained by solving simultaneously and iteratively the equations (2) and (3). This is the so-called "flip-flop" algorithm.

In this model (model I), if $n > \max(p+q, T)$ then the maximum likelihood estimation equations given by (2) and (3) will always converge to the unique maximum (Srivastava et al. 2008).

The following iterative steps are suggested to obtain the maximum likelihood estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$.

Step 1. Get the initial covariance matrix $\boldsymbol{V}$ in the form

$$\hat{\boldsymbol{V}} = \frac{1}{n(p+q)} \sum_{j=1}^{n} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'. \tag{4}$$

Step 2. On the basis of the initial covariance matrix $\hat{\boldsymbol{V}}$ compute the matrix $\hat{\boldsymbol{\Sigma}}$ given by (2).

Step 3. Compute the matrix $\hat{\boldsymbol{V}}$ from equation (3) using the matrix $\hat{\boldsymbol{\Sigma}}$ obtained in Step 2.

Step 4. Repeat Steps 2 and 3 until convergence is attained. We have selected the following stopping rule. Compute two matrices: (a) a matrix of difference between two successive solutions of $\hat{\boldsymbol{\Sigma}}$, and (b) a matrix of difference between two successive solutions of $\hat{\boldsymbol{V}}$. Continue the iteration procedure until the maxima of the absolute values of the elements of the matrices in (a) and (b) are smaller than the prespecified quantities.

As noted in the literature (see, e.g., Galecki (1994); Naik and Rao (2001)), since

$$(c\boldsymbol{\Sigma}) \otimes (c^{-1}\boldsymbol{V}) = \boldsymbol{\Sigma} \otimes \boldsymbol{V},$$

all the parameters of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are not defined uniquely. But in our case, estimation of the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ is not an aim in itself.

The resulting estimates are used to construction the canonical variables. Canonical variables considered in this paper are functions of $\hat{\boldsymbol{\Sigma}} \otimes \hat{\boldsymbol{V}}$, instead of $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{V}}$ separately, and hence only $\hat{\boldsymbol{\Sigma}} \otimes \hat{\boldsymbol{V}}$ needs to be unique under the model (I) with $\boldsymbol{\Sigma} > 0$ and $\boldsymbol{V} > 0$.

5. Analysis results

We are interested in the relationship between the vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. We build estimators of the matrices $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ and $\boldsymbol{\Sigma}_{12}$, and we then find the non-zero eigenvalues $\hat{\alpha}_k$ and corresponding eigenvectors $\hat{\boldsymbol{u}}_k$ of the matrix $\hat{\boldsymbol{A}} = \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21}$, and the non-zero eigenvalues $\hat{\alpha}_k$ and corresponding eigenvectors $\hat{\boldsymbol{v}}_k$ of the matrix $\hat{\boldsymbol{B}} = \hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}$, where $\hat{\boldsymbol{\Sigma}}_{21} = \hat{\boldsymbol{\Sigma}}_{12}'$, $k = 1, \ldots, 7 = \min(p, q)$.

The non-zero eigenvalues $\hat{\alpha}_k$ are shown in Figure 1.

The characteristics of higher education and quality of life and human capital characteristics are moderately correlated with the canonical correlation coefficient $\hat{\rho}_1 = 0.38$.

Eigenvectors $\hat{\boldsymbol{u}}_k$ of the matrix $\hat{\boldsymbol{A}}$ are shown in Table 2, and eigenvectors $\hat{\boldsymbol{v}}_k$ of the matrix $\hat{\boldsymbol{B}}$ are shown in Table 3.

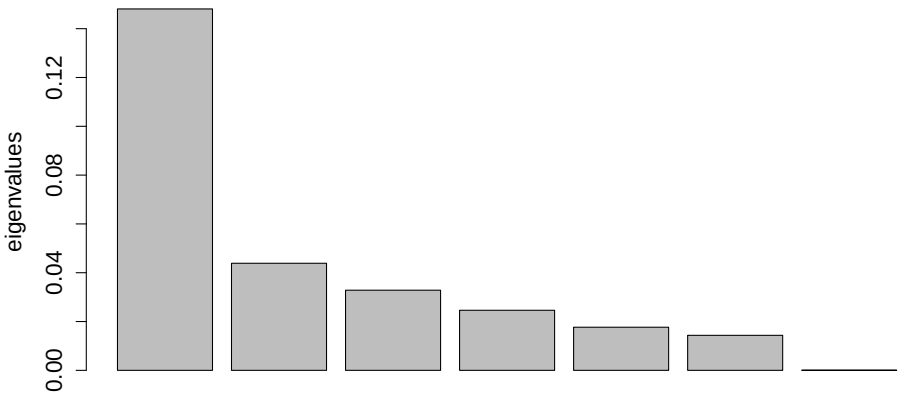


Figure 1: The non-zero eigenvalues $\hat{\alpha}_k$

Table 2. Eigenvectors $\hat{\boldsymbol{u}}_k$ of the matrix $\hat{\boldsymbol{A}}$

	$\hat{\boldsymbol{u}}_1$	$\hat{\boldsymbol{u}}_2$	$\hat{\boldsymbol{u}}_3$	$\hat{\boldsymbol{u}}_4$	$\hat{\boldsymbol{u}}_5$	$\hat{\boldsymbol{u}}_6$	$\hat{\boldsymbol{u}}_7$
1	-0.0703	0.0564	-0.0274	0.0037	0.1379	-0.0045	-0.0086
2	-0.0159	0.0451	0.0054	-0.0045	-0.0140	-0.0130	0.0055
3	-0.0015	-0.1640	0.0371	0.0010	0.0312	-0.0068	0.0044
4	0.1027	0.1848	0.3343	0.4695	-0.4270	-0.0586	-0.1198
5	-0.6819	-0.9531	-0.7931	-0.8710	0.8034	0.9929	0.9859
6	-0.0857	-0.1140	-0.0884	0.0079	-0.0762	-0.0032	-0.0082
7	-0.7155	0.1108	0.4992	-0.1446	-0.3825	0.1021	-0.1163

Table 3. Eigenvectors $\hat{\boldsymbol{v}}_k$ of the matrix $\hat{\boldsymbol{B}}$

	$\hat{\boldsymbol{v}}_1$	$\hat{\boldsymbol{v}}_2$	$\hat{\boldsymbol{v}}_3$	$\hat{\boldsymbol{v}}_4$	$\hat{\boldsymbol{v}}_5$	$\hat{\boldsymbol{v}}_6$	$\hat{\boldsymbol{v}}_7$
1	-0.5126	-0.2076	0.1924	0.1080	0.2955	0.3694	0.2669
2	-0.0164	0.0178	0.0785	-0.0231	0.1125	-0.0218	0.1561
3	0.0008	0.0002	0.0004	-0.0004	0.0001	0.0002	0.0001
4	-0.4536	0.4884	0.3444	-0.5500	-0.4824	0.3384	-0.3656
5	0.0428	-0.2987	-0.1727	-0.2816	-0.2534	0.0425	0.3910
6	-0.0445	-0.7583	0.2534	-0.2278	0.1240	-0.1116	-0.7477
7	0.5255	0.2089	-0.7506	0.2628	0.4246	0.8400	-0.1666
8	0.5013	-0.1010	0.4252	0.6964	-0.6383	0.1696	0.1762

Eigenvalues of matrix $\boldsymbol{I}_{13}$ are identical and equal to one. The corresponding eigenvectors are of the form $\boldsymbol{w}_j = (0,\ldots,1,\ldots,0)'$, where the only non-zero element is the number 1 in the $(14-j)$th position, $j = 1,\ldots,13$.

Hence, the eigenvalues of matrix $\hat{\boldsymbol{A}} \otimes \boldsymbol{I}_{13}$ are equal

$$\underbrace{\hat{\rho}_1^2,\ldots,\hat{\rho}_1^2}_{13 \text{ times}},\underbrace{\hat{\rho}_2^2,\ldots,\hat{\rho}_2^2}_{13 \text{ times}},\ldots,\underbrace{\hat{\rho}_7^2,\ldots,\hat{\rho}_7^2}_{13 \text{ times}}.$$

The corresponding eigenvectors are of the form:

$$\hat{\boldsymbol{\gamma}}_{ij} = \hat{\boldsymbol{u}}_i \otimes \boldsymbol{w}_j,\ i = 1,\ldots,7,\ j = 1,\ldots,13.$$

The lth element of the vector $\hat{\boldsymbol{\gamma}}_{ij}$ has the form:

$$\hat{\gamma}_{ij,l} = \begin{cases} \hat{u}_{il}, & \text{if } l = i(14-j), \\ 0, & \text{if } l \neq i(14-j), \end{cases}$$

$i = 1,\ldots,7,\ j = 1,\ldots,13,\ l = 1,\ldots,7.$

Replacing the vectors $\hat{\boldsymbol{u}}_i$ by $\hat{\boldsymbol{v}}_i$ we obtain similar results for the matrix $\hat{\boldsymbol{B}} \otimes \boldsymbol{I}_{13}$.

If we choose the system $(U_{1,13}, V_{1,13})$, then the location of the various provinces in this system include data from 2002 (see Fig. 2). If, however, we choose the system (U_{11}, V_{11}), then the location of the various provinces in this system includes data from 2014 (see Fig. 3). It results from the vectors $\boldsymbol{\gamma}_{1j}$ and $\boldsymbol{\psi}_{1j}$, $j = 1,\ldots,T$.

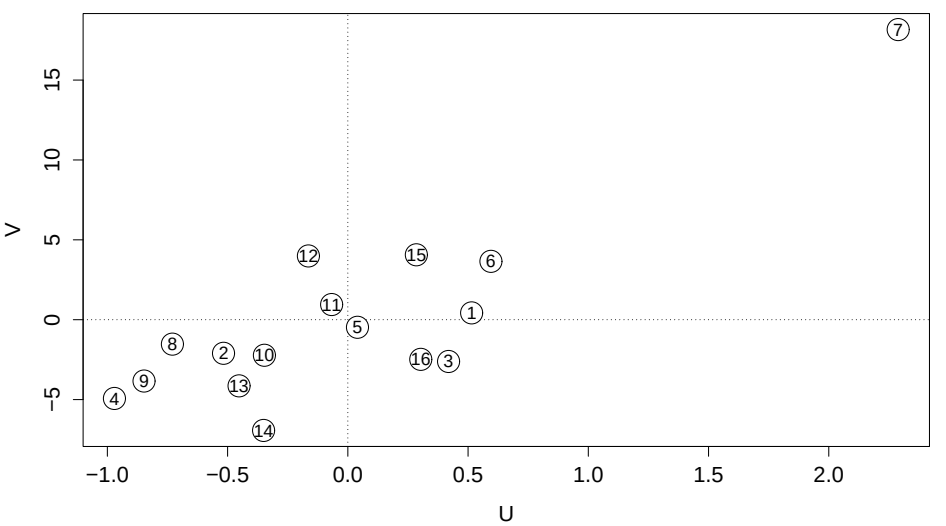


Figure 2: The location of the various provinces in 2002

For example

$$\boldsymbol{\gamma}_{11} = (0,\ldots,u_{11},0,\ldots,u_{12},\ldots,0,\ldots,u_{17})',$$

and

$$\boldsymbol{\gamma}_{1,13} = (u_{11},\ldots,0,\ldots,u_{12},\ldots,0,\ldots,u_{17},\ldots,0)'.$$

Note that all of these systems of canonical variables correspond to the same value of the canonical correlation coefficient $\hat{\rho}_1 = 0.38$.

In Fig. 2, on the one hand, one can see provinces with bad (low) values of characteristics of higher education and bad (low) values of quality of life and human capital characteristics such as Lubuskie (4), Podkarpackie (9) and Opolskie (8), and on the other hand, one can see provinces with high values of characteristics of higher education and good (high) values of quality of life and human capital characteristics such as Mazowieckie (7), Małopolskie (6), and Dolnośląskie (1). The absolute leader is Mazowieckie (7) province. Comparing 2014 to 2002 in Fig. 3, a change in the location of some provinces can be observed. For example, Opolskie (8) and Pomorskie (11) provinces improved their position, but the position of Zachodniopomorskie (16) and Warmińsko-Mazurskie (14) deteriorated.

During the calculations we used R (R Core Team (2017)) software. The R source code is available on request at the third co-author.

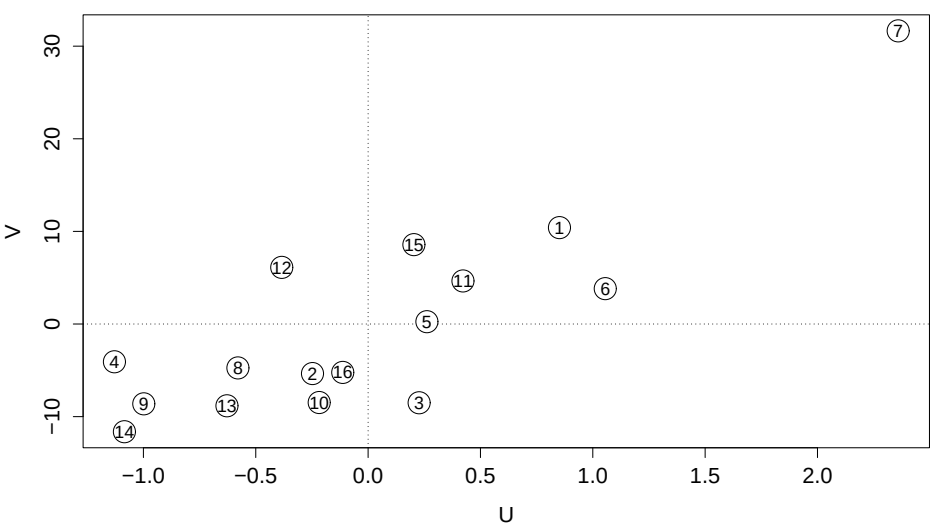


Figure 3: The location of the various provinces in 2014

REFERENCES

DERĘGOWSKI, K., KRZYŚKO, M., (2009). Principal component analysis in the case of multivariate repeated measures data, Biometrical Letters, 46 (2), pp. 163–172.

GALECKI, A. T., (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis, Communications in Statistics – Theory and Methods, 23, pp. 3105-3119.

GIRI, N. C., (1996). Multivariate Statistical Analysis, Marcel Dekker, New York.

HOTELLING, H., (1936). Relations between two sets of variates, Biometrika, 28, pp. 321–377.

KRZYŚKO, M., SKORZYBUT, M., (2009). Discriminant analysis of multivariate repeated measures data with a Kronecker product structured covariance matrices, Statistical papers, 50, 817-835.

KRZYŚKO, M., MĄDRY, W., PLUTA, S., SKORZYBUT, M., WOŁYŃSKI, W., (2010). Analysis of multivariate repeated measures data, Colloquium Biometricum, 40, pp. 117–133.

KRZYŚKO, M., SKORZYBUT, M., WOŁYŃSKI, W., (2011). Classifiers for doubly multivariate data, Discussiones Mathematicae. Probability and Statistics, 31, pp. 5–27.

KRZYŚKO, M., ŚMIAŁOWSKI, T., WOŁYŃSKI, W., (2014). Analysis of multivariate repeated measures data using a MANOVA model and principal components, Biometrical Letters, 51 (2), pp. 103–124.

LANCASTER, P., TISMENETSKY, M., (1985). The Theory of Matrices, Second Edition: With Applications. Academic Press, Orlando.

MCCOLLUM, R., (2010). Canonical correlation analysis for longitudinal data. Ph.D. thesis, Old Dominion University.

NAIK, D. N., RAO, S., (2001). Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix, J. Appl. Statist., 28, pp. 91–105.

ORTEGA, J. M., (1987). Matrix Theory: A Second Course. Plenum Press, New York.

R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

ROY, A., KHATTREE, R., (2005). On discrimination and classification with multivariate repeated measures data, Journal of Statistical Planning and Inference, 134, pp. 462–485.

ROY, A., KHATTREE, R., (2008). Classification rules for repeated measures data from biomedical research. In: Khattree, R., Naik, D. N. (eds) Computational methods in biomedical research, Chapman and Hall/CRC, pp. 323–370.

SRIVASTAVA, J., Naik, D. N., (2008). Canonical correlation analysis of longitudinal data, Denver JSM 2008 Proceedings, Biometrics Section, pp. 563–568.

SRIVASTAVA, M.S., VON ROSEN, T., VON ROSEN, D., (2008). Models with a Kronecker product covariance structure: estimation and testing, Math. Methods Stat., 17 (4), pp. 357–370.

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 87–117, DOI 10.21307/stattrans-2018-006

SEARCHING FOR CAUSES OF NECROTISING ENTEROCOLITIS. AN APPLICATION OF PROPENSITY MATCHING

Nicholas T. Longford[1]

ABSTRACT

Necrotising enterocolitis (NEC) is a disease of the gastrointestinal tract afflicting preterm-born infants in the first few weeks of their lives. We estimate the effect of changing the feeding regimen of infants in their first 14 postnatal days by analysing the data from the UK National Neonatal Research Database. We avoid some problems with drawing causal inferences from observational data by reducing the analysis to the infants who spent the first 14 postnatal days (or longer) in neonatal care and for whom NEC was not suspected in this period. This reduction enables us to use summaries of the feeding regimen in this period as background variables in a potential outcomes framework. Large size of the cohort is a distinct advantage of our study. Its results inform the design of a randomised clinical trial for preventing NEC, and the choice of its active treatment(s) in particular.

Key words: causal analysis, National Neonatal Research Database, necrotising enterocolitis, potential outcomes framework, preterm birth, propensity matching.

1. Introduction

Necrotising enterocolitis (NEC) is a gastrointestinal disease that afflicts mainly preterm-born infants with low birthweight (Neu and Walker, 2011; Patel and Shah, 2012). The aetiology of the disease is poorly understood because preterm births are infrequent (about 10% of all births, World Health Organization, 2011) and the disease is rare even in the highest-risk subpopulation of extreme preterm-born infants (incidence up to 10%). Clinical trials on newborns are difficult to design, organise and have them approved because they involve high ethical costs and standards. Difficulties in recruitment are also frequently encountered. A variety of concerns has to be addressed in the treatment of preterm-born neonates in the first few weeks of their lives, and involving them in a clinical trial is in most cases an unwelcome distraction to both parents and the clinical staff providing neonatal care.

The design of a randomised clinical trial (RCT) in such a vulnerable population has to draw on the information available in all the relevant sources, so as to maximise the chances of an unequivocal result that would be easy to interpret and implement in future practice, while requiring as small a sample as possible

[1] Imperial College London. Great Britain. E-mail: n.longford@imperial.ac.uk

and being least invasive or disruptive in the normal course of providing (intensive) care. Departures from the study protocol are likely as medical staff and parents constantly reassess the appropriateness of the treatment prescribed by the study protocol and, unhesitatingly abandon its strictures if the protocol appears to be in conflict with the (perceived) wellbeing of the infant.

We study the effects of early feeding exposures on NEC using the data from the UK National Neonatal Research Database (NNRD) in 2012 and 2013, originating from 162 neonatal units organised in 23 networks. The principal difficulty in such an analysis is the observational nature of the data, generated without applying any experimental control, and collected not for the purpose of our analysis. The treatment variables we consider are derived from the feeding regimen in the first few postnatal days of the infants detained in neonatal care units. The regimen, prescribed for an infant by a neonatal consultant or dietician, is informed by frequent observations of the infant, and can in no way be regarded as being assigned at random. In contrast, the regimen would be randomised in a RCT.

The importance of the feeding regimen in the first few weeks of an infant's life is beyond any contention. Early feeding exposures are likely to alter the microbiome (the microbial composition) of an infant's gut and influence the susceptibility to NEC (Neu, 2015). The feeding regimen is a key modifiable risk factor for NEC and it is paramount that feeding guidelines be based on relevant evidence and be congruous with contemporary clinical practice. For other elements of care, such as hygiene, ambient temperature and lighting, there are generally accepted standards. International recommendations (Arslanoglu *et al.*, 2013; American Academy of Pediatry, 2012; World Health Organization, 2011) endorse the use of human donor milk (HDM) in preterm-born infants when maternal breast milk (MBM) is not available. However, these recommendations are based on weak evidence, predominantly from trials conducted prior to the routine use of multi-component bovine milk-derived human milk fortifiers, which are now accepted as standard clinical practice (Quigley and McGuire, 2014). Only one of the six trials included in the meta-analysis of Schanler *et al.* (2005) investigated the effect of nutrient-fortified donor milk given as a supplement to MBM. Interest has been also growing in the exclusive human-milk diet which includes human rather than bovine-derived fortifier (Sullivan *et al.*, 2010).

Adequately powered RCTs would establish whether nutrient-fortified HDM is a better supplement to MBM than preterm formula milk, and whether human milk-derived or bovine-derived fortifier has lower risk of NEC. Their results could form the basis for comprehensive guidelines. Observational population-based studies are an alternative to RCT. Their advantages include readily available timely data at a relatively low cost. Their drawback is the necessity to record background variables, control for them in the analysis, and the uncertainty as to whether this list of variables is complete — whether they render the treatment assignment (feeding regimen) ignorable in the sense of Rubin (1976). In contrast, analysis of RCT is simple, but only when the study protocol is complied with fully and the studied population is well represented in the study.

Quigley and McGuire (2014) reviewed nine randomised trials for comparing the effects on preterm and low birthweight infants of HDM and formula as supplements to MBM. The total of their sample sizes was only 1070. For the inferences we seek, comparing two small proportions, we would need far greater sample sizes. NNRD in 2012 – 13 contains records of over 14 000 infants with gestational age at birth (GA) of up to 28 (completed) weeks. This sample size is reduced by two criteria, being detained in a neonatal unit for at least the first 14 postnatal days and not being under suspicion of having NEC during these 14 days, to just under 12 000 infants, from which two matched treatment groups of around 3 000 infants are formed by propensity matching.

The variables in NNRD can be classified as holding information about

- the mother (her age, previous pregnancies, smoking habit, health status, and the like);
- the birth (mode of delivery, birthweight, GA, fetus order, and the like); and
- the daily feeding regimen, indicating the following types of nutrition (or their absence): parenteral nutrition, MBM, HDM, formulas, fortifiers, and no feeding by mouth.

Further, it includes a dichotomous variable that indicates whether NEC is suspected. This variable is also recorded daily, and the assessment is in general not made by the same consultant in a sequence of a few days. The assessment is not straightforward and differences in opinion and judgement of consultants are likely, although they cannot be observed because only one assessment is made and recorded every day. Similar databases are maintained in other countries, but collection of daily data is a unique feature of NNRD.

The feeding regimen is chosen in response to several concerns, of which NEC is not always the foremost. The amount of food intake is set, rising gradually from 40ml in the first few days of the infant's life to 150ml per kilogram of body mass (weight). This daily amount may be composed of several types of nutrition. For instance, if the volume of MBM is not sufficient it may be supplemented by HDM or formulas.

Suspicion of NEC naturally influences the feeding regimen which, together with medication, is the principal set of options for responding to the concern. At the same time, any meaningful analysis of treatments (possible interventions) for NEC has to use covariates derived from the feeding regimen. The purpose of such an analysis would be to propose a change or adaptation of the regimen that would reduce the risk of developing NEC. Standard methods for relating the outcome y (incidence of NEC) to the values of the explanatory variables $\mathbf{x}$ would fail if the values of $\mathbf{x}$ were set in response to the anticipated values of y. Autonomy in how the values of $\mathbf{x}$ are set is a key assumption of most models and methods for relating y, or its conditional expectation $f(\mathbf{x}) = \mathrm{E}(y|\mathbf{x})$, to $\mathbf{x}$. Adjusting for the lack of autonomy is feasible only when the process of setting $\mathbf{x}$, described by $\mathrm{E}(y|\mathbf{x})$, is known or can be inferred with some confidence. Such information is scant because we are not privy to the

decision process involved — the clinical judgement and balancing of a whole array of concerns about the survival and wellbeing of the infant.

The suspicion of NEC is not an accurate indicator of developing NEC in the future. Many more infants are suspected to have NEC (in the first 14 days) than the actual number of cases identified later. Instances of isolated days on which NEC is suspected are an indication of either disagreement or inconsistency in the assessments. At the same time, many future cases of NEC are not suspected until its unambiguous symptoms are observed.

There is no consensus in the literature on the mechanisms by which feeding influences NEC, although some limited insights are widely shared. On the one hand, nutrients are essential for the preterm born infant. Feeding, and breastfeeding in particular, is the obvious way to provide them. On the other hand, processing the feed introduces stress on the immature gastrointestinal tract. The balance of these two considerations remains a fine art in neonatal care. Failure to maintain it, and match it to the state of the infant, is a likely risk factor.

Just as different consultants may disagree with one another about NEC, alternating consultants may introduce more changes in the daily feeding regimen than if one person were in control. Even though the options for feeding the infants are limited (MBM when available, HDM, formulas, and their combinations), variation in the patterns of feeding in the 23 neonatal networks in England is so wide that the policies followed are unlikely to be equally effective.

Even among the extreme preterm-born infants, with GA of 27 weeks or earlier, NEC is a rare condition, but the mortality among the affected is high, and the time between the onset of symptoms of the disease and death is often too short for an effective surgical intervention. The diagnosis is not always clinical. An infant may have the disease without being diagnosed or detected, and may be cured while being treated for a different indication. Thus, the assessment of the quality of the 'suspicion' is itself problematic.

The database from which we extract data for the analysis is described in Section 2. In Section 3, we describe the potential outcomes framework (Holland, 1986; Imbens and Rubin, 2015), also known as Rubin's causal model, and discuss its advantages over some established alternatives for the analysis of causes of NEC. The target of an analysis in this framework is the same as in a (hypothetical) clinical trial for comparing two treatments:

> How would the outcomes of a group of patients who received one treatment change on average *if* they had received the other treatment?

In the framework a subset of each treatment group is selected so that the subsets are tightly matched (balanced) on all the background variables, as they would be in a study with the treatments allocated at random. The details of how these matched groups are selected are given in Section 4. The outcomes of these subsets are then compared straightforwardly, by a method that would be applied in a randomised study. This general approach can be described as *post-observational* design. Of course, background variables that are not observed remain as potential

confounders. Our only protection against this is that the list of background variables recorded in NNRD is quite comprehensive.

Section 5 applies the potential outcomes framework to the data from NNRD. Section 6 discusses the results and how they might inform a randomised clinical trial, by the choice of the alternative treatments in it, and by other elements of the design.

The feeding regimen in the first two postnatal weeks is an important source of background variables, because the feed taken is the first suspect in any gastrointestinal disease. However, the regimen is disqualified as background by the suspicion of NEC because it causes the neonatal consultant to alter the feeding regimen. After all, that is an important means of treating the infant. We resolve this problem by excluding from the analysis all the infants who were suspected to have NEC or to develop it in their first 14 postnatal days. For the retained infants, the feeding regimen is suitable for defining background variables. The choice of 14 days is a compromise. On the one hand, we prefer to have more extensive background (more days); on the other, the number of infants who fall under suspicion of having NEC increases as more days are considered, and then we would lose more cases.

The outcome variable is defined as a positive diagnosis of NEC, made either at a surgery (laparotomy) or determined as the cause of death (after postnatal day 14). An established alternative, called the Vermont-Oxford Network criterion, based on radiological and clinical observations of the infant, is derived from the Bell's staging criteria (Bell *et al.*, 1978; Kliegman and Walsh, 1987), but these signs are not recorded in the database. The criterion defines a dichotomous variable that is in general more liberal than our definition. Battersby *et al.* (2017a) present a proposal based on essentially the same observations as the Vermont-Oxford Network criterion but also incorporating GA.

The results of this study are presented and their implications discussed by Battersby *et al.* (2017b) for a clinical (medical) audience. This paper focuses on the statistical and computational aspects. The method applied is not new, but novel is its application in neonatology.

From the original population of 14 666 infants we excluded 353 infants whose records were not released to us for logistic reasons and 88 infants with empty records. From the remaining 14 225 infants, which include 441 cases of NEC, we excluded 1402 infants (9.9%) who were released from care in a neonatal unit (or died) prior to their 14th postnatal day. In the remainder (12 823 infants) there are 278 cases of NEC; of these, 58 fell under suspicion on at least one of the 14 days. Of the non-cases, 826 infants were suspected on at least one day. Thus, the rate of NEC among all the infants we consider is (278/12 823 =) 2.2%, and among those retained for the analysis it is (220/11 939 =) 1.8%. The feeding regimen can be regarded as background for these infants.

For orientation, Table 1 displays the numbers of cases and non-cases within the groups defined by GA in completed weeks. It shows that the rate of NEC is higher in extreme-preterm born infants (GA up to 26 weeks), but more cases occur among

Table 1: Cases of NEC within gestational age groups (weeks) in NNRD; 2012 – 13.

	Gestational age group (weeks)										
	22	23	24	25	26	27	28	29	30	31	All
All infants (available data)											
No NEC	11	300	643	770	1005	1303	1677	2036	2555	3484	13 784
NEC	0	34	85	75	63	53	67	20	25	19	441
% NEC	0.0	10.2	11.7	8.9	5.9	3.9	3.8	1.0	1.0	0.5	3.1
All infants in neonatal units at the age of 14 days											
No NEC	6	173	487	637	889	1171	1513	1916	2437	3316	12 545
NEC	0	22	57	49	44	31	47	13	11	4	278
% NEC	0.0	11.3	10.5	7.1	4.7	2.6	3.0	0.7	0.4	0.1	2.2
Infants in the analysis											
No NEC	6	156	439	573	795	1079	1391	1778	2312	3190	11 719
NEC	0	16	44	41	37	27	32	10	9	4	220
% NEC	0.0	9.3	9.1	6.7	4.4	2.4	2.2	0.6	0.4	0.1	1.8

the medium-preterm born (at 27 – 29 weeks), who are more numerous.

2. Data

The NNRD database contains information at two levels, related to infants, their mothers and the births (B-data), and summarising the care provided on each day when the infant concerned was detained in a neonatal unit (D-data).

The variables in the B-dataset that we consider in the analysis are summarised in Table 2. A few variables related to the outcome of the stay in a neonatal unit are added. The birthweight z-score is defined by relating the birthweight to the distribution of birthweights within the GA group defined in completed weeks (integers). Let b be the birthweight of an infant born in GA week w, m_w the mean of the birthweights within this GA group, and s_w the standard deviation of these birthweights. Then the z-score is defined as $(b - m_w)/s_w$.

A non-trivial number of values is missing for the background variables 'Previous pregnancies' and 'Smoking in pregnancy'. We define a dichotomous (0/1) variable that indicates nonresponse for them, and regard nonresponse as a separate category. For four mothers with unknown ages, the ages are recoded to 30 (years). Also, the age is truncated to be between 14 and 45 years. For a small number of mothers, the recorded age is outside this range, and we believe it is incorrect.

Table 2: Background variables defined for infants.

Variable	Values	Mean	Median	Percent	Missing values
Mother					
Mothers' age (years)	14 – 45	30.35	31	—	4
— converted from year of birth; missing values are recoded to 30					
Ethnicity	0/1	—	—	32.1	0
— 0: White; 1: other					
Previous pregnancies	0 – 15	1.30	0	—	2204
— number of previous pregnancies; missing/available defined as a dichotomous variable					
Steroids taken	0/1	—	—	10.4	0
Smoking in pregnancy	0/1/M	—	—	69.4; 17.6; 13.0	1555
— values 0: No; 1: Yes; M: missing					
Antibiotics taken by mother	0/1	—	—	25.2	0
— values 0: No; 1: Yes					
The newborn					
Month of the birth	1 – 24	12.40	12	—	0
— integer values from 1: January 2012 to 24: December 2013					
Gestational age (weeks)	22.43 – 31.86	29.16	29.71	—	0
— values converted from number of days					
Birthweight (kg)	0.14 – 3.17	1.24	1.25	—	0
Birthweight z-score	-5.10 to 4.58	0.01	0.09	—	0
— the standardised birthweight within gestational age group					
Mode of delivery	1, 2, 3	—	—	38.9; 5.9; 55.2	0
— 1: vaginal delivery; 2: elected (planned) Caesarean section; 3: emergency Caesarean section					
Gender	1/2	—	—	45.4	6
— values: 1: boy; 2: girl; missing gender recoded at random					
Fetus number	1 – 5	1.30	1	—	1
— relevant for multiple births only; 1 for single births; missing value recoded to 1					
Pyrexia	0/1	—	—	4.9	0
— values 0: No; 1: Yes					
APGAR1	1 – 10	5.68	6	—	0
— APGAR score 1 minute after birth; integers					
Outcomes					
Severe NEC	0/1	—	—	1.8	0
— whether diagnosed with NEC: 0: no; 1: yes					
Discharge outcome	1, 2, 3	—	—	92.4; 3.8; 3.8	0
— 1: release (home); 2: move to another care unit; 3: death					

2.1. Feeding regimen

The variables in D-data relate to the feeding regimen and the suspicion of NEC. We work with these variables for the first 14 postnatal days, and organise their values in strings of length 14, with a code for each day. The code is 0 for absence of the particular type or mode of feeding, and 1 for its presence. The mode includes also no nutrition provided by mouth (Nbms), coded as 0 if some food is given, and 1 if none is given. Suspicion of NEC is coded as 0 (not made) and 1 (made). Missing values are represented by the code 9.

Recorded is for each day whether the newborn received parenteral nutrition, formulas, fortifiers, MBM, HDM, and whether it was not fed by mouth at all. Formulas include any formula milk, not distinguishing among its types and varieties. The record is (multivariate) dichotomous (Y/N), with a fair number of missing values. The missing values arise not only as deficiencies in data collection; infants may leave the care unit for procedures at a different hospital, return to the care unit after a brief spell at home, and the like.

The missing data have a simple pattern. Daily records for parenteral nutrition are missing in 144 records (out of $11\,939 \times 14 = 167\,146$ records), 97 of them from a single network. Six networks have no missing items and 12 others have only one or two missing items each. The quintets of entries for MBM, HDM, Nbms, formulas and fortifiers are either all recorded or all missing. There are 5647 missing quintets (3.4%). The rates of missing values within the networks range from 1.3% to 5.8%, except for one network with 8.0%. The frequency of missing entries decreases from the first postnatal day (1419 entries, 25.1%) to the 14th (164, 2.9%). We note that the first day of life is the day of birth, and so its records are for a variable period shorter than 24 hours. For parenteral nutrition the missing entries are distributed fairly uniformly across the days ($5-15$ entries per day). The other four D-variables, fed, antibiotics, central line and Pda (*patent ductus arteriosus*) medication, are recorded with no entries missing. On a typical day, 3.6% of infants have incomplete records (one of the ten entries is not recorded), but 13.1% of the 14-digit strings have a missing entry. Over the 14 days, 29.1% of the infants have at least one missing entry.

We store the daily values for a variable (a mode of feeding) as a sequence (string) of 14 digits. An example of such a sequence is

$$00011119911111\,. \tag{1}$$

For such sequences, we define the following summaries. The *onset* of a mode of feeding is defined as the day from which on the mode is applied every single day until day 14. The onset is set to 15 if the mode is not applied on day 14.

The *offset* of a mode is defined as the day before the first day on which the mode is not applied. The value of the offset is set to zero if the mode is not applied on day 1. A mode is said to be *present* in a sequence of days if it is applied on at least one day of the sequence. A mode is said to be in a *majority* if it is applied on

more than half of the days; that is, on at least eight of the 14 days.

2.2. Imputation for missing values

Values are imputed for missing entries when there is little or no ambiguity about the missing value. A missing entry (variable-day record for an infant) is said to be *isolated* if there is a valid entry for the variable and infant on both the previous and the next day. For example, the second and third missing entries in the sequence 9001901 1191100, on days 5 and 10, are isolated because there are valid entries for days 4, 6, 9 and 11. The first missing entry, on day 1, is not isolated because it does not have a precedent. We impute for an isolated missing entry the value of its two neighbours if they are identical. That is, we change all substrings 090 to 000 and all substrings 191 to 111; substrings 091 and 190 are left unchanged. In the example, we impute value 1 for day 10, but do not impute for days 1 or 5, so the (partially) completed sequence is 9001901 1111100. Table 9 in the Appendix gives details of the imputations for isolated missing values.

Two consecutive missing entries are said to be an isolated pair if they are preceded by at least two valid entries and are also followed by two valid entries. In the example in (1), the two missing items are an isolated pair. We impute for an isolated pair the preceding and following pair of items if all four items coincide. That is, substrings 009900 and 119911 are changed to 000000 and 111111, respectively. Application to the string in (1) yields the (completed) sequence 0001111 1111111. Table 10 displays information about the sequences and imputations for them in a format similar to Table 9. There are 1008 isolated pairs, far fewer than isolated single entries, and imputations are performed for 687 pairs in 679 14-digit sequences. They involve 179 infants. The imputations are performed first for isolated missing items, and then for isolated pairs. The neighbours of an isolated pair may be altered by imputation for an isolated missing item, so the order of imputation is not innocuous.

Isolated triplets and quadruplets of missing items are defined similarly to pairs. There are 69 isolated triplets, 34 of them occurring on days $2-4$ or $3-5$, so no valid entries could be imputed for them. There are only 14 isolated quadruplets, eight of them occurring entirely within the first week. We have not imputed any values for the isolated triplets of quadruplets. Values that remain missing after the imputations are treated as a barrier to offset and onset. For the presence and majority, only positive entries are counted; missing values are treated as negative.

Table 3 lists the variables defined for presence and majority. They are supplemented by variables that indicate the presence in the first two days (48 hours) of the infant's life. Bovine products comprise formulas and fortifiers. For their presence, the presence of one kind is sufficient. Thirteen values are missing for the variables related to the first 48 hours; negatives are imputed for all of them.

The variables defined with offset and onset are summarised by their histograms in Figure 1. The diagram shows that 2066 infants in the analysis (17.3%) did not receive MBM on day 14 (their onset is on day 15). Some of them may have received

Table 3: Dichotomous variables for summarising the feeding regimen.

Variable	Percent	
	All (12 823)	No suspicion* (11 939)
Presence		
Parenteral nutrition	78.31	76.93
No feeding by mouth	84.08	82.98
Formulas	40.84	42.36
Fortifiers	10.96	11.58
Antibiotics	31.20	31.21
Central line	74.78	73.20
Pda medication	2.00	1.79
Bovine products	47.45	49.33
Majority		
Parenteral nutrition	51.20	48.69
No feeding by mouth	6.01	3.95
Formulas	17.18	18.22
Fortifiers	0.23	0.25
Breast feeding	77.15	79.79
First 48 hours		
Fed at all	43.39	44.27
No feeding by mouth	30.41	30.89
Donor breast milk	6.60	6.77

Note: * — believed to be free of NEC throughout the first 14 days.

MBM on some of the days 1 – 13, but the sequence of ones, if any, was interrupted on (at least) day 14. Exclusions due to suspicion of NEC in the first two weeks are concentrated in this category (361 infants, 40.8% of the exclusions) and they are very rare among infants with an early onset of MBM. This observation cannot be interpreted as a support for the generally adopted view that MBM is the best diet for a newborn.

The middle panel shows that parenteral nutrition is provided to many infants on the first day, but not on the second (offset of one day). No feeding by mouth (bottom panel) is applied for the first or the first two days to (2893+3712=) 6605 infants in the analysis (55.3%), but 266 infants in the analysis are not fed by mouth for the first six days (their offset is 7 days or greater).

3. Potential outcomes framework

In the potential outcomes framework, we consider a treatment variable, usually a dichotomy, such as the presence of a mode of feeding, and an outcome variable, in our case a positive diagnosis of NEC at any point after the first 14 postnatal

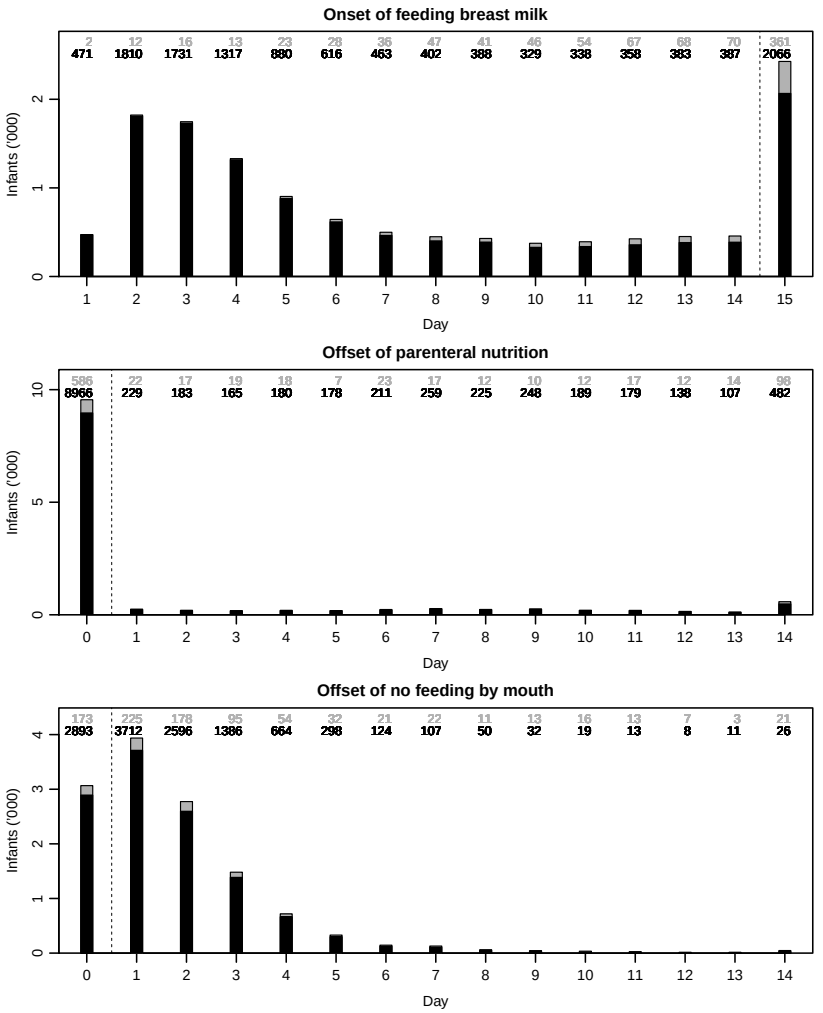


Figure 1: Distribution of the offset and onset variables. The counts of infants in the categories of each variable are given at the top, printed in black for the infants who spent the first 14 days in a neonatal unit and for whom NEC was not suspected on any of these 14 days. Grey colour is used for infants excluded from the analysis because NEC was suspected on at least one of the first 14 days.

days. We are interested in the effect of the treatment variable on the outcome. The effect is defined in a particular perspective (setting) in which the treatment can be manipulated. That is, even though an infant received one treatment, it could have received the other treatment instead. This is an essential property of any treatment we consider, because the desired result of the analysis is a proposal for altering the process of selecting a treatment for each infant. One possible result is that all the infants should be assigned to one (specified) treatment.

No infant can be subjected to more than one treatment. If one treatment is applied, then the outcome with the other treatment on the same infant cannot be established because the infant has been irrevocably altered by the first treatment. Neither can the passage of time be reversed when the treatment is related to a particular age (in days) of the infant. As an aside we note the difficulties that arise in the design and implementation of crossover trials (Jones and Kenward, 1989), which assume that each unit (subject) can be restored to the state prior to applying the first treatment.

The outcome variable has the property of increasing reward. That is, its higher values are more desirable. Equivalently, lower values may be more desirable; by multiplying a variable by a negative constant we do not alter its suitability for being an outcome variable, although we have to alter the associated values of what we regard as more desirable.

A variable is said to be *background* if its value for an infant (an observational unit) would not be altered if a treatment different from the one applied were used. Variables defined prior to considering which treatment to apply are *prima facie* background. Contention as to whether a variable is background or not can be resolved by a careful elucidation of the perspective in which manipulation of the treatment is considered.

Since NNRD is our sole data source, we are not at liberty to specify background variables for the analysis, except by transformations and recoding of the variables recorded in NNRD. In principle, all available variables that qualify as background should be included in propensity matching. Matching on a wider set of background variables makes the analysis more credible; ideally, good match should be achieved on all background variables, including those not observed, irrespective of how important they are for predicting the outcome.

Even though the outcome can be observed for at most one treatment, we consider two variables, $Y^{(\mathrm{A})}$ and $Y^{(\mathrm{B})}$, for respective treatments A and B. They are called *potential outcomes*. The qualifier *potential* signifies that only one of them can be observed. The outcome variable often considered is their mixture

$$Y^{\dagger} = (1 - I_{\mathrm{B}})\, Y^{(\mathrm{A})} + I_{\mathrm{B}}\, Y^{(\mathrm{B})},$$

where I_{B} is the indicator of having received treatment B; $I_{\mathrm{B}} = 1$ if treatment B was received and $I_{\mathrm{B}} = 0$ otherwise. A drawback of the observed outcome $Y^{\dagger}$ is that it can be meaningfully thought of only in connection with (or, conditionally on) the treatment applied. It mixes, and therefore confuses, the effect of a treatment with

the treatment assignment (selection) process. Any comparison of the values of $Y^{\dagger}$ for the group of units that received treatment A and the group that received treatment B is problematic if the two groups are not equivalent in their backgrounds — if the comparison is not of *like with like*. Otherwise the background remains a plausible explanation (a confounder) for the difference in the rates of NEC between the two treatment groups. The purpose of matching is to reduce this plausibility; a perfectly conducted RCT eliminates it altogether.

The unit-level effect of treatment B over treatment A on outcome variable Y is defined as the difference

$$\Delta Y_u = Y_{\mathrm{B}u} - Y_{\mathrm{A}u};$$

u denotes the unit. It is a variable defined in $\mathscr{V}$, the set of all units. Its size is denoted by $N_{\mathscr{V}}$. Instead of the difference another contrast can be applied, such as the ratio (for variables with positive values), or the contrast can be replaced by a comparison, which has values 'better', 'worse' and 'same'. The average effect for a set of units $\mathscr{U}$ is defined as the average of the unit-level effects for the units in $\mathscr{U}$; that is,

$$\overline{\Delta}Y = \frac{1}{N_{\mathscr{U}}} \sum_{u \in \mathscr{U}} \Delta Y_u \,.$$

The set may comprise all units that were exposed to treatment B (or A), or a specific population, such as all preterm-born infants born in a particular set of neonatal units (in a country) in a given period of time. For a comparison, the treatment effect in a population is summarised by the composition (percentages) of winners (B better than A), losers (A better than B) and ties (A and B having the same effect). An important strength of this framework is that no assumption is made about the distribution of the effect (the pattern of its values). A constant effect, assumed in some (linear) models for the observed outcomes $Y^{\dagger}$, is in general a far too restrictive assumption.

The fundamental difficulty in estimating an average treatment effect is that the contrast (or comparison) cannot be observed for any unit. A solution to this problem can be motivated by regarding it as involving missing values — values for the unrealised unit-treatment pairs. This suggests imputation for the missing values, and *multiple* imputation (Rubin, 2002; Carpenter and Kenward, 2013) to reflect the uncertainty about the completion of the dataset. A dataset completed by imputation comprises a pair $(Y_u^{(\mathrm{A})}, Y_u^{(\mathrm{B})})$ for each unit $u \in \mathscr{U}$. The dataset is analysed by evaluating the contrast of the within-treatment means. If the target is the average effect for the infants included in the study, $\mathscr{U} \in \mathscr{V}$, then the completed data analysis (CDA) involves no sampling variation, because a hypothetical replication of the study would involve the same set of units, with the same (pairs) of values of the potential outcomes. The combination of imputation (completion) and CDA involves variation (uncertainty) only owing to the imputation process, and this is captured by the variation among the results based on replications of the imputation process. This highlights the need for multiple imputation.

Inference for a (super-) population involves another element of uncertainty, due

to representing a population (incompletely) by a sample. In our context, the population we study is enumerated, admittedly with some compromises that make the analysis feasible: excluding all infants who were suspected of having NEC in the first two weeks of their lives and all those who were released from care on day 14 or earlier.

We pose the following question. Suppose all the infants who received treatment A would have received treatment B instead. How much better would the outcomes be? In our setting, the outcome is 'contracting severe NEC', a dichotomy, and the desire is for this variable to be negative. Thus, we ask how many instances of the disease would be avoided if all infants received treatment B.

An important assumption about the treatment assignment is that the units (infants and their families) do not 'interfere' with one another. That is, the treatment received by one infant has no impact on the outcome of another infant in the study. In general, a set of potential outcomes is defined for each treatment assignment, and there are $2^{N_{\mathscr{U}}}$ such assignments. The assumption of no interference amounts to the reduction of these $2^{N_{\mathscr{U}}}$ sets to just two, influenced for each unit solely by the treatment received. This assumption is known by the acronym SUTVA — stable unit-treatment value assignment (Rubin, 1978). We assume that it holds, even though it is obviously violated for multiple births, and for mothers who meet and exchange their experiences and act upon them.

4. Propensity analysis and matching

We adhere to the general standard of comparing two groups only when they are matched on the set of available background variables, that is, when the comparison is of like with like. From the two treatment groups, A and B, we select subsets of units of equal size so that the distributions of their backgrounds are close to being identical. This is done by forming a set of matched pairs; each pair comprises an infant from treatment group A and one from B. In the analysis, we consider several pairs of groups (A, B).

For given treatment groups A and B, infants are paired by matching on their fitted propensities. Propensity of a treatment is defined as the probability of being assigned the treatment, expressed as a function of the background variables. Its central role for matching was identified by Rosenbaum and Rubin (1983). Using fitted (estimated) propensities is justified by Rubin and Thomas (1996).

Thus, we fit a model for the treatment (as the binary outcome) to the background variables. This propensity model is merely a vehicle for arranging a close agreement of the distributions of the background variables in the two treatment groups. Such a match can be motivated as selecting from the observational units a subset, as large as possible, which has the appearance of a dataset that might have arisen in an (hypothetical) randomised trial, and can be analysed as such. The comparison of the matched pairs (the completed dataset) is straightforward — evaluating the contrast of the within-group means or proportions, as appropriate. It involves no background variables.

Table 4: Composition of the propensity groups (deciles).

Treatment group	Decile										All
	1	2	3	4	5	6	7	8	9	10	
A	1083	1012	930	810	691	588	467	325	136	8	6050
B	111	182	264	384	503	605	727	869	1058	1186	5889
Matched pairs	109	179	256	373	438	435	410	308	133	8	2649
All	1194	1194	1194	1194	1194	1193	1194	1194	1194	1194	11 939

Note: A — no bovine products; B — some bovine products given in the first 14 days of life.

We consider three GA groups; extremely preterm, born at GA of 26 weeks or earlier, medium preterm (weeks 27 – 29) and later preterm (weeks 30 and 31). We regard the (neonatal) network and GA group as primary background variables. The fitted propensities are split into deciles (ten groups of equal size) and matched pairs are formed within these deciles with the constraint that every matched pair has to be from the same network and the same GA group. The impact of this restriction is illustrated on the following example.

The composition of the propensity groups is given in Table 4 and presented by the within-treatment group histograms in Figure 2. The table and diagram show that treatment group A (no bovine products) dominates in the low deciles and is in a minority in the high deciles. If we matched solely within the propensity groups, we would obtain $111 + 182 + \cdots + 136 + 8 = 2968$ matched pairs, accounting for 5936 infants (49.7%). Further 'losses' are incurred because we match also on network and GA group. We obtain only 2649 matched pairs, accounting for 44.4% of the infants, even though the two (original) treatment groups have similar sizes, 6050 and 5889, prior to matching. The number of matched pairs is listed in the third row of Table 4. The additional matching on network and GA group results in $(111 - 109 =)$ two fewer matches in the first propensity decile, three fewer in the second, and so on, and none in the tenth decile. The largest loss is in the sixth decile, $(605 - 435 =)$ 170 pairs, and altogether $(2968 - 2649 =)$ 319 pairs are lost. That is the sacrifice for a more refined matching of the two treatment groups.

Instead of ten, we also consider six propensity groups, as part of a sensitivity analysis. Its purpose is to explore the impact of the details of the matching procedure, some of which entail some arbitrariness, and hopefully confirm that the impact is very small and can be ignored.

Discarding so many infants from the analysis is justified by our emphasis on unbiased estimation, which is promoted by matching. The discarded infants are not a random sample from the set of all infants originally considered. Many of them are not matched because the configuration of their backgrounds is rare in the other treatment group or has been used up in matches with other units. Among the infants

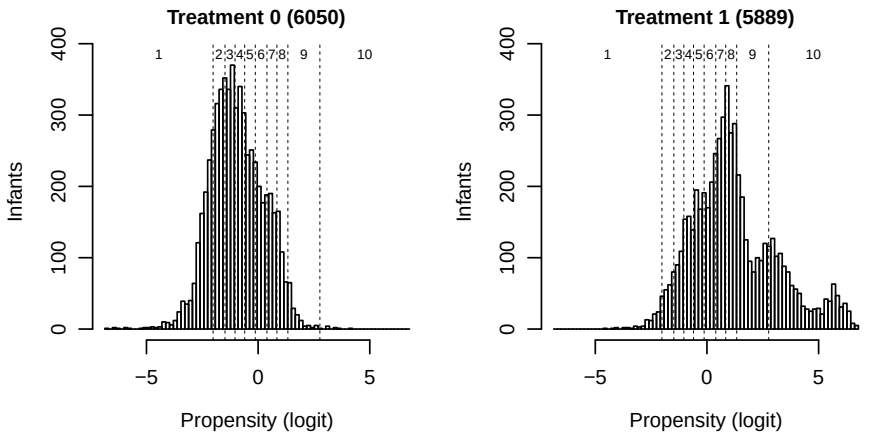


Figure 2: Fitted propensities within the treatment groups for the presence of bovine products. The vertical dashes mark the deciles which separate the propensity groups.

with extreme configurations, and those with the highest and lowest propensities in particular, there are many that would be influential observations in a regression analysis that relates the outcome to the treatment and background variables. Yet, they are least relevant to the comparisons we want to make. In brief, by forming matched pairs, we distil from the original sample a subset of units relevant for the target of the analysis. This can be motivated as an attempt to find a subset that could be analysed by methods for studies with randomised allocation of the treatments.

In a study with treatment assigned by randomisation, the two treatment groups are well balanced on all background variables, unless one or both groups are so small that nontrivial differences between the two groups can arise by chance. That is, they would not arise in (some) other replications of the (randomised) treatment assignment and the imbalance averaged over many replications would be very small for every background variable. In contrast, the balance in matched treatment groups in an observational study is only approximate, and it is arranged only for the variables recorded in the study.

The propensity analysis does not guarantee a good balance of the two treatment groups. We check the balance by evaluating the following summaries. For a categorical variable, we evaluate the differences of the proportions within the treatment groups. For a continuous (ordinal) variable, we evaluate the difference of the means within the treatment groups and scale it by their standard deviation pooled across the groups. We also compare the standard deviations within the treatment groups. We evaluate the logarithm of their ratio, so that the reference (ideal) value is zero.

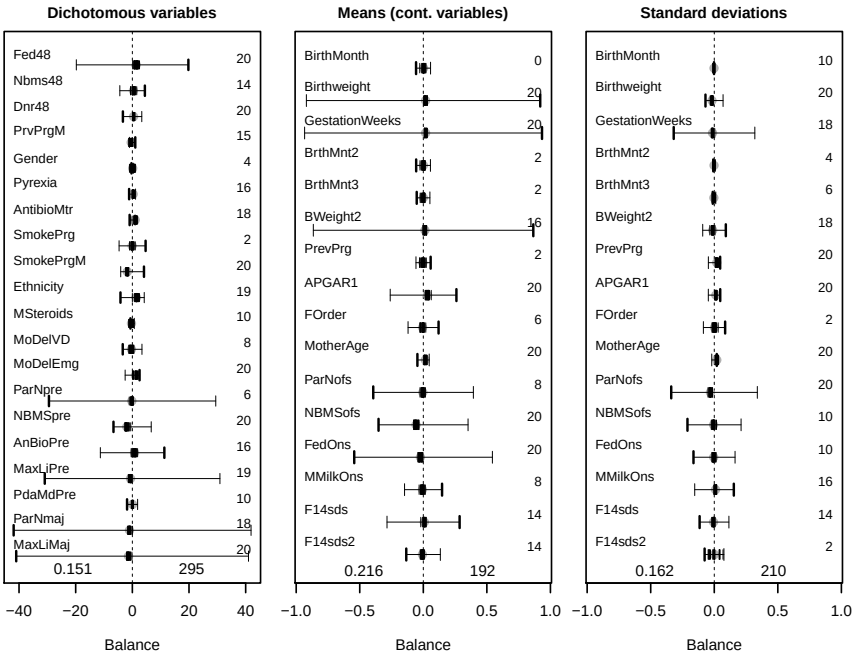


Figure 3: Balance plot for the propensity model for the presence of bovine products.

The balance for a variable is defined as the absolute value of this summary. The total of these summaries characterises the overall balance for a particular propensity model. We start by the propensity model with all the background variables, and then search systematically by adding their interactions and squares of continuous variables, one at a time, and retain a term when the corresponding overall balance is improved. Figure 3 displays the balances for 20 replicate sets of matched pairs.

The balances for the categorical variables (all of them are binary) are plotted in the left-hand panel. Each variable is represented by a horizontal segment that extends from its balance in the original dataset (the difference of the proportions of the variable in the two treatment groups) to its negative. The balance is marked by a fuller vertical tick and its negative by a thinner tick. The balances for 20 replicate sets of matched pairs are marked by shorter vertical ticks, and their average by a grey disc in the background. In some cases, the disc is almost entirely obscured by the tightly packed ticks for the replicate balances.

The replicate balances for a variable are summarised by their mean. A coarser alternative is the balance of the signs, equal to the absolute difference of the number of positive and negative balances. For example, the original (raw) balance for variable Fed48 is 0.198, obtained as the difference $0.543 - 0.345$ of the proportions of Fed48 in the groups of infants who received some bovine products in the first

14 days and those who received none. The balances in the 20 replicate matched groups range from 0.0011 to 0.0253, and their average is 0.0137. All the balances are positive, so the balance of the signs is 20, displayed at the right-hand margin. The balance on Fed48 is improved by matching substantially, but remains imperfect.

For the continuous variables, we study the balance of the means and standard deviations. The balance of the means for a variable is defined as the difference of the within-treatment group means, scaled (divided) by the pooled standard deviation. The means are compared in the middle panel. For each continuous variable, the horizontal line connects the balance for the original dataset (full vertical tick) with its negative (thin tick), and the balances for the 20 replicate matched groups are marked by shorter vertical ticks. Their average is marked by a grey disc. The balances for the matched groups are far superior to the raw balances, although the balances of signs are extreme (close or equal to 20) for several variables. Thus, some residual bias remains, but it is of smaller order of magnitude than in the original (unmatched) groups. The balances of the standard deviations are represented similarly. They are based on $\log(s_{\mathrm{B}}/s_{\mathrm{A}})$, where s_{A} and s_{B} are the within-treatment group standard deviations of the background variable.

The balances are summarised by their totals within the panels. These totals are 0.151 for the dichotomous variables (on the scale of probabilities, not percentages), 0.216 for the means of the continuous variables and 0.162 for their standard deviations. The balances of the signs are summarised similarly by their totals, 295, 192 and 210, for the proportions, means and standard deviations, respectively. They are printed at the bottom of each panel.

The model for propensities is found by a systematic search, aiming to minimise the total of the balances (0.528 in Figure 3). First we evaluate the summary of balance for the model with all the background variables and no interactions. Then we fit models with one interaction added at a time, retaining interactions that yield a lower value of the overall balance. We started with the summary $(0.188 + 0.470 + 0.444 =)$ 1.102 and by adding 25 interactions and five polynomial terms we gradually reduced it to 0.528. The total of the balances for the signs was reduced from 881 to 697.

The value of the overall balance should ideally be such that it could plausibly arise in a randomised study with the same background variables and the same units as in the realised study. This value, a random variable, can be established by simulation, reassigning all the units (infants) to synthetic treatment groups with the same within-treatment sample sizes, and evaluating the balance of these groups. Figure 4 presents the balance plot for a set of such synthetic re-assignments. This 'synthetic' balance is much better than for the matched datasets; compare with Figure 3. The synthetic balances add up to $(0.030 + 0.052 + 0.049 =)$ 0.131, about four times less than for the sets of matched groups, 0.528. However, the corresponding statistic for the original dataset is 9.036, so the matched groups are much better balanced. In brief, the analysis of the matched pairs is unlikely to be without bias, but this bias is of a smaller order of magnitude than the comparison of the raw rates.

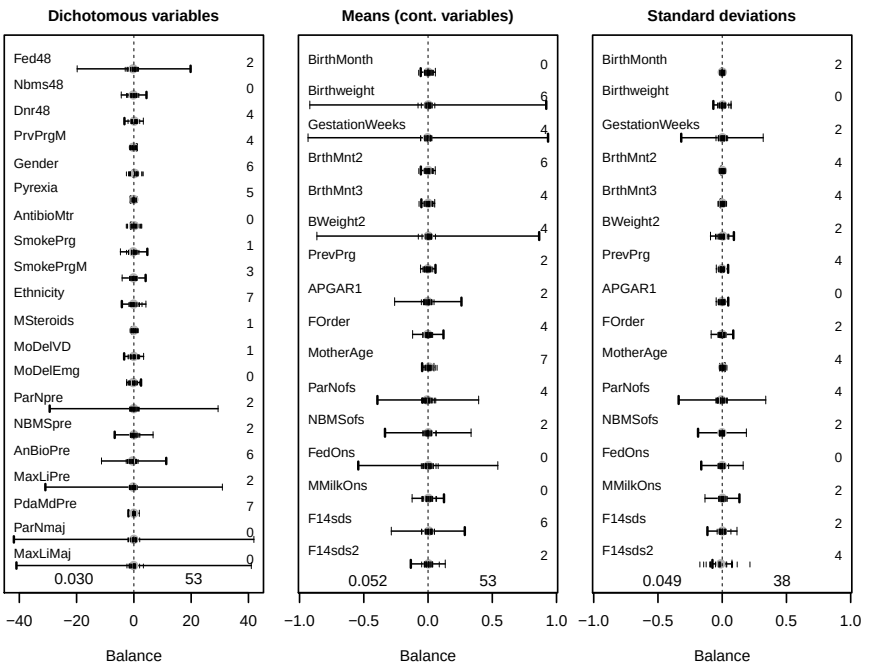


Figure 4: Balance plot for synthetic matched groups generated by would-be randomisation.

5. Results

We estimate the average effect of the presence of bovine products in the feeding regimen on the probability (risk) of contracting NEC. For a set of matched pairs, we evaluate the contrast of the outcomes, $\bar{y}_{\mathrm{B}} - \bar{y}_{\mathrm{A}}$. The estimate of the average treatment effect is their average over the replicate sets of matched pairs. The sampling variance is estimated by the sample variance of the replicate estimates. We evaluate these statistics not only for the final propensity model, but also for intermediate models, to assess the stability (robustness) of the results with respect to the details of the propensity model. We note that the propensity model, and the model fit have no inferential value; their sole purpose is to obtain a better balance.

The approach to minimising the balance has several refinements. First, the summary (total) of the balances can be evaluated with weights, giving greater emphasis to some variables than to others. The extreme of this is to insist on a perfect balance — each matched pair has to have the same value of a variable. We match perfectly on the network and GA group. They are omitted from the balance plot in Figure 3 because their balance is perfect (equal to zero) by construction.

The 20 replicate estimates we obtained are 0.680, 0.491, . . . , 0.680 and 0.868, in percentages. Their mean, the estimate of the average treatment effect, is 0.647%.

Table 5: Rates of NEC in the population and matched pairs.

	All infants (11 939)			Matched pairs			
Treatment	All	NEC	%	All	NEC*	%*	St. error
No bovine products	6050	160	2.64	2649	26.5	1.00	(0.18)
Some bovine products	5889	60	1.02	2649	43.7	1.65	(0.05)
Difference			1.62		−17.2	−0.65	(0.18)

Note: * — average over 20 replications of matching.

The associated standard error is estimated by 0.185. We contrast this with the (biased) estimates based on all the (12 823) infants cared for during their first 14 postnatal days, –2.03%, and on the (11 939) infants not suspected to have NEC in the first 14 days, –1.62%; see Table 5. These estimates have zero standard errors because they are based on the entire population of interest.

In the process of matching, 2649 pairs are formed, accounting for only 44.4% of the population. The comparison of the rates of NEC in the two treatment groups is reversed. Now the estimated rate among infants who received some bovine products, 1.65%, is higher than in the matched group of infants who received no bovine products, 1.00%. The average treatment effect is 0.65%, with estimated standard error 0.19%. These results are based on propensity deciles. We obtained very similar results by matching within six and fifteen propensity groups of equal size.

The trivial (biased) estimates differ substantially from the estimates based on matched groups because bovine products tend to be given to infants who are developing fast and whose gastrointestinal tract is judged to be sufficiently mature. The estimates with the last few propensity models differ very little, suggesting that even if better propensity models were found, the estimates would not differ substantially from the one we obtained. The (provisional) estimates obtained with the last five propensity models are in the range $0.577 - 0.675$, with estimated standard errors in the range $0.159 - 0.185$.

The replicate sets of matched pairs involve only 70 cases of NEC on average, out of 220 in the population. A majority of the cases among infants who received some bovine products are included in the matched groups (73%), whereas only 17% of the cases among those who received no bovine products are included. The rate of matching is so low because inclusion of bovine products in the diet is closely related to the background variables. Infants with certain backgrounds are nearly always given some bovine products in the first 14 days and others are almost never given — most of them cannot be matched, and are not relevant for the analysis. A disconcerting conclusion is that there are configurations (profiles) of background associated with absence of bovine products in the diet, in which NEC is prevalent. If we rule out the possibility that consultants and dieticians are consistently mak-

Table 6: Rates of NEC in the original data and matched pairs; early and late onset of breastfeeding.

	All infants (11 939)			Matched pairs			
Treatment	All	NEC	%	All	NEC*	%*	St. error
Early onset	7835	97	1.24	3081	60.8	1.97	(0.11)
Late onset	4104	123	3.00	3081	87.9	2.85	(0.10)
Difference			−1.76		−27.1	−0.88	(0.14)

Note: * — average over 20 replications of matching.

ing an error by prescribing bovine products, then we would have to conclude that presence or absence of bovine products is not an important factor in preventing NEC among background profiles for which the choice is largely unanimous, where matches across the treatment groups are difficult or impossible to find.

Among profiles where there is a disagreement and matched pairs can be formed, the average effect of bovine products is negative, and the estimated number of additional cases of NEC is 17. It is questionable whether such a small estimated benefit would warrant a clinical trial to confirm it. However, the beneficial effect applies also to some infants who were excluded from the study because they were suspected to have NEC in the first 14 days. Note that our analysis is without a proposal for how the bovine products should be replaced.

Breastfeeding and NEC

We define the treatment variable by the onset of breastfeeding. The focal treatment is onset on days 1 – 7 (early onset). The reference treatment (late onset) includes onset not only in the second week of life, but also later, or even never. The joint distribution of this treatment variable and the outcome is given in Table 6 for all infants in the analysis and for matched pairs, discussed below. The rate of NEC among the infants with early onset is much lower (1.24% vs. 3.00%). If the infants who were at some point in the first fortnight suspected of having NEC are included, the rates of NEC differ even more, 1.28% vs. 3.39%, because there are more additional cases among the infants with late onset of breastfeeding.

The fitted propensities are plotted in Figure 5. They are derived from a model with six interactions added to the 37 background variables listed in Tables 2 and 3, after excluding the variables whose status as background is not compatible with the outcome variable. Excluded are the following variables: onset of feeding (FedOns), offset of Nbms (NBMSofs), and central line installed on majority of the days (MaxLiMaj), because a change of onset of breastfeeding would lead to alteration of these variables. Matching within propensity deciles, network and GA group yields 3081 matched pairs. Matching solely on the deciles would yield 3342 matched pairs.

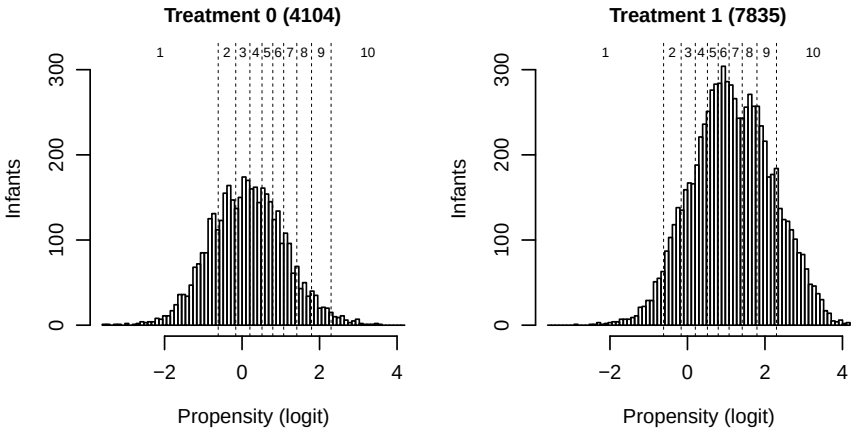


Figure 5: Fitted propensities within the treatment groups for the early onset of breastfeeding. The vertical dashes mark the deciles which separate the propensity groups.

Imputation for missing daily entries for breastfeeding either leaves the onset unchanged or alters it to an earlier day, so some infants are re-classified to the early-onset group. For example, the sequence 0011111 9911111 (onset on day 10) is completed to 0011111 1111111 (onset on day 3). By imputation, 419 infants were reclassified, reducing the late-onset group from 4523 to 4104 infants. In cases in which we did not impute valid entries for missing values, the error in the onset is limited. The error with subsequences 091 and 190 is by one day at most. Similarly, the error with isolated pairs left unchanged is by two days at most.

The balance plot in Figure 6 shows that overall a balance much finer than for the original (unmatched) treatment groups has been achieved. The summaries of the balance are (0.333, 573) for the adopted propensity model, developed by two rounds of systematic search, starting with the model with no interactions, for which the summaries of the balance are (0.497, 720).

The estimated rates of NEC in the matched pairs are 1.97% and 2.85% for the reference (early onset) and focal group (late onset), respectively, yielding the estimate of the average treatment effect in the matched groups 0.88%. It is associated with (estimated) standard error 0.14. The estimated percentages correspond to 60.8 and 87.9 infants (averages over 20 replications), so we estimate that about 27 cases of NEC would be avoided by switching from later to earlier onset of breastfeeding. The matched pairs include on average 149 cases of NEC; 71 cases are not matched, 36 with early onset and 35 with late onset of breastfeeding. These infants, together with 5706 non-cases, are not involved in matched pairs because, in effect, the alternative treatment would be considered for them very rarely or not at all, and therefore the counterfactual question of what would have happened if they were subjected to the other treatment is not well posed in our setting.

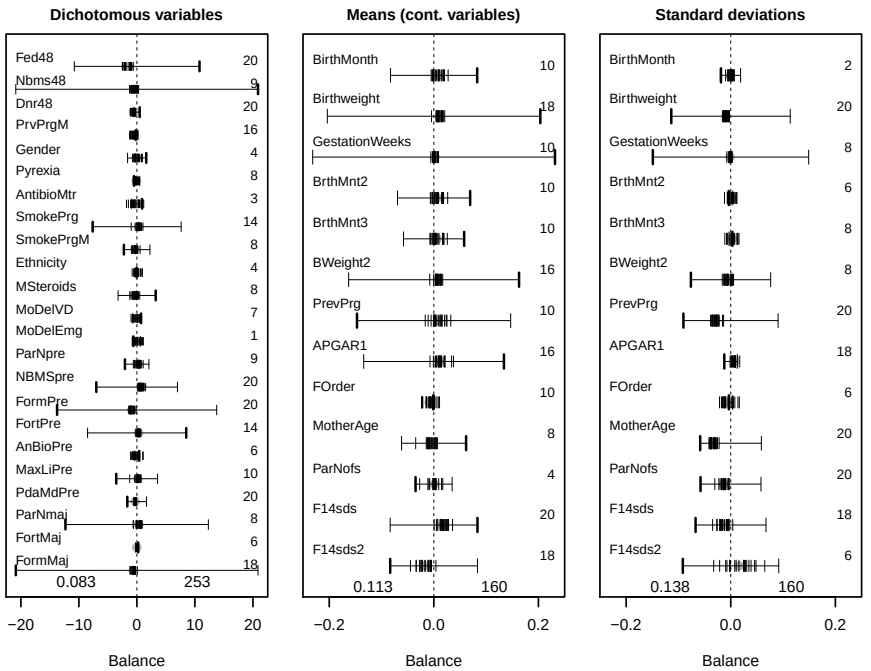


Figure 6: Balance plot for the propensity model for the early onset of breastfeeding.

These results were obtained with matching within propensity deciles. Without the imputations for isolated missing entries and pairs, the estimate of the treatment effect would be somewhat lower, 0.56%.

Changes in the feeding regimen

Lack of stability in the feeding regimen may be a cause of problems with the gastrointestinal tract, and ultimately of NEC. We study this issue by defining a treatment variable that reflects the changes in the regimen. For each type of feeding we mark the days on which a change has occurred, and discard the first change. For example, for the pattern 111$\underline{0}$0$\underline{1}$1 $\underline{0}$000000 for a type of feed over 14 days there are changes on days 4, 6 and 8 (underlined), so days 6 and 8 are marked. For the variables that describe the daily feeding, we count the number of marked days, omitting any duplicates. The distribution of this variable for cases and non-cases of NEC is given in Table 7. We define the (dichotomous) outcome variable as the indicator of four or more changes.

The rates of NEC among the infants cared for for at least the first 14 days are 1.49% (111/7344) for those with four or fewer changes and 3.01% (109/4375) for those with more than four changes.

Table 7: Number of changes in the daily feeding regimen and NEC.

	Number of changes												
	0	1	2	3	4	5	6	7	8	9	10	11	12
Cared for on day 14													
No NEC	955	1786	2153	2028	1884	1437	1043	668	351	166	56	17	1
NEC	14	27	38	40	43	37	34	19	11	11	2	2	0
% NEC	1.44	1.49	1.73	1.93	2.23	2.51	3.16	2.77	3.04	6.21	3.45	10.53	0.00
Not suspected to have NEC by day 14													
No NEC	935	1730	2050	1905	1736	1303	947	594	306	152	45	15	1
NEC	14	21	30	30	37	26	29	15	9	6	2	1	0
% NEC	1.48	1.20	1.44	1.55	2.09	1.96	2.97	2.46	2.86	3.80	4.26	6.25	0.00

The propensity model with 13 interactions yields the overall balance (0.471, 536), reduced from $(0.693, 704)$ for the model with no interactions. Based on the adopted propensity model, and matching additionally on the network and GA group, 2968 matched pairs are formed, 49.7% of the studied population. Without matching on the network and GA 3158 matched pairs would be formed. The matched pairs contain 125 cases of NEC, just over half of all cases. The estimate of the average treatment effect is –0.07% (higher probability of NEC with four or more changes), with estimated standard error 0.14. This provides too weak support for the proposal to reduce the number of changes in the daily feeding regimen.

6. Discussion and conclusion

We applied the potential outcomes framework to estimate the effect of treatments on the incidence of NEC among infants born at GA of 32 weeks or earlier. The clear separation of matching, which does not involve the outcomes, and the comparison, in which the background variables are not involved, is a great conceptual advantage over methods based on regression which carry the additional baggage of the model assumptions (Rubin, 2005). With the potential outcomes framework, there are only two essential assumptions; that all the relevant background variables are recorded and that a balance of the matched groups as good as with randomisation has been achieved. The first assumption is common to all regression-based approaches. The second can be assessed directly by comparing the distributions of the background variables within the matched treatment groups.

The appeal of the framework is in comparing matched groups of infants in two treatment groups, for which the same analysis can be applied as in a hypothetical randomised study. The matched treatment groups are selected after a systematic search of propensity models. Some imbalance remains, and therefore also some residual bias in estimating the treatment effect. We conjecture that the bias is small

because the estimates of the treatment effects with propensity models that yield slightly inferior balance differ only slightly.

The estimated treatment effect is interpreted as the change (reduction) in the rate of NEC that would have resulted from the corresponding change in the treatment. Removal of bovine products (formulas and fortifiers) might in principle be easy to implement, but these products are invaluable for the growth of infants who are not threatened by NEC. Therefore, the constituency of the treatment has to be carefully qualified. Early onset of breastfeeding is generally encouraged, and the only issue is whether it is given sufficient priority. The constituency of infants is defined principally by availability of MBM. These sources of bias can be interpreted as imperfections in the definition of the target population and the treatments, because the treatments we defined cannot be manipulated for all infants.

A further source of bias in our analysis is the exclusion of infants who were suspected of having (or developing) NEC in the first 14 days. For them, the feeding regimen could not be used as background because it is (indirectly) affected by the outcome.

We selected for the background the period of 14 days because it is generally regarded as a landmark in neonatal care. Preterm born infants are rarely discharged earlier, but those deemed not to require intensive care are discharged soon thereafter. Earliest cases of NEC tend to be recorded after 21 postnatal days. Definitions of some of the treatments would be less natural if a period that differed from two weeks by a few days were selected.

A randomised clinical trial is regarded as the gold standard for comparing alternative treatments. Our analysis informs its design by obtaining an estimate that can be regarded as preliminary and can be used as input in a sample size calculation. We note that a clinical trial is likely to encounter difficulties that undermine its full potential. First, recruitment of a large number of preterm-born infants is difficult and requires a long period of concentrated effort to enlist many neonatal care units and agree with them on the terms of the cooperation. Second, clinical priorities and parents' wishes may result in dropout and other forms of noncompliance. The target population (inclusion criteria) and the details of the treatment options have to be defined with care, so that randomisation would be acceptable and either treatment could be applied.

Dawid (2015) and Dawid, Musio and Fienberg (2016) have challenged the potential outcomes framework on several counts, foremost that it cannot identify causes (treatments or interventions), merely compare them. We agree that our analysis is concerned with a search for causes, but we have a short list of candidates that can be fitted into the framework. A strong suit of this approach is its appeal to the clinical community who are acquainted with clinical trials and find analyses that are closely related to them appealing.

Data for the analysis described in this paper were extracted from NNRD using SAS procedures. The R language and environment for statistical computing and graphics was used for all the analysis. The computer code, in the form of R functions compiled specifically for this project is available on request from the author.

Acknowledgements

This paper presents independent research funded by the National Institute for Health Research, UK (NIHR), under its Programme Grants for Applied Research Programme (Grant Reference Number RP–PG–0707–10010). The views expressed are those of the author and not necessarily those of the National Health Service, NIHR or the UK Department of Health. Assistance of Daniel Gray and Eugene Statnikov with the extraction of the data from NNRD is acknowledged. The paper has benefited from invaluable cooperation with Cheryl Battersby.

REFERENCES

AMERICAN ACADEMY OF PEDIATRY, (2012). Breastfeeding and the use of human milk. Policy statement. Pediatrics 129, pp. e827–e841.

ARSLANOGLU, S., CORPELEIJN, W., MORO, G., BRAEGGER, G., CAMPOY, C., COLOMB, V., DECSI, T., DOMELLOF, M., FEWTRELL, M., HOSAK, I., MIHATSCH, W., MOLGAARD, C., SHAMIR, R., TURCK, D., VAN GOUDOEVER, J., and ESPGHAN COMMITTEE ON NUTRITION, (2013). Donor human milk for preterm infants: current evidence and research directions. Journal of Pediatric Gastroenterology and Nutrition 57, pp. 535–542.

BATTERSBY, C., LONGFORD, N., COSTELOE, K., and MODI, N., (2017a). Development of a gestational age-specific case definition for neonatal necrotizing enterocolitis. Journal of American Medical Association 171, pp. 256–263.

BATTERSBY, C., LONGFORD, N., MANDALIA, S., COSTELOE, K., and MODI, N., (2017b). Incidence and enteral feed antecedents of severe neonatal necrotising enterocolitis across neonatal networks in England, 2012–13: a whole-population surveillance study. The Lancet Gastroenterology and Hepatology 2, pp. 43–51.

BELL, M.J., TERNBERG, J.L., FEIGIN, R.D., KEATING, J.P., MARSHALL, R., BARTON, L., ET AL., (1978). Neonatal necrotizing enterocolitis: therapeutic decisions based upon clinical staging. Annals of Surgery 187, pp. 1–7.

CARPENTER, J. R., KENWARD, M. G., (2013). Multiple Imputation and its Application. Wiley, New York.

DAWID, A.P., (2015). Statistical causality from a decision-theoretical perspective. Annual Review of Statistics and Its Application 2, pp. 273–303.

DAWID, A. P, MUSIO, M., FIENBERG, S. E., (2016). From statistical evidence to evidence of causality. Bayesian Analysis 11, pp. 725–752.

HOLLAND, P. W., (1986). Statistics and causal analysis. Journal of the American Statistical Association 81, pp. 945–960.

IMBENS, G. W., RUBIN, D. B., (2015). Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction. Cambridge University Press, New York.

JONES, B., KENWARD, M. G., (1989). Design and Analysis of Crossover Trials. 2nd ed. Chapman and Hall/CRC, London.

KLIEGMAN, R.M., WALSH, M.C., (1987). Neonatal necrotizing enterocolitis: pathogenesis, classification, and spectrum of disease. Current Problems in Pediatrics 17, pp. 243–288.

LITTLE, R. J. A., RUBIN, D. B., (2002). Statistical Analysis with Missing Data. 2nd ed. Wiley, New York.

LONGFORD, N. T., (2004). Missing Data and Small-area Estimation. Modern Analytical Equipment for the Survey Statistician. Springer, New York.

NEU, J., (2015). Preterm infant nutrition, gut bacteria, and necrotizing enterocolitis. Current Opinion in Clinical Nutritional Metabolism Care 18, pp. 285–288.

NEU, J., WALKER, W. A., (2011). Necrotizing enterocolitis. New England Journal of Medicine 364, pp. 255–264.

PATEL B. K., SHAH, J. S., (2012). Necrotizing enterocolitis in very low birth weight infants: a systemic review. Gastroenterology, PMC3444861.

QUIGLEY, M., MCGUIRE, W., (2014). Formula versus donor breast milk for feeding preterm or low birthweight infants. Cochrane Database of Systematic Reviews, CD002971; 14th April 2014.

ROSENBAUM, P.R., RUBIN D.B., (1983). The central role of propensity score in observational studies for causal effects. Biometrika 70, pp. 41–55.

RUBIN, D. B., (1976). Inference and missing data. Biometrika 63, pp. 581–592.

RUBIN, D. B., (1978). Bayesian inference for causal effects: The role of randomization. Annals of Statistics 6, pp. 34–58, pp. 961–962.

RUBIN, D. B., (2002). Multiple Imputation for Nonresponse in Surveys. 2nd ed. Wiley, New York.

RUBIN, D. B., (2005). Causal inference using potential outcomes: design, modeling, decisions. 2004 Fisher Lecture. Journal of the American Statistical Association 100, pp. 322–331.

RUBIN, D. B., THOMAS, N., (1996). Matching using estimated propensity scores: Relating theory to practice. Biometrics 52, pp. 249–264.

SCHANLER, R. J., LAU, C., HURST, N. M., SMITH, E. O., (2005). Randomized trial of donor human milk versus preterm formula as substitutes for mothers' own milk in the feeding of extremely premature infants. Pediatrics 116, pp. 400–406.

SULLIVAN, S., SCHANLER, R. J., KIM, J. H., PATEL, A. L., TRAWÖGER, R., KIECHL-KOHLENDORFER, U., CHAN, G. M., BLANCO, C. L., ABRAMS, S., COTEN, C. M., LAROIA, N., EHRENKRANTZ, R.A., DUDELL, G., CRISTOFALO, E. A., MEIER, P., LEE ,M. L., RECHTMAN, D. J., LUCAS, A., (2010). An exclusively human milk-based diet is associated with a lower rate of necrotizing enterocolitis than a diet of human milk and bovine milk-based products. The Journal of Pediatrics 156, pp. 562–567.

VAN BUUREN, S., (2012). *Flexible Imputation of Missing Data.* Chapman and Hall/CRC, London.

WORLD HEALTH ORGANIZATION, (2011). Guidelines on optimal feeding of birthweight infants in low- and middle-income countries. World Health Organization, Geneva, Switzerland.

APPENDIX

Table 8 summarises the missing values in the daily regimens of the infants. It lists for each network the number of infants in the analysis (N), and the averages of the numbers of items missing, as well as days, variables and infants who have these missing items. Table 9 lists the numbers relevant to imputations for isolated missing items. There are 28 379 missing items; 11 244 of them are isolated, 9400 of them are imputed in 8368 distinct 14-digit sequences. They involve 1987 infants. Table 10 displays similar information about imputations for isolated pairs of missing values. There are 7100 missing entries on day 1. By definition, they are not isolated, and imputation is not performed for them.

Table 8: Missing entries in the daily feeding regimen.

		Percent incomplete			
Network	N	Items	Days	Variables	Infants
BedHer	319	1.28	2.84	12.14	26.96
Kent	391	0.81	1.83	9.07	20.72
LDNsw	364	1.49	3.36	14.44	31.59
NTrent	524	0.59	1.31	6.68	14.69
SurSx	495	2.64	5.86	19.76	43.84
CheMer	256	1.31	2.90	12.46	27.73
LDNnc	344	1.27	3.09	12.45	27.62
LanSCu	325	1.95	4.31	16.11	35.38
North	657	0.95	2.10	10.77	23.74
Trent	392	2.49	5.52	19.67	43.62
Easter	619	1.63	3.67	13.75	30.37
LDNne	864	1.42	3.19	11.30	24.88
MidBl	493	1.00	2.20	7.65	16.84
Penins	276	1.51	3.31	11.69	25.72
West	583	3.65	8.10	17.76	39.45
GManch	719	1.26	2.77	12.09	26.56
LDNnw	653	1.45	4.55	11.10	26.19
Midcn	632	1.83	4.05	16.00	35.44
SouCN	470	1.30	2.90	11.61	25.74
Yorks	803	1.93	4.25	17.46	38.61
LDNse	534	1.24	2.78	12.80	28.28
Midsw	643	1.16	2.59	11.51	25.51
SouCS	583	1.52	3.79	13.36	29.67
All	11939	1.56	3.57	13.12	29.11
Minimum		0.59	1.31	6.68	14.69
Maximum		3.65	8.10	19.76	43.84

Table 9: Summary of imputations for isolated missing items in the feeding regimen.

	Counts					
Network	Missing	Isolated	Imputed	Changes	Infants	Percent
BedHer	631	221	175	153	39	12.2
Kent	483	236	199	185	44	11.3
LDNsw	831	330	267	249	61	16.8
NTrent	480	295	256	235	54	10.3
SurSx	2012	842	726	585	132	26.7
CheMer	516	256	219	206	49	19.1
LDNnc	650	240	197	182	44	12.8
LanSCu	976	456	377	322	78	24.0
North	962	400	304	289	76	11.6
Trent	1501	576	475	394	93	23.7
Easter	1547	597	510	456	107	17.3
LDNne	1882	746	624	580	138	16.0
MidBl	760	300	255	232	53	10.8
Penins	640	165	143	130	30	10.9
West	3274	484	418	390	91	15.6
GManch	1391	721	616	555	129	17.9
LDNnw	1272	346	286	262	69	10.6
Midcn	1780	865	728	621	144	22.8
SouCN	937	390	311	282	70	14.9
Yorks	2380	1115	923	795	187	23.3
LDNse	1017	481	404	368	86	16.1
Midsw	1142	577	491	453	107	16.6
SouCS	1315	605	496	444	106	18.2
All	28 379	11 244	9400	8368	1987	16.6
Minimum	480	165	143	130	30	10.3
Maximum	3274	1115	923	795	187	26.7

Note: The columns contain the counts of: Missing — missing values in the eleven variables that indicate the daily elements of the feeding regimen; Isolated — missing values that are preceded *and* followed by recorded (valid) values; Imputed — imputations made for isolated missing entries (with agreement of the adjacent values); Changes — 14-digit records altered; Infants — infants involved in these records; Percent — percentage of the infants with changes.

Table 10: Summary of imputations for isolated pairs of missing items in the feeding regimen.

	Counts			
Network	Pairs	Imputations	Changes	Infants
BedHer	35	24	24	6
Kent	15	11	11	3
LDNsw	20	10	10	3
NTrent	10	10	10	2
SurSx	85	50	50	13
CheMer	25	20	20	5
LDNnc	15	12	12	3
LanSCu	65	53	53	12
North	10	8	8	2
Trent	55	41	41	10
Easter	80	59	59	16
LDNne	75	47	47	13
MidBl	10	10	10	2
Penins	5	3	3	1
West	45	28	28	7
GManch	55	31	31	9
LDNnw	43	16	16	8
Midcn	115	86	86	21
SouCN	40	28	20	5
Yorks	75	50	50	13
LDNse	35	23	23	7
Midsw	35	30	30	7
SouCS	60	37	37	11
All	1008	687	679	179
Minimum	5	3	3	1
Maximum	115	86	86	21

Note: The columns contain the counts of: Pairs — isolated pairs of missing items — missing values that are preceded *and* followed by at least two recorded (valid) values each; Imputed — imputations made for isolated missing entries (with agreement of the adjacent pairs of values); Changes — 14-digit records altered; Infants — infants involved in these records.

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 119–134, DOI 10.21307/stattrans-2018-007

INDICATORS OF INTERNATIONAL TRADE ORIENTATION OF UKRAINE IN THE CONTEXT OF ASSESSMENT OF THE EFFECTIVENESS OF ITS EXPORT RELATIONS

N. Reznikova[1], **O. Osaulenko**[2], **V. Panchenko**[3]

ABSTRACT

The approach to study the significance of trade relations between countries by analysing economic vulnerability, economic sensitivity, symmetry and asymmetry of the established economic links is proposed in the paper. This approach is adapted to an analysis of the trade dependence of Ukraine. The estimated interdependence ratios for Ukraine and its largest trade partners – the EU, the Russian Federation, post-Soviet countries, China, the USA and Brazil and India as emerging economies – are compared with the respective ratios of Ukraine's dependence on these countries' markets. The analysed dynamics of Ukraine's GDP dependence on Ukraine's trade partners shows a growing relative weight of the countries that have not had a substantial role in the foreign trade of Ukraine. The proposed approach for estimating the quality of the established trade relations is supposed to contribute to the radical transformation of Ukraine's foreign trade.

Key words: economic policy, export orientation, trade relations, trade dependence

1. Introduction

A relatively high level of Ukraine's economy integration causes the objective necessity in building up the national economic policy as a response to globalization challenges. It needs to be based on adequate understanding of the mechanism of interaction between the national economy performance and exports as "a channel" linking the country with the global economy. This link between exports and economic growth has two essential dimensions:

1. Direct causality between exports and economic growth, with exports considered as a key factor for the national economy development. This idea was laid by some countries in the strategy of export-led growth.
2. The causality between growth in exports and dynamics and structure of national GDP: GDP is the determinant of exports.

[1] Doctor of Economics, Professor of the Chair of World Economy and International Economic Relations of Institute of International Relations Taras Shevchenko National University of Kyiv. Ukraine.
[2] Doctor of Public Administration, Professor, Corresponding Member of National Academy of Sciences of Ukraine, Rector of National Academy of Statistics, Accounting and Audit, Kiev. Ukraine.
[3] Post-PhD at Mariupil State University. Ukraine.

Yet, in practice national economic strategies are set by countries through combining elements of the first and the second approach, with the significance of exports as a factor of economic growth and the correlation between GDP and exports revised in view of various internal and external economic and political factors.

2. Export structure quality and economic growth

2.1. Theoretical Framework

According to IMF experts, foreign trade policy and optimization of the structure of trade partners is a foremost factor behind economic development and convergence in developing countries. North American and European analysts of current economic practices observe positive correlation and strong impact of export-oriented strategies on economic development of countries with transitional economy. Empirical studies of correlation between exports and economic development have become a common matter for experts in international economy since the axiom “exports lead to economic development” was put forward by C. P. Kindleberger [Kindleberger 1962]. A. Krueger puts emphasis on empirical evidence to a strong positive impact of the development of trade, diversification of trade partners and a clear export-oriented strategy for economic growth [Krueger 1988]. According to J. E. Stiglitz, the most part of empirical regressions demonstrates a strong correlation between measures of external openness, i.e. foreign trade, stimulation of exports, tariff, indexes of price distortions, and growth of incomes per capita [Stiglitz 1999]. Due to the data accuracy problems, modern researches tend to use various empirical strategies to study economic openness versus economic growth. These strategies include: (i) the use of openness indicators (D. Dollar [Dollar 1992], J. Sachs and A. Warner [Sachs 1995]); (ii) reliability testing by the use of a wide range of openness criteria, including subjective indicators (S. Edwards [Edwards 1993; Edwards 1998]); (iii) comparisons of convergence practices in groups of liberalized and non-liberalized countries (D. Ben-David [Ben-David 1993]). P. Romer proposes to use the spatial component as a tool to find out the impact of grade on incomes level [Romer 1986]. J. Sachs and A. Warner attempt to measure the index of openness, combining information on several aspects of trade policy, by surveys in 79 countries [Sachs 1995]. It follows from their results that an economy is considered closed once five criteria are met: (i) average tariff rates lower than 40%; (ii) non-tariff barriers applied to more than 40% of the imports; (iii) the economic system is socialist; (iv) government monopoly on a major part of exports; (v) the share of the shadow economy is larger than 20%. The researchers come to the conclusion that the above five criteria have 2.44 percent negative impact on economic growth. Significance t-test is 5.5, and the probability of error is lower than 0.1%. As a significant change would not occur when the first three criteria were not applied, it is the scopes of the shadow economy and the government monopoly which have essential negative effects on economic growth. A. Harrison studies correlation between trade policy and economic development and observes effects from trade liberalization in many countries [Harrison 1991]. He uses seven indicators of trade policy, including the share of the shadow market, the level of

trade prices and trade liberalization index of the World Bank. These are indicators having strong correlation with economic development of a country. P. Wacziarg, determining indicators of economic openness that have impact on economic growth, constructed trade policy index as the combination of three indicators: average import rate, share of non-tariff constraints and Sachs-Warner indicator [Wacziarg 2001]. The effectiveness of the established foreign economic relations of a country was measured by R. M. Kunst, D. Marin through analysing the causality between labour productivity and exports [Kunst 1989]. From the analysis of the output in industrial sectors they were able to find out that while exports had no impact on productivity, productivity did have impact on exports. J. A. Hatemi and M. Irandoust found a causal relation between exports and two factors, labour productivity and total factor productivity growth, by the use of data for five developed countries [Hatemi 2001]. A review of scientific publications devoted to the impact of export structure on economic development demonstrates that the problem remains to be important, but insufficiently explored; it, therefore, requires further studies.

2.2. Ukrainian international trade orientation versus global trends

In the years following 1991, when Ukraine gained independence, its exports were comparable with some of the European countries. In the following 20 years or more, each of these countries could increase exports to a significant extent: in Poland exports grew by 14.1 times, in Hungary by 11.1 times, in Turkey by 9.4 times. Yet, in Ukraine it was only 4.7 times.

Given that Ukrainian exports fell by 30.14% in 2015, in absolute figures they amounted to 37.8 billion USD, which is 1.9 billion USD lower than in the crisis year of 2009, when there was an unprecedented decline in exports of 40.7%. From the macroeconomic perspective, in the years of independence Ukraine failed to achieve significant success in economic policy reforms: its results were mostly bad except for years of good market conjuncture for key commodities groups of Ukrainian exports (see Table 1).

Table 1. GDP and foreign trade of Ukraine

	2000	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Real GDP, % to the previous year	5.9	2.7	7.3	7.9	2.3	-15.1	4.1	5.2	0.2	-4.6	-7.5	-10.0	2.3
Exports of goods and services, % to the previous year	25.8	4.8	12.1	28.5	35.8	-40.7	-30.7	-11.2	-7.4	-5.2	-27.6	-30.14	-4.1
Imports of goods and services, % to the previous year	17.8	24.6	24.6	34.6	41.1	-46.9	-26.8	-2.2	-0.8	-3.4	-26.5	-27.6	3.7
Trade balance, % of GDP	1.97	1.52	-2.67	-5.05	-7.54	-1.18	-2.22	-4.03	-5.12	-3.53	3.92	1.24	0.36

The argument that all the troubles of Ukraine are caused by the withdrawal of a part of its territory, warfare and destruction of the largest industrial region seems to be rather controversial. The most rapid decline of the national economy was in 90s of the past century: over the earliest five years of independence Ukraine lost nearly 60% of GDP. This rate of collapse is twice higher than the rate of the American economy's decline in the times of Great Depression.

These disappointing results of economic performance in Ukraine, foreign trade in particular, over the long time justify applications of unconventional methods for analysis of interdependence of countries that are trade partners, in order to find the challenges faced by the Ukrainian economy and demonstrate the need to diversify its export structure and destinations. This determines the objective of the study.

The structure of exports of goods and services in a country is conditional on the impact of international demand on them, and the performance and profile of its economy. The structure of Ukrainian exports of goods only partly corresponds with the global one. While the global exports are dominated by mechanical equipment (23.7%), mineral products (18.8%), transport vehicles (9.9%), chemicals (8.8%) and non-precious metals, the share of mechanical equipment in Ukrainian exports (10.5%) is twice lower than the global average. The share of non-precious metals and products made thereof in the total Ukrainian exports (29.8%) is essentially higher than the global average. The share of mineral goods is lower than the global average (12.5% for Ukraine against 18.8% global average), although the difference is rather small compared with other commodity groups. At the same time, the share of plant products in Ukrainian exports (11.9%) is nearly four times higher than the global average. Ukrainian fats and oils account for 3.5% of the global market, plant products – 1.6%. The rates of growth in Ukrainian exports of agricultural and food products outpace the global ones, which confirms that the global demand for these products has been stable and their producers have not been exposed to crisis-specific pressure of the Ukrainian economy.

Basically, Ukrainian exports feature a relatively low share of industrial products with a high value added and larger share of basic metals, agricultural and food products. Ukraine is a global leader in exports of selected commodity groups (see Table 2).

Table 2 Positions and shares of Ukrainian exports at selected global commodity markets in 2016

Commodity position	Global position	Market share, %	Main importers
Crops	7	5,5	Egypt, Spain, Saudi Arabia, China
Fats and oils	6	4,0	India, China, Iran, Spain
Ores	10	1,5	China, Czech Republic, Poland, Austria, Slovakia
Ferrous metals	11	3,0	Turkey, Russia, Italy, Egypt, Poland

Ukrainian exports of services are dominated by transport (40.9%) and business (14.0%) services; travel services (19.8%). The three categories of

services, although in slightly different proportion, dominate at the global market: travel services (24.1%), business services (19.8%), and transport services (18.5%). The share of services on processing of material resources in Ukrainian exports (8,5%) is more than thrice higher than the global average (2.5%). Growth in Ukrainian exports of these services outpaced the global average in absolute (growth rates) and relative (export shares) terms.

As shown by the analysis, nearly 19.9% of the output of Ukrainian goods and services is exported. Export orientation is the strongest in manufacturing industry, where the export share is 41.0%: the largest share of exports is in mechanical engineering, textiles and basic metals, whereas the smallest one is in coke and other non-metal products. Mining industry exports nearly 27.8% of the output, with the share of exports being the highest for metal ores. Mining industry is followed by agriculture, where export orientation (the share of exports in the output) is 23.1%. The smallest share of exports is in services, where only 8.1% of the output is exported; the service sector in Ukraine is, therefore, strongly oriented on the domestic market.

3. Methodology for quality assessment of Ukraine's trade with partner countries

To our opinion, interdependence should be interpreted in view of the two critical characteristics: *sensitivity* and *vulnerability*. *Sensitivity* refers to direct and primary costs that can be imposed by one of the partner countries by changing interdependent relations between two partner countries. Sensitivity is associated with the severity of losses resulting from an unpredictable change. Vulnerability, on the other hand, is conditional on the country's capability to recover after losses resulting from the change in the policy of another country. R. Cooper elaborates on conceptual differences between sensitivity and vulnerability, and addresses these concepts as the two parallel definitions to separate forms of interdependence. Interdependence associated with vulnerability refers to the costs that a country has to bear (when the economic relations are disrupted), in order to do without trade transactions with its already former trade partner. These costs are classified in the public costs met by a country to the extent of its capacities, once it could adapt to the new situation.

On the other hand, interdependence associated with sensitivity acts as a tool for short-term corrections of public costs that a government has to impose on foreign policy measures in response to departures from established standards or economic practices. Therefore, while interdependence associated with sensitivity involves the costs related to *maintenance* of economic relations with another country, interdependence associated with vulnerability refers to the costs required for *disruption* of such relations.

Yet, this theoretical modelling cannot solve the problem related to the manifestation of these costs' effects. The concept of interdependence cannot be systematized unless the causal factors behind these benefits or final costs are found out, because it would be too difficult to extract systematically the vulnerability component without understanding the factor causing these costs. An in-depth analysis of the most typical variations in cross-country interactions gives reaffirming arguments of the essential modification in the meaning of the

dependence phenomenon, caused by endogenous and exogenous factors. Manoeuvring between economic vulnerability and sensitivity, between internal and external dependence allows us to interpret the condition of economic interdependence as the intermediate and equidistant case between the two extreme cases of full dependence and full dominance.

The concept of significance refers to the importance of trade relations relative to other trade relations. The significance of trade for one country in bilateral trade relations will not be always similar to its trade partner. For example, in the case of trade relations between Ukraine and the EU, their significance is much higher for Ukraine than for the EU.

The reduction in Ukrainian exports of goods to the EU to as low as 3.98 billion USD, or by 23.5%, in 2015 can be partially explained by the stoppage of industrial activities on occupied territories, because before the warfare in Lugansk and Donetsk regions started these regions' share in the national exports had reached 27%.

Insignificance of exports of Ukrainian goods and services for the EU market is confirmed by their share in the total imports of the EU, ranging from 0.27 to 0.36%: rarely found across the EU, Ukrainian goods and services do have low priority for the EU market.

The deepened and comprehensive free trade zone between Ukraine and the EU was launched in January 2016, which was supposed to push up modernization of the Ukrainian economy due to the increasing scopes of trade and improved regulatory mechanisms in Ukraine in conformity with the European practice.

When measured by ratio of exports of goods and services to GDP, the Ukrainian economy is even more open than the EU economies. The average export share of Ukraine was higher than the EU by 6.4 percentage points in 2005–2015 (except for 2013), which is an indication of a high dependence of the Ukrainian economy on global market conjuncture (see Table 3).

Table 3 Indicators of exports of goods and services from Ukraine to the EU in 2005–2015

Indicator	2005	2008	2009	2010	2011	2012	2013	2014	2015
Exports of goods to the EU, billion USD	10233	18130	9499	13052	17970	17081	16759	17003	13015
% to the previous year	92.9	130.3	52.4	137.4	137.7	95.1	97.8	102.6	76.5
Exports of services to the EU, billion USD	1766	4066	3021	3117	3525	3745	4196	3992	2928
% to the previous year	113.4	136.5	74.3	105.6	113.2	106.4	111.9	95.1	73.4
The share of Ukrainian exports of goods and services in the total imports of the EU, %	0.29	0.36	0.27	0.30	0.34	0.36	0.35	0.35	0.31

The key aspect that we are going to emphasize when interpreting the concept of interdependence is *symmetry* in cross-country relations. It is argued that the significance of economic relations can vary in the dyad of countries, whereas the symmetry indicates the relative equality of their economic interdependence. A potential case of the ideal symmetry is when both countries are equally dependent on each other. The ideal asymmetry occurs when one country is fully dependent on its trade partner, but this partner is almost independent from the former country. Yet, considering that each country's dependency is a function of the total exports and imports between them, and this total does not equal zero for one country in the dyad, the total will not be zero for the other country as well. Therefore, the case when one country is absolutely independent from the other country can only occur when the other country is also fully independent.

The interdependence establishes the relative importance of bilateral trade relations for each of the countries compared with the amounts of their total trade (in both cases imports and exports are accounted for). For two countries (Country *i* and Country *j*), $TradeShare_{ij}$ measures the ratio of economic exchange between countries *i* and *j*, and the exchange of County *i* with all the partners.

$$TradeShare_{ij} = \frac{DyadicTrade_{ij}}{TotalTrade_{i}} \tag{1}$$

Where $DyadicTrade_{ij}$ is the total imports and exports between Country *i* and Country j, $TotalTrade_{ij}$ is the total imports and exports of Country i with all the partners.

This ratio can range between 0 and 1, with 0 indicating absence of imports or exports between Country *i* and Country *j*, and 1 showing that Country i has international trade relations only with Country j. Using the basic share of trade derived by (1), the significance of interdependence between two countries can be estimated by multiplying the share of $TradeShare_{ij}$ for both countries and taking square root from the product by the formula:

$$Salience_{ij} = \sqrt{TradeShare_{ij} \cdot TradeShare_{ji}} \tag{2}$$

The low level of dependence for one country decreases the overall significance of the relations in a dyad of countries. The overall significance for each of the two countries can be estimated by the use of $TradeShare_{ij}$ for each country.

Estimating the dependence of Country i on Country j on Gross Domestic Product (GDP) of Country is calculated according to the formula:

$$Depend_{ij,t} = \frac{X_{ij,t} + M_{ij,t}}{GDP_{i,t}} \quad (3)$$

where $Depend_{ij,t}$ is the estimate of dependence of Country i on Country j, $X_{ij,t}$ is the exports from Country i to Country j at the moment of time t, and $M_{ij,t}$ is the imports to Country i from Country j in the moment of time t.

3.1. Estimating the ratio of economic exchange between Ukraine and its selected trade partners

Estimation of $TradeShare_{ij}$ for Ukraine and its selected trade partners – the EU, the Russian Federation, post-Soviet countries, China, the USA, and the group of countries consisting of Brazil and India – allows for the following conclusions (see Table 4, Table 5):

Table 4. Ratios of interdependence between Ukraine and its largest trade partners

TradeShareij						
Date	**Russian Fed.**	**Post-Soviet countries**	**EU**	**USA**	**China**	**BRIC (not incl. Russian Fed.)**
01.01.1996	0.318	0.088	0.173	0.021	0.030	0.007
01.01.1997	0.277	0.092	0.209	0.023	0.042	0.010
01.01.1998	0.236	0.049	0.206	0.026	0.031	0.008
01.01.1999	0.222	0.054	0.198	0.023	0.031	0.011
01.01.2000	0.287	0.097	0.265	0.033	0.030	0.013
01.01.2001	0.254	0.106	0.275	0.027	0.027	0.010
01.01.2002	0.229	0.091	0.286	0.024	0.028	0.011
01.01.2003	0.289	0.099	0.369	0.027	0.033	0.021
01.01.2004	0.318	0.098	0.364	0.037	0.025	0.021
01.01.2005	0.262	0.093	0.290	0.021	0.018	0.020
01.01.2006	0.255	0.114	0.323	0.024	0.016	0.019
01.01.2007	0.285	0.136	0.351	0.024	0.018	0.020
01.01.2008	0.258	0.157	0.347	0.035	0.025	0.022
01.01.2009	0.205	0.062	0.235	0.008	0.015	0.012
01.01.2010	0.322	0.089	0.291	0.023	0.028	0.028
01.01.2011	0.344	0.101	0.308	0.026	0.034	0.030

Table 4. Ratios of interdependence between Ukraine and its largest trade partners (cont.)

TradeShareij						
Date	**Russian Fed.**	**Post-Soviet countries**	**EU**	**USA**	**China**	**BRIC (not incl. Russian Fed.)**
01.01.2012	0.240	0.078	0.231	0.021	0.025	0.024
01.01.2013	0.214	0.065	0.245	0.020	0.059	0.021
01.01.2014	0.166	0.071	0.281	0.019	0.060	0.023
01.01.2015	0.127	0.061	0.291	0.020	0.063	0.024

Table 5. Ratios of interdependence between Ukraine and its largest trade partners

TradeShareij						
Date	**Russian Fed.**	**Post-Soviet countries**	**EU**	**USA**	**China**	**BRIC (not incl. Russian Fed.)**
01.01.1996	0.318	0.088	0.173	0.021	0.030	0.007
01.01.1997	0.277	0.092	0.209	0.023	0.042	0.010
01.01.1998	0.236	0.049	0.206	0.026	0.031	0.008
01.01.1999	0.222	0.054	0.198	0.023	0.031	0.011
01.01.2000	0.287	0.097	0.265	0.033	0.030	0.013
01.01.2001	0.254	0.106	0.275	0.027	0.027	0.010
01.01.2002	0.229	0.091	0.286	0.024	0.028	0.011
01.01.2003	0.289	0.099	0.369	0.027	0.033	0.021
01.01.2004	0.318	0.098	0.364	0.037	0.025	0.021
01.01.2005	0.262	0.093	0.290	0.021	0.018	0.020
01.01.2006	0.255	0.114	0.323	0.024	0.016	0.019
01.01.2007	0.285	0.136	0.351	0.024	0.018	0.020
01.01.2008	0.258	0.157	0.347	0.035	0.025	0.022
01.01.2009	0.205	0.062	0.235	0.008	0.015	0.012
01.01.2010	0.322	0.089	0.291	0.023	0.028	0.028
01.01.2011	0.344	0.101	0.308	0.026	0.034	0.030
01.01.2012	0.240	0.078	0.231	0.021	0.025	0.024
01.01.2013	0.214	0.065	0.245	0.020	0.059	0.021
01.01.2014	0.166	0.071	0.281	0.019	0.060	0.023
01.01.2015	0.127	0.061	0.291	0.020	0.063	0.024

1) Trade relations of Ukraine with the EU and the Russian Federation can be referred to as significant and indicative of the vulnerability of the Ukrainian economy to their dynamics.
2) The relative vulnerability of trade relations between Ukraine and the Russian Federation have been gradually decreasing.
3) The trade interdependence of Ukraine and post-Soviet countries features high volatility and the decreasing vulnerability.
4) Regarding the interrelations of Ukraine and the EU, the period from 2000 to 2008 stands out as one demonstrating most clearly the growing share of the EU in the total exports and imports of Ukraine.
5) Ukrainian-American trade relations do not feature dynamics.

An nnalysis of data for 2015 shows the continuingly decreasing trade dependence of Ukraine on the Russian Federation due to the sanctions (0.127 in 2015; 0.214 in 2013, against 0.318 in 1996); in parallel, estimates of trade dependence for Ukraine in the posts-crisis year of 2009 marking the shrinking global demand show that markets in post-Soviet countries could adapt to the consumption of Ukrainian products.

Beginning with 2012, the dependence of Ukraine on the Russian Federation and post-Soviet countries was notably decreasing, contrary to the markets of the EU and China, which, given the high volatility (turning points of growths and recessions), could retain stability. In parallel, the decreasing dependence of Ukraine on the main trade partners in 2012–2015 is an indication of the growing relative weight of the third countries, which did not have a substantial role in Ukraine's foreign trade. It is true that Egypt or Turkey, whose figures of trade with Ukraine are beyond the scope of our analysis, could increase their shares in the foreign trade with Ukraine beginning with 2014.

A remarkable long-term tendency in Ukrainian exports is the falling share of the Commonwealth of Independent States (CIS) beginning with 2011 (from 38.27% in 2012 to 16.62% in 2016) in parallel with the increasing share of EU-28, Asian and African countries. This reorientation is caused by the aggravation of trade and political contradictions between Ukraine and Russia, and the need to seek for new export markets (see Table 6).

Table 6. Geographic structure of Ukrainian exports of goods in 2005–2016, %

Year	CIS	Europe	EU-28	Asia	Africa	America	Australia and Oceania
2005	30.77	31.79	30.07	25.06	6.99	5.35	0.04
2006	32.19	32.91	31.71	22.01	6.19	6.65	0.05
2007	36.69	29.97	28.44	22.07	5.66	5.45	0.03
2008	34.59	29.47	27.28	23.72	5.83	6.19	0.10
2009	33.94	25.86	23.97	30.56	6.62	2.83	0.05
2010	36.46	26.90	25.46	26.68	5.87	3.89	0.06
2011	38.27	26.96	26.35	25.93	4.89	3.73	0.04
2012	36.78	25.31	24.88	25.69	8.19	3.79	0.07

Table 6. Geographic structure of Ukrainian exports of goods in 2005–2016, % (cont.)

Year	CIS	Europe	EU-28	Asia	Africa	America	Australia and Oceania
2013	34.87	26.95	26.47	26.55	8.05	3.42	0.06
2014	27.61	31.77	31.54	28.48	9.46	2.55	0.04
2015	20.47	34.75	34.14	32.47	9.98	2.06	0.04
2016	16.62	37.55	37.11	33.34	10.21	2.24	0.04

3.2. Estimating trade dependence for Ukraine and its selected trade partners

Estimation of $Depend_{ij,t}$ for Ukraine and its selected trade partners – the EU, the Russian Federation, post-Soviet countries, China, the USA, and the group of countries consisting of Brazil, China and India – allows for the following conclusions (see Table 7):

- the dependence of Ukraine's GDP growth on trade relations of the Russian Federation decreased; interrelations between Ukraine and the Russian Federation have five explicit phases of economic activity, correlating closely with the political climate in Ukraine.
- the contribution of post-Soviet countries in Ukraine's GDP growth rapidly decreased (dependence ratio 0.114 as of 1 January 2008, against 0.062 as of 1 January 2016).
- although the impact of trade relations between Ukraine and EU countries on growth of Ukraine's GDP features relative stability (dependence ratio 0.319 for 2003; 0.254 for 2006; 0.259 for 2011), in 2015 EU countries (with dependence ratio of 0.295 recorded for the second time after 2000, the year when the significance of trade relations with this group of countries was dominant for Ukraine's GDP dynamics) became the trade partner for Ukraine with the most essential impact on the dynamics of Ukraine's GDP. However, given that the indicators of dependence of Ukrainian trade on EU countries are analysed considering the waves of EU enlargement (enlarging significantly the number of EU members), the change in Ukraine-EU relations is not explicit.
- given that China joined the top three trade partners of Ukraine by the results of 2016, its impact on the dynamics of Ukraine's GDP gives evidence of gradual transformations in China-Ukraine relations: its nearly zero impact on Ukraine's GDP at early phases of Ukraine's state building (0.019 dependence ratio as of 1 January 1996) was gradually increasing to catch up with the dependence estimates for the group of post-Soviet countries, for which the significance of trade was rapidly falling (dependence ratio 0.064 for China and 0.062 for the group of post-Soviet countries as of 1 January 2016, against 0.026 for China and 0.090 for the group of post-Soviet countries as of 1 January 2006).

Table 7. Ratios of dependence of Ukrainian GDP growth on trade relations with Ukraine's partners

	$Depend_{ij,t}$					
	Russian Fed.	**Post-Soviet countries**	**EU**	**USA**	**China**	**Brazil, China, India**
01.01.1996	0.312	0.087	0.170	0.021	0.019	0.007
01.01.1997	0.223	0.074	0.169	0.018	0.024	0.008
01.01.1998	0.230	0.048	0.201	0.025	0.020	0.008
01.01.1999	0.244	0.060	0.218	0.026	0.026	0.012
01.01.2000	0.289	0.097	0.266	0.034	0.042	0.013
01.01.2001	0.241	0.101	0.262	0.026	0.019	0.010
01.01.2002	0.216	0.086	0.271	0.023	0.022	0.011
01.01.2003	0.249	0.086	0.319	0.023	0.029	0.018
01.01.2004	0.268	0.083	0.307	0.031	0.023	0.018
01.01.2005	0.228	0.081	0.252	0.019	0.028	0.017
01.01.2006	0.201	0.090	0.254	0.019	0.026	0.015
01.01.2007	0.198	0.094	0.244	0.017	0.025	0.014
01.01.2008	0.187	0.114	0.251	0.025	0.033	0.016
01.01.2009	0.179	0.094	0.205	0.013	0.034	0.018
01.01.2010	0.252	0.069	0.228	0.018	0.043	0.022
01.01.2011	0.289	0.085	0.259	0.022	0.050	0.025
01.01.2012	0.247	0.081	0.237	0.021	0.053	0.025
01.01.2013	0.201	0.061	0.230	0.019	0.056	0.019
01.01.2014	0.171	0.073	0.289	0.020	0.061	0.023
01.01.2015	0.128	0.062	0.295	0.020	0.064	0.024

Given the strong impact from the USA on shaping the geopolitical vector of Ukraine's development, the existing trade relations between the two countries indicate unchanged positions (dependence ratio 0.021 as of 1996 and 0.020 as of 2015).

Yet, the estimates of $TradeShare_{ij}$ *and* $Depend_{ij,t}$ demonstrate the quality of economic exchange between Ukraine and its partners in a more representative way, which allows for the following statements:

- while the impact of Ukraine's foreign trade with the EU on Ukraine's GDP changed from negative (-0.132 in 2012) to positive (0.204), in the case of foreign trade with the Russian Federation (-0.359) and the USA (-0.447) the situation is too bad.
- the impact of Ukraine's foreign economic relations with developing countries (Brazil, India, China) on Ukraine's GDP growth is positive (0.534);
- high estimates of dependence show insufficient structural diversification of the Ukrainian economy, disregard to the need for the import substitution policy implementation, which would change commodity positions of Ukrainian exports and imports.

In the case of Ukraine (given its economic dependence on the EU (0.295 as of the end of 2015) and the Russian Federation (0.128 as of the end of 2015, against 0.217 as of the end of 2013)), the estimates give evidence of the skewed trade structure and orientation towards the group of selected partners.

4. Analysis and discussion of results

We have built the equation of regression for the total exports and imports by country, which looks representative.

$$Y_1 = -2.758\mathrm{X}_6 + 4{,}922\mathrm{X}_9 + 2{,}042\mathrm{X}_{12} + 16{,}524\mathrm{X}_{15} - 35{,}289\mathrm{X}_{18} + 23{,}273\mathrm{X}_{21} + 54687205 \quad (4)$$

coefficient of determination $R^2 = 0.940$;

where X_6 – total imports and exports with the Russian Federation;

X_9 – total imports and exports with post-Soviet countries (not including the Russian Federation);

X_{12} – total imports and exports with EU countries;

X_{15} – total imports and exports with China;

X_{18} – total imports and exports with the USA;

X_{21} – total imports and exports with Brazil, China and India.

The results lead to the following conclusions:

- Ukraine's dependence on foreign economic relations with the Russian Federation has a negative impact on the growth rates of Ukraine's GDP (growth in the trade relations by 1000 UAH reduces the GDP by 2758 UAH);
- Ukraine's dependence on foreign economic relations with the USA has an extremely negative impact on the growth rates of Ukraine's GDP (growth in the trade relations by 1000 UAH reduces the GDP by 35289 UAH);
- Ukraine's dependence on foreign economic relations with China has a positive impact on the growth rates of Ukraine's GDP (growth in the trade relations by 1000 UAH increases the GDP by 16524 UAH);
- Ukraine's dependence on foreign economic relations with post-Soviet countries has positive impact on the growth rates of Ukraine's GDP (growth in the trade relations by 1000 UAH increases the GDP by 4922 UAH);
- Ukraine's dependence on foreign economic relations with EU countries has a positive impact on the growth rates of Ukraine's GDP (growth in the trade relations by 1000 UAH increases the GDP by 2042 UAH);

- Trade leaders with a positive impact on the growth rates of Ukraine's GDP (growth worth of 23273 UAH per each 1000 UAH) are Brazil China, and India.

5. Summary and conclusion

The proposed methodology for computing indexes of dependence and interdependence, measures of symmetry, sensitivity and vulnerability of relations between partner countries can be useful for the analysis of established relations, to reveal the comparative dynamics of change in partner countries with different economic capacities, and in countries with similar economic structures. It should be borne in mind, however, that once a partner country pursues imports substitution policy or, say, reshoring, which changes its economic structure and, consequently, the structure of its demand for goods at the global market, this can have a tangible effect on the quality of established relations that will undergo gradual transformations: when imports substitution policy is adopted by a country that is an outsider in relations, the asymmetries will be decreasing; when reshoring policy is adopted by countries that are leaders of relations, the explicit asymmetries will be aggravating.

The dynamics of countries' interdependence is conditional not only on endogenous factors (structure of economy, structure of demand, macroeconomic stability in a country), but also exogenous ones (rate of the global economy growth, conjuncture at global commodity markets, conditions for access to capital markets and intellectual property markets, etc.). Thus, if GDP of a partner country grows significantly, the unchanged figures of its trade relations with selected countries cannot be evidence of these relations' decline.

The reorientation of Ukraine's trade flows from CIS to the EU and Asia, confirmed by the assessment, is a long-term trend that has been strengthened as a consequence of recent events and Ukraine and beyond. Ukraine has leading positions at the markets of agricultural goods, ores and metal, i.e. the so called "stock exchange" goods with prices very sensitive to global conjuncture fluctuations. Once the share of goods with high value added is increased, export earnings will be more stable. Ukrainian exports are concentrated; because this also increases their sensitivity to shocks, their scope and price can be to a significant extend volatile. The indexes of dependence, derived for Ukraine, show that Ukraine has sensitivity-based interdependence relations with its trade partners, except for Russia, with which Ukraine has interdependence associated with vulnerability, because it refers to deliberate *disruption* of the existing relations and minimization of Russia's role as exporter and importer.

Considering the already existing economic capacities and sectoral structure of Ukraine, it needs to be noted that Ukraine faces objective challenges on the way to integration in the global market that has undergone powerful globalization processes involved in coordination of interests by entities participating in international value added chains. We believe that Ukraine needs to act in a multi-vector way, in the five mainstream directions (technological, financial, infrastructural, structural and diplomatic), to optimize its foreign economic relations. The export pattern of Ukraine, based on a significant share of primary commodities, confirms its low productivity and non-competitiveness at the global

market. A comprehensive analysis of trade relations of Ukraine by computing RVA index and dependence index confirms that the less diversified the economic structure of a country is and the more similar the economic structures of partner countries are, the more stable their relations are. Accordingly, if even the trade between such partner countries diminishes, the quality of their relations will remain unchanged, with the implicit asymmetric or symmetric dependence. Moreover, the change will be mutual if even asymmetric relations are preserved.

The interdependence of the partner countries' economies can be caused by the symmetrically growing demand for goods that they offer if even technological gaps between them are preserved. Thus, the increasing imports of technologies by one of the partner countries can be symmetrically accompanied by the increasing exports of its primary commodities to the partner country's market. Therefore, it would be too difficult to substantiate the quality of such trade relations without a detailed study of the structure of commodity exports and imports.

It can be concluded that the decreasing interdependence of partner countries, in parallel with establishing more diversified trade relations and/or reorientation to production of alternative goods/services with the respective growth in exports is a sign of economic development of a country.

REFERENCES

BEN-DAVID, D., (1993). Equalizing exchange: trade liberalization and income convergence, Quarterly Journal of Economics, Vol. 108, No. 3, pp. 653–679.

COOPER, R., KENEN, P. B., JONES, R. W., (1985). Economic Interdependence and Coordination of Economic Policies, http://econpapers.repec.org/bookchap/eeeintchp/2-23.htm.

DOLLAR, D., (1992). Outward-oriented developing economies really do grow more rapidly: evidence from 95 LDCs, 1976-85, Economic Development and Cultural Change, pp. 523–544.

EDWARDS, S., (1993). Openness, trade liberalization, and growth in developing countries, Journal of Economic Literature, Vol. 31, No 3, pp. 1358–1393.

EDWARDS, S., (1998). Openness, productivity and growth: what do we really know? Economic Journal, Vol. 108, pp. 383–398.

HARRISON, A., (1991). Openness and growth: a time-series, cross-country analysis for developing countries, NBER Working Paper, No. WPS 809, pp. 6–45.

HATEMI, J. A., IRANDOUST, M., (2001). Productivity performance and export performance: a time series perspective, Eastern Economic Journal, Vol. 27, No. 2, pp. 149–164

INTERNATIONAL MONETARY FUND, (1997). Annual report of the Executive Board for the financial year, http://www.imf.org/external/pubs/ft/ar/97/pdf/file06.pdf.

KINDLEBERGER, C. P., (1962). Foreign trade and the national economy, New Heaven; Yale University Press.

KRUEGER, A. O., (1988). Why trade liberalization is good for growth, The Economic Journal, Vol. 88, pp.1513–1522.

KUNST, R.M., MARIN, D., (1989). On exports and productivity: a causal analysis, Review of Economics and Statistics, Vol. 71, No. 4, pp. 699–703.

ROMER, P., (1986). Increasing returns and long–run growth, Journal of Political Economy, Vol. 94, pp. 1002–1038.

SACHS, J., WARNER, A., (1995). Economic reform and the process of global integration, Brookings Papers on Economic Activity, Vol. 1, pp. 1–118.

STIGLITZ, J. E., (1998). Towards a new paradigm for development: strategies, policies, and processes, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.199.9708&rep=rep1&type=pdf.

WACZIARG, R., (2001). Measuring the dynamic gains from trade, World Bank Economic Review, Vol. 15, No. 3, pp. 393–429.

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 135–148, DOI 10.21307/stattrans-2018-008

POWER ISHITA DISTRIBUTION AND ITS APPLICATION TO MODEL LIFETIME DATA

Kamlesh Kumar Shukla [1], **Rama Shanker** [2]

ABSTRACT

A study on two-parameter power Ishita distribution (PID), of which Ishita distribution introduced by Shanker and Shukla (2017 a) is a special case, has been carried out and its important statistical properties including shapes of the density, moments, skewness and kurtosis measures, hazard rate function, and stochastic ordering have been discussed. The maximum likelihood estimation has been discussed for estimating its parameters. An application of the distribution has been explained with a real lifetime data from engineering, and its goodness of fit shows better fit over two-parameter power Akash distribution (PAD), two-parameter power Lindley distribution (PLD) and one-parameter Ishita, Akash, Lindley and exponential distributions.

Key words: Ishita distribution, moments, hazard rate function, stochastic ordering, maximum likelihood estimation, goodness of fit.

1. Introduction

The probability density function (pdf) of Ishita distribution introduced by Shanker and Shukla (2017 a) is given by

$$f_1(y;\theta) = \frac{\theta^3}{\theta^3+2}\left(\theta + y^2\right)e^{-\theta y} \quad ; y > 0,\ \theta > 0 \tag{1.1}$$

$$= p\, g_1(y;\theta) + (1-p)\, g_2(y;\theta) \tag{1.2}$$

where

$$p = \frac{\theta^3}{\theta^3+2}$$

$$g_1(y;\theta) = \theta e^{-\theta y} \ ; y > 0, \theta > 0$$

[1] First author: Department of Statistics, College of Science, Eritrea Institute of Technology, Asmara, Eritrea. E-mail: kkshukla22@gmail.com.
[2] Corresponding author: Department of Statistics, College of Science, Eritrea Institute of Technology, Asmara, Eritrea. E-mail: shankerrama2009@gmail.com.

$$g_2(y;\theta)=\frac{\theta^3}{\Gamma(3)}e^{-\theta y}y^{3-1}\ ;y>0,\theta>0$$

The pdf in (1.1) reveals that the Ishita distribution is a two-component mixture of an exponential distribution (with scale parameter θ) and a gamma distribution (with shape parameter 2 and scale parameter θ), with mixing proportion $p=\frac{\theta^3}{\theta^3+2}$. Shanker and Shukla (2017 a) have discussed some of its mathematical and statistical properties including its shapes for varying values of the parameter, moments, skewness, kurtosis, hazard rate function, mean residual life function, stochastic ordering, mean deviations, order statistics, Bonferroni and Lorenz curves, Renyi entropy measure, stress-strength reliability, and the applications of the distribution for modelling lifetime data from engineering and medical science. However, there are some situations where the Ishita distribution may not be suitable from either theoretical or applied point of view. Shukla and Shanker (2017) have also obtained a Poisson mixture of Ishita distribution and named it Poisson-Ishita distribution, and studied its various statistical properties, estimation of parameter and the goodness of fit with some real count data sets.

The corresponding cumulative distribution function (cdf) of (1.1) is given by

$$F_1(y;\theta)=1-\left[1+\frac{\theta y(\theta y+2)}{\theta^3+2}\right]e^{-\theta y};\ y>0,\theta>0 \tag{1.3}$$

Recall that the pdf and the cdf of two-parameter power Akash distribution (PAD) introduced by Shanker and Shukla (2017 b) and two-parameter power Lindley distribution (PLD) introduced by Ghitany *et al.* (2013) are respectively given by

$$f_2(x;\theta,\alpha)=\frac{\alpha\theta^3}{\theta^2+2}\left(1+x^{2\alpha}\right)x^{\alpha-1}e^{-\theta x^{\alpha}}\ ;x>0,\theta>0,\alpha>0 \tag{1.4}$$

$$F_2(x;\theta,\alpha)=1-\left[1+\frac{\theta x^{\alpha}\left(\theta x^{\alpha}+2\right)}{\theta^2+2}\right]e^{-\theta x^{\alpha}};\ x>0,\theta>0,\alpha>0 \tag{1.5}$$

$$f_3(x;\theta,\alpha)=\frac{\alpha\theta^2}{\theta+1}\left(1+x^{\alpha}\right)x^{\alpha-1}e^{-\theta x^{\alpha}}\ ;x>0,\theta>0,\alpha>0 \tag{1.6}$$

$$F_3(x;\theta,\alpha)=1-\left[1+\frac{\theta x^{\alpha}}{\theta+1}\right]e^{-\theta x^{\alpha}}\ ;x>0,\theta>0,\alpha>0 \tag{1.7}$$

A detailed study regarding various properties, estimation of parameters and applications of PAD and PLD can be seen from Shanker and Shukla (2017 b) and

Ghitany *et al.* (2013) respectively. At $\alpha = 1$, PAD reduces to Akash distribution introduced by Shanker (2015) having pdf and cdf given by

$$f_4(x;\theta) = \frac{\theta^3}{\theta^2+2}\left(1+x^2\right)e^{-\theta x}\ ; x>0, \theta>0 \tag{1.8}$$

$$F_4(x;\theta) = 1-\left[1+\frac{\theta x(\theta x+2)}{\theta^2+2}\right]e^{-\theta x}; x>0, \theta>0 \tag{1.9}$$

Shanker (2015) has a detailed study about various statistical and mathematical properties of Akash distribution, estimation of parameter and applications for modelling lifetime data from engineering and medical science and showed that Akash distribution gives better fit than both exponential and Lindley distributions. Shanker (2017) has also obtained a Poisson mixture of Akash distribution and named Poisson-Akash distribution and discussed important statistical properties, estimation of parameter using both the method of moments and the method of maximum likelihood and the application for modelling count data.

Similarly, at $\alpha = 1$, PLD reduces to Lindley distribution introduced by Lindley (1958) having pdf and cdf given by

$$f_5(x;\theta) = \frac{\theta^2}{\theta+1}(1+x)e^{-\theta x}; x>0, \theta>0 \tag{1.10}$$

$$F_5(x;\theta) = 1-\left[1+\frac{\theta x}{\theta+1}\right]e^{-\theta x}; x>0, \theta>0 \tag{1.11}$$

Ghitany *et al*. (2008) have a detailed study about various properties of Lindley distribution, estimation of parameter and application for modelling waiting time data from a bank and it has been shown that it gives better fit than exponential distribution. Shanker *et al.* (2016) have a detailed and critical comparative study of modelling real lifetime data from engineering and biomedical sciences using Akash, Lindley and exponential distribution and observed that each of these one-parameter distribution has some advantage over the other but none is perfect for modelling all real lifetime data. Since Ishita distribution gives better fit than Akash, Lindley and exponential distribution, it is expected and hoped that the two-parameter power Ishita distribution (PID) will provide a better model over two-parameter power Akash distribution (PAD) and power Lindley distribution (PLD) and one-parameter Ishita, Akash, Lindley and exponential distributions.

In this paper, a two-parameter power Ishita distribution (PID), which includes one-parameter Ishita distribution, has been introduced and its various properties including shapes for varying values of the parameters, survival function, hazard rate function, moments, stochastic ordering have been studied. The maximum likelihood estimation has been discussed for estimating its parameters. Finally, applications and goodness of fit of PID has been illustrated with a real life time data and fit has been found better over two-parameter power Akash distribution

(PAD) of Shanker and Shukla (2017 b), two-parameter power Lindley distribution (PLD) of Ghitany *et al.* (2013), and one-parameter Ishita, Akash, Lindley and exponential distributions.

2. Power Ishita distribution

Taking the power transformation $X = Y^{1/\alpha}$ in (1.1), pdf of the random variable X can be obtained as

$$f_6(x;\theta,\alpha) = \frac{\alpha\theta^3}{\theta^3+2}\left(\theta + x^{2\alpha}\right)x^{\alpha-1}e^{-\theta x^{\alpha}} \; ; x>0, \theta>0, \alpha>0 \tag{2.1}$$

$$= p\, g_3(x;\theta,\alpha) + (1-p)\, g_4(x;\theta,\alpha) \tag{2.2}$$

where $p = \dfrac{\theta^3}{\theta^3+2}$

$$g_3(x;\theta,\alpha) = \alpha\theta x^{\alpha-1}e^{-\theta x^{\alpha}} \; ; x>0, \alpha>0, \theta>0$$

$$g_4(x;\theta,\alpha) = \frac{\alpha\theta^3 x^{3\alpha-1}e^{-\theta x^{\alpha}}}{2}; x>0, \alpha>0, \theta>0$$

We would call the density in (2.1) "Power Ishita distribution (PID)" and denote it as PID(θ,α). It is obvious that the PID is also a two-component mixture of Weibull distribution (with shape parameter α and scale parameter θ), and a generalized gamma distribution (with shape parameters 3, α and scale parameter θ) introduced by Stacy (1962) with their mixing proportion $p = \dfrac{\theta^3}{\theta^3+2}$. The corresponding cumulative distribution function (cdf) of (2.1) can be obtained as

$$F_6(x;\theta,\alpha) = 1 - \left[1 + \frac{\theta x^{\alpha}\left(\theta x^{\alpha}+2\right)}{\theta^3+2}\right]e^{-\theta x^{\alpha}} \; ; x>0, \theta>0, \alpha>0 \tag{2.3}$$

Graphs of the pdf and the cdf of PID for varying values of the parameters have been drawn and presented in Figures 1 and 2 respectively. If $\alpha = 1$, the pdf of PID is monotonically decreasing for increasing values of the parameter θ. But for $\alpha > 1$ and increasing values of the parameter θ, the shapes of the pdf of PID become negatively skewed, positively skewed, symmetrical, platykurtic and mesokurtic; and this means that PID can be used for modelling lifetime data of various nature.

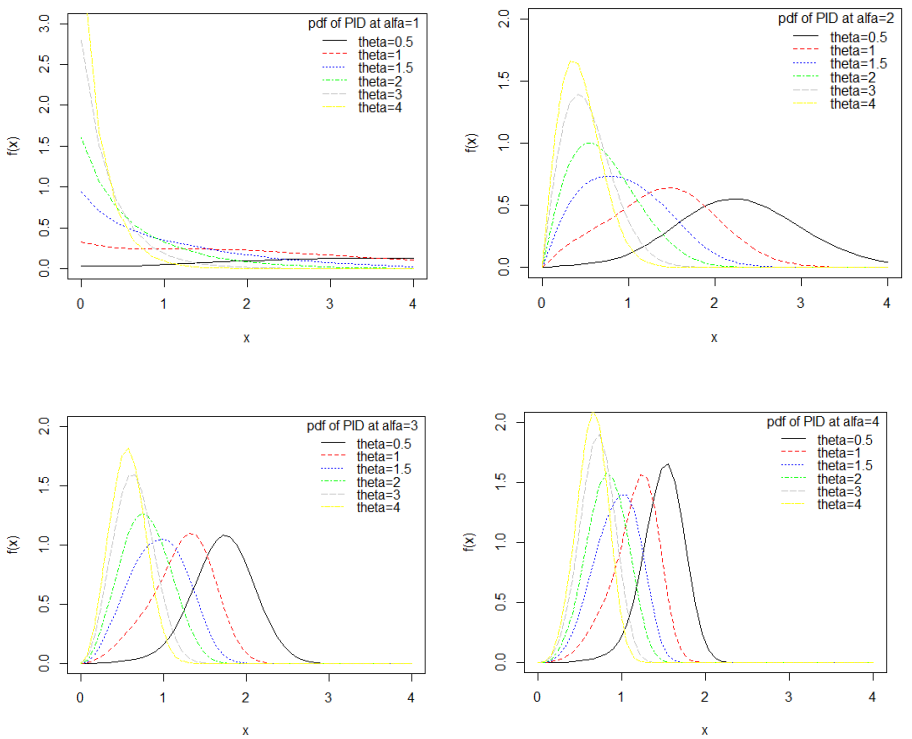
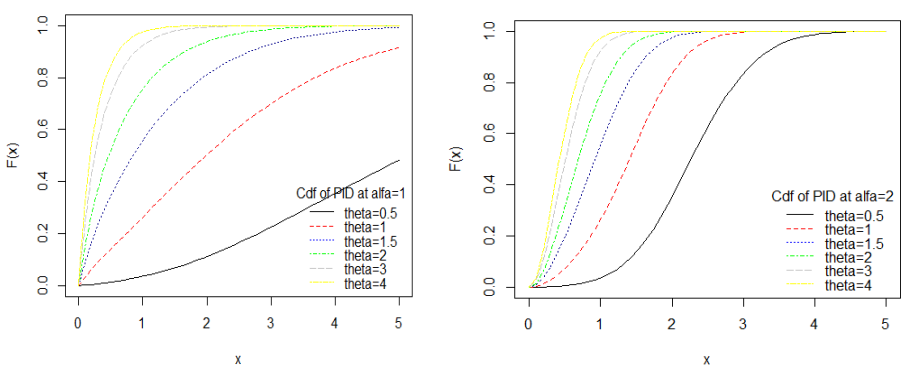


Figure.1. Graphs of pdf of PID for varying values of parameters θ and α



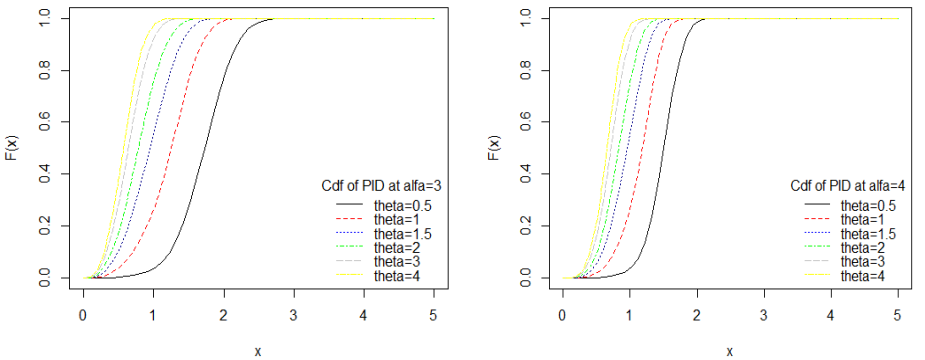


Figure 2. Graphs of cdf of PID for varying values of parameters θ and α

3. Survival and hazard rate functions

The survival function, $S(x)$ and hazard rate function, $h(x)$ of the PID can be obtained as

$$S(x;\theta,\alpha)=1-F_6(x;\theta,\alpha)=\left[\frac{\theta x^{\alpha}\left(\theta x^{\alpha}+2\right)+\left(\theta^3+2\right)}{\theta^3+2}\right]e^{-\theta x^{\alpha}};\ x>0,\theta>0,\alpha>0 \tag{3.1}$$

$$h(x;\theta,\alpha)=\frac{f_6(x;\theta,\alpha)}{S(x;\theta,\alpha)}=\frac{\alpha\theta^3\left(1+x^{2\alpha}\right)x^{\alpha-1}}{\theta x^{\alpha}\left(\theta x^{\alpha}+2\right)+\left(\theta^3+2\right)};\quad x>0,\ \theta>0,\ \alpha>0 \tag{3.2}$$

The nature and behaviour of $h(x)$ of the PID for varying values of the parameters θ and α are shown graphically in Figure 3. It is obvious from the graphs of $h(x)$ that it is monotonically decreasing and increasing for increased values of the parameters θ and α.

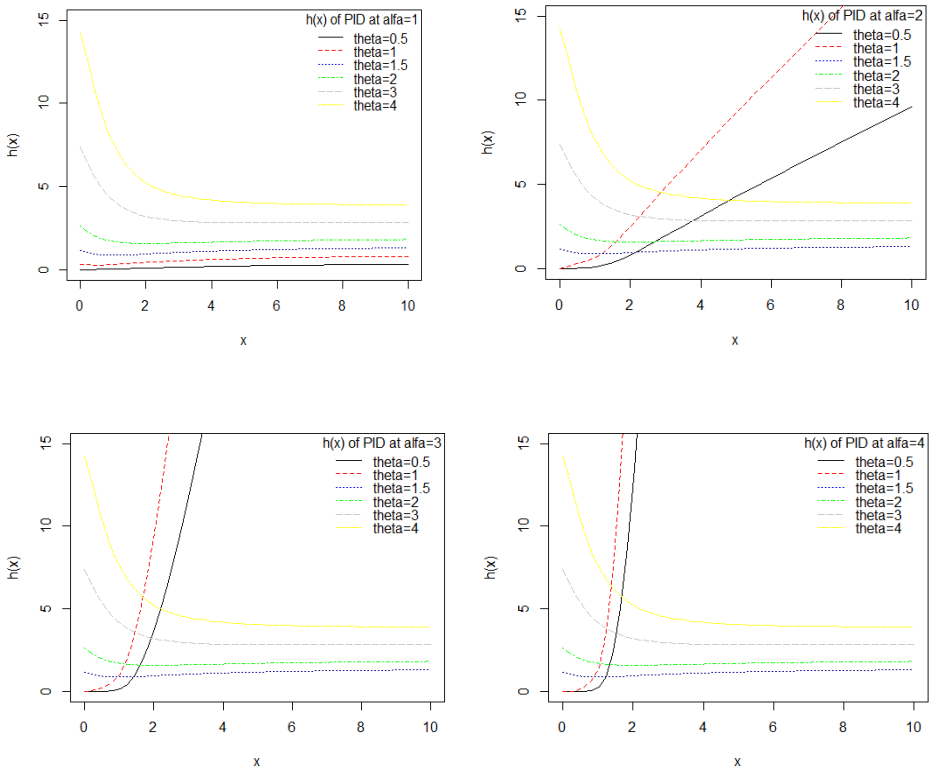


Figure 3. Graphs of $h(x)$ of PID for varying values of the parameters θ and α

4. Moments and related measures

Using the mixture representation (2.2), the r th moment about origin of the PID can be obtained as

$$\mu_r' = E\left(X^r\right) = p\int_0^\infty x^r g_3(x;\theta,\alpha)\,dx + (1-p)\int_0^\infty x^r g_4(x;\theta,\alpha)\,dx$$

$$= \frac{r\Gamma\left(\frac{r}{\alpha}\right)\left[\alpha^2\theta^3 + (r+\alpha)(r+2\alpha)\right]}{\alpha^3\theta^{r/\alpha}\left(\theta^3+2\right)}; r = 1,2,3,.... \qquad (4.1)$$

It should be noted that at $\alpha = 1$, the above expression will reduce to the r th moment about origin of Ishita distribution and is given by

$$\mu_r' = \frac{r!\left[\theta^3 + (r+1)(r+2)\right]}{\theta^r\left(\theta^3+2\right)}; r = 1,2,3,...$$

Therefore, the mean and the variance of the PID are obtained as

$$\mu_1' = \frac{\Gamma\left(\frac{1}{\alpha}\right)\left[\alpha^2\theta^3 + (\alpha+1)(2\alpha+1)\right]}{\alpha^3\theta^{1/\alpha}\left(\theta^3+2\right)}$$

$$\sigma^2 = \frac{2\Gamma\left(\frac{2}{\alpha}\right)\left[\alpha^2\theta^3 + 2(\alpha+1)(\alpha+2)\right]\alpha^3\left(\theta^3+2\right) - \left(\Gamma\left(\frac{1}{\alpha}\right)\right)^2\left[\alpha^2\theta^3 + (\alpha+1)(2\alpha+1)\right]^2}{\alpha^6\theta^{2/\alpha}\left(\theta^3+2\right)^2}$$

The coefficient of skewness and the coefficient of kurtosis of PID, upon substituting for the raw moments and standard deviation (σ), can be obtained using following expressions

$$\text{Coefficient of Skewness} = \frac{\mu_3' - 3\mu_2'\mu_1' + 2\left(\mu_1'\right)^3}{\sigma^3}$$

$$\text{and Coefficient of Kurtosis} = \frac{\mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\left(\mu_1'\right)^2 - 3\left(\mu_1'\right)^4}{\sigma^4}.$$

5. Stochastic ordering

Stochastic ordering of positive continuous random variables is an important tool for judging their comparative behaviour. A random variable X is said to be smaller than a random variable Y in the

(i) stochastic order $\left(X \le_{st} Y\right)$ if $F_X(x) \ge F_Y(x)$ for all x

(ii) hazard rate order $\left(X \le_{hr} Y\right)$ if $h_X(x) \ge h_Y(x)$ for all x

(iii) mean residual life order $\left(X \le_{mrl} Y\right)$ if $m_X(x) \le m_Y(x)$ for all x

(iv) likelihood ratio order $\left(X \le_{lr} Y\right)$ if $\frac{f_X(x)}{f_Y(x)}$ decreases in x.

The following important interrelationships due to Shaked and Shanthikumar (1994) are well known for establishing stochastic ordering of distributions

$$X \leq_{lr} Y \Rightarrow X \leq_{hr} Y \Rightarrow X \leq_{mrl} Y$$

$$\Downarrow$$

$$X \leq_{st} Y$$

The PID is ordered with respect to the strongest 'likelihood ratio ordering' as shown in the following theorem:

Theorem: Let $X \sim \text{PID}(\theta_1, \alpha_1)$ and $Y \sim \text{PID}(\theta_2, \alpha_2)$. If $\theta_1 > \theta_2$ and $\alpha_1 = \alpha_2$ (or $\alpha_1 < \alpha_2$ and $\theta_1 = \theta_2$) then $X \leq_{lr} Y$ and hence $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$.

Proof: From the pdf of PID (2.1), we have

$$\frac{f_X(x,\theta_1,\alpha_1)}{f_Y(x;\theta_2,\alpha_2)} = \left(\frac{\alpha_1 \theta_1^3 \left(\theta_2^3 + 2\right)}{\alpha_2 \theta_2^3 \left(\theta_1^3 + 2\right)}\right)\left(\frac{\theta_1 + x^{2\alpha_1}}{\theta_2 + x^{2\alpha_2}}\right) x^{\alpha_1 - \alpha_2} e^{-\left(\theta_1 x^{\alpha_1} - \theta_2 x^{\alpha_2}\right)} \quad ; \; x > 0$$

Now

$$\ln \frac{f_X(x,\theta_1,\alpha_1)}{f_Y(x;\theta_2,\alpha_2)} = \ln\left(\frac{\alpha_1 \theta_1^3 \left(\theta_2^3 + 2\right)}{\alpha_2 \theta_2^3 \left(\theta_1^3 + 2\right)}\right) + \ln\left(\frac{\theta_1 + x^{2\alpha_1}}{\theta_2 + x^{2\alpha_2}}\right) + \left(\alpha_1 - \alpha_2\right) \ln x - \left(\theta_1 x^{\alpha_1} - \theta_2 x^{\alpha_2}\right)$$

This gives

$$\frac{d}{dx}\left\{\ln \frac{f_X(x,\theta_1,\alpha_1)}{f_Y(x;\theta_2,\alpha_2)}\right\} = \frac{2\left(\alpha_1 \theta_2 x^{2\alpha_1 - 1} - \alpha_2 \theta_1 x^{2\alpha_2 - 1}\right) + 2\left(\alpha_1 - \alpha_2\right) x^{2(\alpha_1 + \alpha_2) - 1}}{\left(\theta_1 + x^{2\alpha_1}\right)\left(\theta_2 + x^{2\alpha_2}\right)} + \frac{\alpha_1 - \alpha_2}{x}$$

$$- \left(\alpha_1 \theta_1 x^{\alpha_1 - 1} - \alpha_2 \theta_2 x^{\alpha_2 - 1}\right)$$

Clearly for $\theta_1 > \theta_2$ and $\alpha_1 = \alpha_2$ (or $\alpha_1 < \alpha_2$ and $\theta_1 = \theta_2$), $\frac{d}{dx}\left\{\ln \frac{f_X(x,\theta_1,\alpha_1)}{f_Y(x;\theta_2,\alpha_2)}\right\} < 0$.

This means that $X \leq_{lr} Y$ and hence $X \leq_{hr} Y$, $X \leq_{mrl} Y$ and $X \leq_{st} Y$. Thus PID follows the strongest likelihood ratio ordering.

6. Maximum likelihood estimation

Let $\left(x_1, x_2, x_3, \ldots, x_n\right)$ be a random sample of size n from $\text{PID}(\theta, \alpha)$. Then, the log-likelihood function is given by

$$\ln L = \sum_{i=1}^{n} \ln f_6\left(x_i; \theta, \alpha\right)$$

$$= n\left[\ln\alpha + 3\ln\theta - \ln\left(\theta^3+2\right)\right] + \sum_{i=1}^{n}\ln\left(\theta + x_i^{2\alpha}\right) + (\alpha-1)\sum_{i=1}^{n}\ln\left(x_i\right) - \theta\sum_{i=1}^{n}x_i^{\alpha}.$$

The maximum likelihood estimates (MLE) $\left(\hat{\theta},\hat{\alpha}\right)$ of (θ,α) of PID (2.1) are the solutions of the following log likelihood equations

$$\frac{\partial \ln L}{\partial\theta} = \frac{3n}{\theta} - \frac{3n\theta^2}{\theta^3+2} + \sum_{i=1}^{n}\frac{1}{\left(\theta + x_i^{2\alpha}\right)} - \sum_{i=1}^{n}x_i^{\alpha} = 0$$

$$\frac{\partial \ln L}{\partial\alpha} = \frac{n}{\alpha} + 2\sum_{x=1}^{n}\frac{x_i^{2\alpha}\ln\left(x_i\right)}{\left(\theta + x_i^{2\alpha}\right)} + \sum_{i=1}^{n}\ln\left(x_i\right) - \theta\sum_{i=1}^{n}x_i^{\alpha}\ln\left(x_i\right) = 0$$

These two log likelihood equations do not seem to be solved directly because these cannot be expressed in closed form. However, Fisher's scoring method can be applied to solve these equations iteratively. Thus, we have

$$\frac{\partial^2 \ln L}{\partial\theta^2} = -\frac{3n}{\theta^2} + \frac{3n\left(\theta^3-4\right)\theta}{\left(\theta^3+2\right)^2} - \sum_{i=1}^{n}\frac{1}{\left(\theta + x_i^{2\alpha}\right)^2}$$

$$\frac{\partial^2 \ln L}{\partial\theta\,\partial\alpha} = -2\sum_{i=1}^{n}\frac{x_i^{2\alpha}\ln\left(x_i\right)}{\left(\theta + x_i^{2\alpha}\right)^2} - \sum_{i=1}^{n}x_i^{\alpha}\ln\left(x_i\right) = \frac{\partial^2 \ln L}{\partial\alpha\,\partial\theta}$$

$$\frac{\partial^2 \ln L}{\partial\alpha^2} = -\frac{n}{\alpha^2} + 4\theta\sum_{i=1}^{n}\frac{\left(x_i\ln\left(x_i\right)\right)^2}{\left(\theta + x_i^{2\alpha}\right)^2} - \theta\sum_{i=1}^{n}x_i^{\alpha}\left(\ln\left(x_i\right)\right)^2$$

The MLE $\left(\hat{\theta},\hat{\alpha}\right)$ of (θ,α) of PID (2.1) are the solution of the following equations

$$\begin{bmatrix} \dfrac{\partial^2 \ln L}{\partial\theta^2} & \dfrac{\partial^2 \ln L}{\partial\theta\,\partial\alpha} \\ \dfrac{\partial^2 \ln L}{\partial\theta\,\partial\alpha} & \dfrac{\partial^2 \ln L}{\partial\alpha^2} \end{bmatrix}_{\substack{\hat{\theta}=\theta_0 \\ \hat{\alpha}=\alpha_0}} \begin{bmatrix} \hat{\theta}=\theta_0 \\ \hat{\alpha}=\alpha_0 \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \ln L}{\partial\theta} \\ \dfrac{\partial \ln L}{\partial\alpha} \end{bmatrix}_{\substack{\hat{\theta}=\theta_0 \\ \hat{\alpha}=\alpha_0}}$$

where θ_0 and α_0 are initial values of θ and α. These equations are solved iteratively until sufficiently close estimates of $\hat{\theta}$ and $\hat{\alpha}$ are obtained. In this paper, R-software has been used to estimate the parameters θ and α for the considered dataset.

7. Applications and goodness of FIT

In this section, we present the goodness of fit of PID using maximum likelihood estimates of parameters to a real data set from engineering and compare its fit with the one-parameter exponential, Lindley, Akash and Ishita distributions and two-parameter PAD and PLD. The following real lifetime data have been considered for the goodness of fit of the considered distributions.

Data Set: The following data represent the tensile strength, measured in GPa, of 69 carbon fibers tested under tension at gauge lengths of 20mm, Bader and Priest (1982)

1.312 1.314 1.479 1.552 1.700 1.803 1.861 1.865 1.944 1.958 1.966 1.997
2.006 2.021 2.027 2.055 2.063 2.098 2.140 2.179 2.224 2.240 2.253 2.270
2.272 2.274 2.301 2.301 2.359 2.382 2.382 2.426 2.434 2.435 2.478 2.490
2.511 2.514 2.535 2.554 2.566 2.570 2.586 2.629 2.633 2.642 2.648 2.684
2.697 2.726 2.770 2.773 2.800 2.809 2.818 2.821 2.848 2.880 2.954 3.012
3.067 3.084 3.090 3.096 3.128 3.233 3.433 3.585 3.585

In order to compare the considered distributions, values of $-2\ln L$, AIC (Akaike Information Criterion), K-S Statistic (Kolmogorov-Smirnov Statistic) and p-value for the real dataset have been computed using maximum likelihood estimates and presented in Table 1. The formulae for computing AIC and K-S Statistics are as follows:

$AIC = -2\ln L + 2k$ and $\text{K-S} = \text{Sup}\left|F_n(x) - F_0(x)\right|$, where k = the number of parameters, n = the sample size, $F_n(x)$ is the empirical (sample) cumulative distribution function and $F_0(x)$ is the theoretical cumulative distribution function. The best distribution is the distribution corresponding to lower values of $-2\ln L$, AIC, and K-S statistics and higher p-value.

Table 1. MLE's, -2ln L, AIC, K-S and p-value of the fitted distributions of the considered dataset

Model	ML Estimates	-2ln L	AIC	K-S	p-value
PID	$\hat{\theta} = 0.18063$ $\hat{\alpha} = 3.00429$	97.84	101.84	0.033	1.00
PAD	$\hat{\theta} = 0.169$ $\hat{\alpha} = 3.061$	98.02	102.02	0.038	0.999
PLD	$\hat{\theta} = 0.050$ $\hat{\alpha} = 3.868$	98.12	102.12	0.044	0.998
Ishita	$\hat{\theta} = 0.39152$	223.14	225.14	0.331	0.003

Table 1. MLE's, -2ln L, AIC, K-S and p-value of the fitted distributions of the considered dataset (cont.)

Model	ML Estimates	-2ln L	AIC	K-S	p-value
Akash	$\hat{\theta} = 0.964726$	224.28	226.28	0.348	0.001
Lindley	$\hat{\theta} = 0.659000$	238.38	240.38	0.390	0.000
Exponential	$\hat{\theta} = 0.407941$	261.74	263.74	0.434	0.000

It is obvious from the goodness of fit based on K-S statistic that PID gives better fit than all the considered distributions and hence it can be considered an important two-parameter lifetime distribution for modelling lifetime data. The variance-covariance matrix and the 95% confidence intervals (CI's) of the ML estimates of the parameters θ and α of PID are presented in Table 2.

Table 2. Variance-Covariance matrix and 95% confidence intervals (CI's) for the parameters $\hat{\theta}$ and $\hat{\alpha}$ of the considered dataset

Parameters	Variance-Covariance Matrix		95% CI	
	$\hat{\theta}$	$\hat{\alpha}$	Lower	Upper
$\hat{\theta}$	0.002233	-0.0117269	0.105235	0.292116
$\hat{\alpha}$	-0.0117269	0.0662314	2.527340	3.53244

The profile of likelihood estimates of parameters $\hat{\theta}$ and $\hat{\alpha}$ of PID for the considered data set is shown in Figure 4. Also, the fitted plots of the considered dataset for PID are shown in Figure 5.

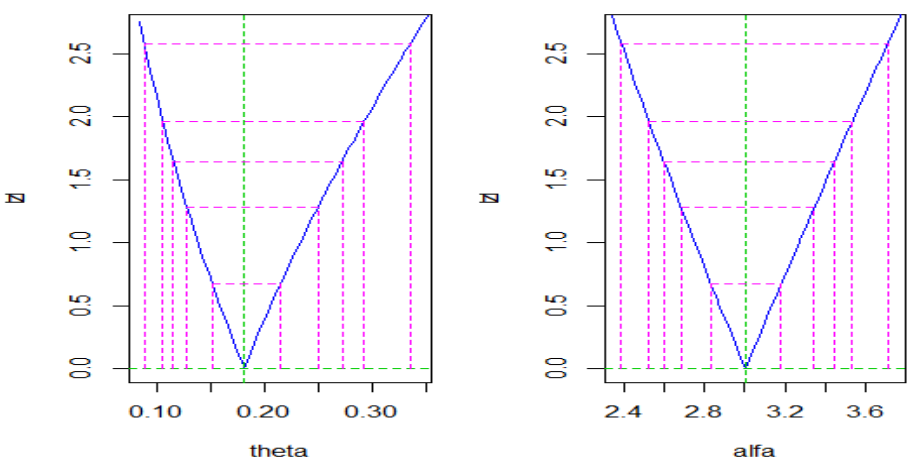


Figure 4. Profile of the likelihood estimates $\hat{\theta}$ and $\hat{\alpha}$ of PID for the considered dataset

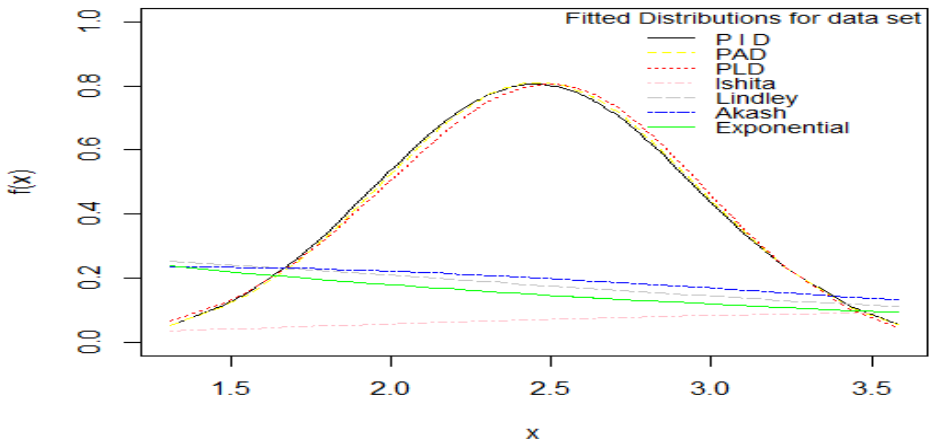


Figure 5. Fitted plots of the considered distributions for the given dataset

8. Concluding remarks

In this paper a two-parameter power Ishita distribution (PID), of which one-parameter Ishita distribution introduced by Shanker and Shukla (2017 a) is a special case, has been introduced and its important statistical properties including shapes of the density, moments, skewness and kurtosis measures and hazard rate function have been discussed. The stochastic ordering of the distribution has been studied. The maximum likelihood estimation has been discussed for estimating its parameters. An application and goodness of fit of PID have been discussed with a real lifetime data set from engineering and the fit has been found quite satisfactory over two-parameter PAD and PLD and one-parameter Ishita, Akash, Lindley and exponential distributions.

Acknowledgment

Authors are grateful to the Editor-In-Chief of the journal and two independent reviewers for their useful comments, which improved the presentation and the quality of the paper.

REFERENCES

BADER, M. G., PRIEST, A. M., (1982). Statistical aspects of fiber and bundle strength in hybrid composites, In: Hayashi, T., Kawata, K., Umekawa, S., (Eds.), Progress in Science in engineering Composites, ICCM-IV, Tokyo, pp. 1129–1136.

GHITANY, M. E., ATIEH, B., NADARAJAH, S., (2008). Lindley distribution and its Application, Mathematics Computing and Simulation, 78, pp. 493–506.

GHITANY, M. E ., Al-MMUTAIRI, D. K., BALAKRISHANAN, N., Al-ENEZI, L. J., (2013). Power Lindley distribution and Associated Inference, Computational Statistics and Data Analysis, 64, pp. 20–33.

LINDLEY, D. V., (1958). Fiducial distributions and Bayes' Theorem, Journal of the Royal Statistical Society, Series B, 20, pp.102–107.

SHAKED, M., SHANTHIKUMAR, J. G., (1994). Stochastic Orders and Their Applications, Academic Press, New York.

SHANKER, R., (2015). Akash Distribution and Its Applications, International Journal of Probability and Statistics, 4 (3), pp. 65–75.

SHANKER, R., HAGOS, F., SUJATHA, S., (2016). On Modeling of Lifetime Data using One-Parameter Akash, Lindley and Exponential Distributions, Biometrics & Biostatistics International Journal, 3 (2), pp. 1–10.

SHANKER, R., (2017). The Discrete Poisson-Akash Distribution, International Journal of Probability and Statistics, 6 (1), pp. 1–10

SHANKER, R., SHUKLA, K. K., (2017a). Ishita Distribution and its Applications**,** Biometrics & Biostatistics International Journal, 5 (2), pp. 1–9.

SHANKER, R., SHUKLA, K. K., (2017b). Power Akash distribution and its Application, to appear in Journal of Applied Quantitative Methods.

SHUKLA, K. K., SHANKER, R., (2017). The Discrete Poisson-Ishita Distribution, Communicated.

STACY, E. W., (1962). A generalization of the gamma distribution, Annals of Mathematical Statistics, 33, pp. 1187–1192.

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 149–158, DOI 10.21307/stattrans-2018-009

SOME RESULTS FROM THE 2013 INTERNATIONAL YEAR OF STATISTICS

Jan Kordos[1]

ABSTRACT

There are presented in this report, seven case studies of the uses of statistics in the past and present. I do not intend these examples to be exhaustive. I intend them primarily as educational examples for readers who would like to know: *What is statistics good for?* Also, to encourage the readers to study detailed reports from the 13 International Year of Statistics given in the notes of this report.

Key words: statistics, International Year of Statistics, Bayesian statistics, frequentist, data quality, official statistics, science of uncertainty, Markov Chain Monte Carlo (MCMC), Big Data.

1. Introduction

In 2013, six professional societies[2] declared an International Year of Statistics to celebrate the multifaceted role of statistics in contemporary society:

a) to raise public awareness of statistics, and;

b) to promote thinking about the future of the discipline.

In addition to these six societies, more than 2,300 organizations from 128 countries participated in the International Year of Statistics. The capstone event for this year of celebration *was the Future of the Statistical Sciences Workshop, held in London on November 11 and 12, 2013*. This meeting brought together more than 100 invited participants for two days of lectures and discussions. The organizers made the freely available lectures and discussions at Internet[3].

In Poland several organizations, societies and universities participated in the celebration. The Central Statistical Office of Poland and the Polish Statistical Association organized on 17-18 October 2013 a scientific conference entitled *Statistics – Knowledge – Development*[4].

[1] Warsaw Management University. Poland. E-mail: jan1kor2@gmail.com.

[2] The major sponsors of the yearlong celebration were: the American Statistical Association, the Royal Statistical Society, the Bernoulli Society, the Institute of Mathematical Statistics, the International Biometric Society, and the International Statistical Institute

[3] Statistics and Science – A Report of the London Workshop on the Future of the Statistical Sciences. http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf

[4] Some papers have been published in Statistics in Transition new series.

The Warsaw Management University and the Polish Statistical Association organized on 25-26 November 2013 a scientific conference entitled *Statistics in Service of Business and Social Sciences*[5].

The year 2013 was a very appropriate one for a celebration of statistics. It was the 300th anniversary of Jacob Bernoulli's *Ars conjectandi* (Art of Conjecturing) and the 250th anniversary of Thomas Bayes' "*An Essay Towards Solving a Problem in the Doctrine of Chances.*" The first of these papers helped lay the groundwork for the theory of probability. The second, little noticed in its time, eventually spawned an alternative approach to probabilistic reasoning that truly come to fruition in the computer age. In very different ways, Bernoulli and Bayes recognized that *uncertainty* is subject to mathematical rules and rational analysis. Nearly all research in science today requires the *management* and *calculation of uncertainty*, and for this reason statistics–*the science of uncertainty*–has become a crucial partner for modern science.

2. Purpose of this report

This report is projected primarily for people who are not experts in statistics. It is intended as a resource:

a) for students who might be interested in studying statistics and would like to know something about the field and where it is going;
b) for policymakers who would like to understand the value that statistics offers to society, and;
c) for people in the general public who would like to learn more about this often misunderstood field.

One common misconception about statisticians is that they are mere data collectors, or "*number crunchers*". That is almost the opposite of the truth. Often, the people who come to a statistician for help–whether they be scientists, CEOs[6], or public servants–either can collect the data themselves or have already collected it. The mission of the statistician is to work with the scientists to ensure that the data will be collected using the optimal method (free from bias and confounding). Then, the statistician extracts meaning from the data, so that the scientists can understand the results of their experiments and the CEOs and public servants can make well-informed decisions.

Another misperception, which is unfortunately all too common, is that the statistician is a person brought in to wave a magic wand and make the data say what the experimenter wants them to say. Statisticians provide researchers the tools to declare comparisons "statistically significant" or not, typically with the implicit understanding that statistically significant comparisons will be viewed as real and non-significant comparisons will be tossed aside. When applied in this way, statistics becomes a ritual to *avoid* thinking about uncertainty, which is again the opposite of its original purpose.

[5] E., Frączak, A. Kamińska, J., Kordos (Eds), (2014). Statistics – Business and Social Sciences Applications (in Polish). Available at: http://www.kaweczynska.pl/wydawnictwo/publikacje/wazniejsze-publikacje.

[6] CEOs communicate, collaborate, and exchange information on Earth observation activities, spurring useful partnerships such as the Integrated Global Observing Strategy (IGOS), http://ceos.org/about-ceos/overview/.

Ideally, statisticians should provide concepts and methods to learn about the world and help people make decisions in the face of uncertainty. If anything is certain about the future, it is that the world will continue to need this kind of "*honest broker*." It remains in question whether statisticians will be able to position themselves not as number crunchers or as practitioners of an arcane ritual, but as data explorers, data diagnosticians, data detectives, and ultimately as answer providers.

Statistics can be most succinctly described as the science of uncertainty. While the words "*statistics*" and "*data*" are often used interchangeably by the public, statistics actually goes far beyond the mere accumulation of data. The role of a statistician is:

- To design the acquisition of data in a way that minimizes bias and confounding factors and maximizes information content.
- To verify the quality of the data after it is collected.
- To analyze data in a way that produces insight or information to support decision-making.

These processes always take into explicit account the stochastic uncertainties present in any real-world measuring process, as well as the systematic uncertainties that may be introduced by the experimental design. This recognition is an inherent characteristic of statistics, and this is why we describe it as the "*science of uncertainty*," rather than the "*science of data.*"

Data are ubiquitous in 21st-century society: they pervade our science, our government, and our commerce. For this reason, statisticians can point to many ways in which their work has made a difference to the rest of the world. However, the very usefulness of statistics has worked in some ways as an obstacle to public recognition. Scientists and executives tend to think of statistics as infrastructure, and like other kinds of infrastructure, it does not get enough credit for the role it plays. Statisticians, with some prominent exceptions, also have been unwilling or unable to communicate to the rest of the world the value (and excitement) of their work.

3. Seven case studies of past "success stories" in statistics continued to the present day.

This report, therefore, begins with something that was mostly absent from the London workshop: seven case studies of past "success stories" in statistics, which in all cases have continued to the present day. These success stories are certainly not exhaustive–many others could have been told–but it is hoped that they are at least representative. They include:

1) The development of the ***randomized controlled trial methodology*** and appropriate methods for evaluating such trials, which are a required part of the drug development process in many countries.
2) The application of "***Bayesian statistics***" to image processing, object recognition, speech recognition, and even mundane applications such as spellchecking.
3) The explosive spread of "***Markov chain Monte Carlo***" methods, used in statistical physics, population modelling, and numerous other applications to

simulate uncertainties that are not distributed according to one of the simple textbook models (such as the "bell-shaped curve").

4) ***The involvement of statisticians in many high-profile court cases over the years***. When a defendant is accused of a crime because of the extraordinary unlikelihood of some chain of events, it often falls to statisticians to determine whether these claims hold water.
5) The discovery through statistical methods of "***biomarkers***"[7] – genes that confer an increased or decreased risk of certain kinds of cancer.
6) A method called "***kriging***[8]", which enables scientists to interpolate a smooth distribution of some quantity of interest from sparse measurements. Application fields include mining, meteorology, agriculture, and astronomy.
7) The rise of "***analytics***" in sports and politics in recent years. In some cases, the methods involved are not particularly novel, but what is new is the recognition by stakeholders (sports managers and politicians) of the value that objective statistical analysis can add to their data.

Statistics was a multidisciplinary science from the very beginning, long before that concept became fashionable. The same techniques developed to analyze data in one application are very often applicable in numerous other situations. One of the best examples of this phenomenon in recent years is the application of Markov Chain Monte Carlo (MCMC) methods. While MCMC was initially invented by statistical physicists who were working on the hydrogen bomb, it has since been applied in settings as diverse as image analysis, political science, and digital humanities. Markov Chain Monte Carlo is essentially a method for taking random samples from an unfathomably large and complex probability distribution.

The original algorithm was designed in the late 1940s by Nicholas Metropolis, Stanislaw Ulam, the Polish statistician, Edward Teller, and others to simulate the motion of neutrons in an imploding hydrogen bomb. This motion is essentially random. However, "random" does not mean "arbitrary." The neutrons obey physical laws, and this makes certain outcomes much more likely than others. The probability space of all plausible neutron paths is far too large to store in a computer, but Metropolis' algorithm enables the computer to pick random plausible paths and thereby predict how the bomb will behave.

In a completely different application, MCMC has been used to analyze models of how politicians vote on proposed legislation or how U.S. Supreme Court justices vote on cases that come before them. The second example is of particular interest because the justices typically say very little in public about their political viewpoints after their confirmation hearings, yet their ideologies can and do change quite a bit during the course of their careers. Their votes are the only indicator of these changes. While political pundits are always eager to "read the tea leaves," their analysis typically lacks objectivity and quantitative rigor.

The International Year of Statistics came at a time when the subject of statistics itself stood at a crossroads. Some of its most impressive achievements

[7] The term "biomarker", a portmanteau of "biological marker", refers to a broad subcategory of medical signs – that is, objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly.

[8] kriging – optimal interpolation based on regression against observed z values of surrounding data points, weighted according to spatial covariance values, http://www.kriging.com/whatiskriging.html.

in the 20th century had to do with extracting as much information as possible from relatively small amounts of data–for example, predicting an election based on a survey of a few thousand people, or evaluating a new medical treatment based on a trial with a few hundred patients.

4. BIG DATA

While these types of applications will continue to be important, there is a new game in town. We live in the era of ***BIG DATA***. Companies such as Google or Facebook gather enormous amounts of information about their users or subscribers. They constantly run experiments on, for example, how a page's layout affects the likelihood that a user will click on a particular advertisement. These experiments have millions, instead of hundreds, of participants, a scale that was previously inconceivable in social science research. In medicine, *the Human Genome Project* has given biologists access to an immense amount of information about a person's genetic makeup. Before Big Data, doctors had to base their treatments on a relatively coarse classification of their patients by age group, sex, symptoms, etc. Research studies treated individual variations within these large categories mostly as "noise." Now doctors have the prospect of being able to treat every patient uniquely, based on his or her *DNA*. Statistics and statisticians are required to put all these data on individual genomes to effective use.

The rise of Big Data has forced the field to confront a question of its own identity The creation of this new job category brings both opportunity and risk to the statistics community. The value that statisticians can bring to the enterprise is their ability to ask and to answer such questions as these:

a) Are the data representative?
b) What is the nature of the uncertainty?
c) It may be an uphill battle even to convince the owners of Big Data that their data are subject to uncertainty and, more importantly, bias.

On the other hand, it is imperative for statisticians not to be such purists that they miss the important scientific developments of the 21st century. "Data science" will undoubtedly be somewhat different from the discipline that statisticians are used to. Perhaps statisticians will have to embrace a new identity. Alternatively, they might have to accept the idea of a more fragmented discipline in which standard practices and core knowledge differ from one branch to another.

Undoubtedly the greatest challenge and opportunity that confronts today's statisticians is the rise of *Big Data*–databases on the human genome, the human brain, Internet commerce, or social networks (to name a few), which dwarf in size any databases statisticians encountered in the past. *Big Data* is a challenge for several reasons:

1) *Problems of scale.* Many popular algorithms for statistical analysis do not scale up very well and run hopelessly slowly on terabyte-scale data sets. Statisticians either need to improve the algorithms or design new ones that trade off theoretical accuracy for speed.

2) *Different kinds of data.* Big Data are not only big, they are complex and they come in different forms from what statisticians are used to, for instance images or networks.
3) *The "look-everywhere effect".* As scientists move from a hypothesis-driven to a data-driven approach, the number of spurious findings (e.g. genes that appear to be connected to a disease but really are not) is guaranteed to increase, unless specific precautions are taken.
4) *Privacy and confidentiality.* This is probably the area of greatest public concern about Big Data, and statisticians cannot afford to ignore it. Data can be anonymized to protect personal information, but there is no such thing as perfect security.
5) *Reinventing the wheel.* Some of the collectors of Big Data—notably, web companies—may not realize that statisticians have generations of experience at getting information out of data, as well as avoiding common fallacies. Some statisticians resent the new term "data science". Others feel we should accept the reality that "data science" is here and focus on ensuring that it includes training in statistics.

Big Data was not the only current trend discussed at different meetings, and indeed there was a minority sentiment that it is an overhyped topic that will eventually fade. Other topics that were discussed include:

i. *The reproducibility of scientific research.* Opinions vary widely on the extent of the problem, but many "discoveries" that make it into print are undoubtedly spurious. Several major scientific journals are requiring or encouraging authors to document their statistical methods in a way that would allow others to reproduce the analysis.
ii. *Updates to the randomized controlled trial.* The traditional RCT[9] is expensive and lacks flexibility. "Adaptive designs[10]" and "SMART trials[11]" are two modifications that have given promising results, but work still needs to be done to convince clinicians that they can trust innovative methods in place of the tried-and-true RCT.
iii. *Statistics of climate change[12].* This is one area of science that is begging for more statisticians. Climate models do not explicitly incorporate uncertainty, so the uncertainty has to be simulated by running them repeatedly with slightly different conditions.
iv. *Statistics in other new venues.* For instance, one talk explained how new data capture methods and statistical analysis are improving (or will improve) our understanding of the public diet. Another participant described how the United Nations is experimenting for the first time with probabilistic, rather than deterministic, population projections.
v. *Communication and visualization.* The Internet and multimedia give statisticians new opportunities to take their work directly to the public.

[9] RCT (Randomized Control Trial) is a type of scientific (often medical) experiment which aims to reduce bias when testing a new treatment.
[10] http://adaptivedesigns.com/about.
[11] SMART-trials – a next generation platform intended for data acquisition in medical research and clinical trials, https://www.cognizant.com/SmartTrials.
[12] http://data.worldbank.org/topic/climate-change.

vi. *Education.* A multifaceted topic, this was discussed a great deal but without any real sense of consensus. Most participants at the meeting seemed to agree that the curriculum needs to be re-evaluated and perhaps updated to make graduates more competitive in the workplace. Opinions varied as to whether something needs to be sacrificed to make way for more computer science–type material, and if so, what should be sacrificed.

vii. *Professional rewards.* The promotion and tenure system needs scrutiny to ensure non-traditional contributions such as writing a widely used piece of statistical software are appropriately valued. The unofficial hierarchy of journals, in which theoretical journals are more prestigious than applied ones and statistical journals count for more than subject-matter journals, is also probably outmoded.

5. Official/government statistics

It is a little-known fact that the word “statistics” actually comes from the root “state”–it is the science of the state. Thus, government or official statistics have been involved in the discipline from the beginning, and, for many citizens, they are still the most frequently encountered form of statistics in daily life.

Several trends are placing new demands on official statisticians. Many governments are moving toward open government, in which all official data will be available online. Many constituents expect these data to be free. However, open access to data poses new problems of privacy, especially as it becomes possible to parse population data into finer and finer units. Free access is also a problem in an era of flat or declining budgets. Though information may want to be free, it is certainly not free to collect and curate.

At the same time, new technologies create new opportunities. There are new methods of collecting data, which may be much cheaper and easier than traditional surveys. As governments move online, administrative records become a useful and searchable source of information. Official statisticians will face a Big Data problem similar to private business as they try to figure out what kinds of usable information might exist in these large volumes of automatically collected data and how to combine them with more traditionally collected data. They also need to think about the format of the data; mounds of page scans or data that are presented out of context may not be very useful. With proper attention to these issues, both old democracies and new democracies can become more transparent, and the citizens can become better informed about what their governments are doing.

But the more time that students spend learning computer science, the less time they will have available for traditional training in statistics. The discussion of what parts of the “core” can be sacrificed, or if there even is a “core” that is fundamental for all students, produced even less agreement. A few voices tentatively called for less emphasis on the abstract mathematical foundations of the subject. However, some attendees felt that the unity of the subject was its strength, and they remembered fondly the days when they could go to a statistics meeting and understand any lecture. Even they acknowledged that things are changing; the trend is toward a field that is more diverse and fragmented. Should

this trend be resisted or embraced? Will the pressure of Big Data be the straw that breaks the camel's back, or the catalyst that drives a long needed change? On questions like these, there was nothing even approaching consensus.

6. Quality of Data

One of the underrated services that statisticians can provide in the world of Big Data is to look at the quality of data with a skeptical eye. This tradition is deeply ingrained in the statistical community, beginning with the first controlled trials in the 1940s. Data come with a provenance. If they come from a double-blind randomized controlled trial, with potential confounding factors identified and controlled for, then the data can be used for statistical inference. If they come from a poorly designed experiment–or, even worse, if they come flooding into a corporate web server with no thought at all given to experimental design–the identical data can be worthless.

In the world of Big Data, someone has to ask questions like the following:

- Are the data collected in a way that introduces bias? Most data collected on the Internet, in fact, come with a sampling bias. The people who fill out a survey are not necessarily representative of the population as a whole.
- Are there missing or incomplete data? In Web applications, there is usually a vast amount of unknown data. For example, the movie website Netflix wanted to recommend new movies to its users using a statistical model, but it only had information on the handful of movies the user had rated. It spent $1 million on a prize competition to identify a better way of filling in the blanks.
- Are there different kinds of data? If the data come from different sources, some data might be more reliable than others. If all the numbers get put into the same analytical meat grinder, the value of the high-quality data will be reduced by the lower-quality data. On the other hand, even low-quality, biased data *might* contain some useful information. Also, data come in different formats–numbers, text, networks of "likes" or hyperlinks. It may not be obvious to the data collector how to take advantage of these less traditional kinds of information.

  Statisticians not only know how to ask the right questions, but, depending on the answers, they may have practical solutions already available.

7. Some conclusions

The Workshop on the Future of Statistics did not end with a formal statement of conclusions or recommendations. However, the following unofficial observations may suffice:

1. The analysis of data using statistical methods is of fundamental importance to society. It underpins science, guides business decisions, and enables public officials to do their jobs.
2. All data come with some amount of uncertainty, and the proper interpretation of data in the context of uncertainty is by no means easy or routine. This is one of the most important services that statisticians provide to society.

3. Society is acquiring data at an unprecedented and ever-increasing rate. Statisticians should be involved in the analysis of these data.
4. Statisticians should be cognizant of the threats to privacy and confidentiality that Big Data pose. It will remain a challenging problem to balance the social benefits of improved information with the potential costs to individual privacy.
5. Data are coming in new and untraditional forms, such as images and networks. Continuing evolution of statistical methods will be required to handle these new types of data.
6. Statisticians need to reevaluate the training of students and the reward system within their own profession to make sure that these are still functioning appropriately in a changing world.
7. In particular, statisticians are grappling with the question of what a "data scientist" is, whether it is different from a statistician, and how to ensure that data scientists do not have to "reinvent the wheel" when they confront issues of uncertainty and data quality.
8. In a world where the public still has many misperceptions about statistics, risk, and uncertainty, communication is an important part of statisticians' jobs. Creative solutions to data visualization and mass communication can go a long way.

We conclude with some observations on statistical education, which was a major topic of discussion at the London workshop, even though there were no formal lectures about it.

Clearly, some students are getting the message that statistics is a useful major, and many of them are undoubtedly attracted by the job possibilities. However, statistics departments need to do a better job of preparing them for the jobs that are actually available and not necessarily to become carbon copies of the professors. Some suggestions include the following:

- *Working on communication skills.* Statisticians have a deep understanding and familiarity with the concept of uncertainty that many other scientists lack. They will only be able to disseminate their knowledge of this critical concept if they can convey it readily and with ease.
- *Working on team projects, especially with non-statisticians*. The workshop itself modeled this behavior, as most of the speakers who were statisticians were paired with a non-statistician who is an expert in the subject-matter area under discussion. In most cases, the two speakers were collaborators.
- *Training on leadership skills.* There was a strong sentiment among some workshop participants that statisticians are pigeonholed as people who support the research of others, rather than coming up with original ideas themselves.
- *Strong training in an application field.* This again may help prepare the students to steer the direction of research, rather than following it.
- *More exposure to real "live" data.* Many students will learn best if they can see the applicability to real-world problems.
- *More exposure to Big Data, or at least reasonably Big Data that cannot be analyzed using traditional statistical methods or on a single computer.* Students need to be prepared for the world that they will be entering, and Big Data seems to be here to stay.

- *More emphasis on computer algorithms, simulation, etc.* To prepare for engineering-type jobs, students need to learn to think like engineers.

To sum up, the view of statistics that emerged from the conferences and workshops was one of a field that, after three centuries, is as healthy as it ever has been, with robust growth in student enrolment, abundant new sources of data, and challenging problems to solve over the next century.

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 159–166, DOI 10.21307/stattrans-2018-010

PRODUCT EXPONENTIAL METHOD OF IMPUTATION IN SAMPLE SURVEYS

Shakti Prasad[1]

ABSTRACT

In this paper, a product exponential method of imputation has been suggested and their corresponding resultant point estimator has been proposed for estimating the population mean in sample surveys. The expression of bias and the mean square error of the suggested estimator has also been derived, up to the first order of large sample approximations. Compared with the mean imputation method, Singh and Deo (Statistical Papers (2003)) and Adapted estimator (Bahl and Tuteja (1991)), the simulation studies show that the suggested estimator is the most efficient estimator.

Key words: imputation methods, bias, mean square error (MSE), efficiency.

1. Introduction

The use of auxiliary information for estimating the finite population mean of the study variable has played an eminent role in sample surveys. The ratio imputation method is employed for missing data if the correlation between the study variable and the auxiliary variable is positive. On the other hand, if this correlation is negative, the product imputation method investigated by Singh and Deo (2003), is quite effective. The application of the product imputation method has too much importance but absolutely has some limitation in medical discipline, industrial and social science, etc. There are several medical or social science related variables which decrease as the people grow up. For example, as the people become older, the following variables have negative correlation with the age: (a) duration of sleeping hours (b) hearing tendency (c) eye sight (d) number of hairs on the head (e) number of love affairs and (f) working hour capacity etc. If information on any of these study variable is missing, but the age of the persons is available, the product imputation method will be beneficial.

It is worth to be noticed that appreciable amount of works carried out under product method of estimation in sample surveys by several authors, but its application very limited in imputation methods. Dated back, Singh and Deo (2003) have used the product imputation method in survey sampling. Motivated with the above work, we study the some product exponential method of imputation in sample surveys.

Let y and x be denoted by the negatively correlated study variable and auxiliary variable respectively. A simple random sample (without replacement) s_n of

[1]Department of Basic & Applied Science National Institute of Technology, Arunachal Pradesh,Yupia, Papum pare-791112, India. E-mail: shakti.pd@gmail.com.

n units is depleted from a finite population $U = (U_1, U_2, ..., U_N)$ of N units to estimate the population mean $\bar{Y}$. Let r be the number of responding units out of sampled n units, the set of responding units by R and the set of non-responding units by R^c. If the units involve the responding unit set, the values on the study variable y_i are observed. If they involve the non-responding unit set, the values on the study variable y_i are missing and hereafter the imputed values are derived for a well known units.

$$y_{.i} = \begin{cases} y_i & \text{if } i\varepsilon R \\ \tilde{y}_i & \text{if } i\varepsilon R^c \end{cases} \tag{1}$$

The general point estimator of population mean $\bar{Y}$ takes the form:

$$\bar{y}_s = \frac{1}{n}\sum_{i\varepsilon s_n} y_{.i} = \frac{1}{n}\left[\sum_{i\varepsilon R} y_{.i} + \sum_{i\varepsilon R^c} y_{.i}\right] = \frac{1}{n}\left[\sum_{i\varepsilon R} y_i + \sum_{i\varepsilon R^c} \tilde{y}_i\right] \tag{2}$$

Here, the value $\tilde{y}_i$ denote the imputed value of the study variable corresponding to the i^{th} non- responding units.
The consequently notations have been approaching in this work:
$\bar{Y}$, $\bar{X}$: The population means of the variables y and x respectively.
$\bar{y}_r$, $\bar{x}_r$: The response means of the respective variables for the sample sizes shown in suffices.
$\bar{x}_n$: The sample mean of the variable x.
ρ_{yx} : The population correlation coefficient between the variables y and x.
$S_x^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$: The population mean square of the variable x.
S_y^2 : The population mean square of the variable y .
C_y and C_x : The coefficients of variation of the variables shown in suffices.

2. Some Existing Estimators

In this section, the several estimators with imputation have been discussed for estimating the population mean in sample surveys.

2.1. Mean Imputation Method

Under this method, After imputation, data take the form:

$$y_{.i} = \begin{cases} y_i & \text{if } i\varepsilon R \\ \bar{y}_r & \text{if } i\varepsilon R^c \end{cases} \tag{3}$$

The resultant point estimator (2) of $\bar{Y}$ becomes

$$\bar{y}_m = \frac{1}{r}\sum_{i=1}^{r} y_i = \bar{y}_r \tag{4}$$

which is known as the response mean estimator $\bar{y}_r$ of population mean $\bar{Y}$.
The variance of the response sample mean $\bar{y}_r$, is given by

$$Var(\bar{y}_r) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2 \tag{5}$$

2.2. Product Imputation Method

Singh and Deo (2003) proposed the product imputation method in sample surveys. After imputation, data take the form:

$$y_{.i} = \begin{cases} y_i & \text{if } i\varepsilon R \\ \bar{y}_r \left[\frac{n\bar{x}_r - r\bar{x}_n}{\bar{x}_n}\right] \frac{x_i}{\sum_{i\varepsilon R^c} x_i} & \text{if } i\varepsilon R^c \end{cases} \tag{6}$$

Under this method of imputation, the resultant point estimator (2) of $\bar{Y}$ becomes

$$\bar{y}_{SD} = \bar{y}_r \frac{\bar{x}_r}{\bar{x}_n} \tag{7}$$

which is analogue of the product estimator proposed by Murthy (1964).
The MSE of estimator $\bar{y}_{SD}$, is given by

$$MSE(\bar{y}_{SD}) = Var(\bar{y}_r) + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2\left(C_x^2 + 2\rho_{yx}C_yC_x\right) \tag{8}$$

3. Adapted Product Exponential Method of Imputation

Following the Bahl and Tuteja (1991), We have adapted product exponential method of imputation and their corresponding estimator for estimating the $\bar{Y}$ in survey sampling.
The adapted imputation method,
After imputation, the data take the form:

$$y_{.i} = \begin{cases} y_i & \text{if } i\varepsilon R \\ \frac{\bar{y}_r}{n-r}\left[n\exp\left(\frac{\bar{x}_r - \bar{X}}{\bar{x}_r + \bar{X}}\right) - r\right] & \text{if } i\varepsilon R^c \end{cases} \tag{9}$$

Under above adapted imputation methods, the resultant point estimators (2) of the population mean $\bar{Y}$ become

$$\bar{y}_{AE} = \bar{y}_r \exp\left[\frac{\bar{x}_r - \bar{X}}{\bar{x}_r + \bar{X}}\right] \tag{10}$$

The MSE of estimator $\bar{y}_{AE}$, is given by

$$MSE(\bar{y}_{AE}) = \left(\frac{1}{r} - \frac{1}{N}\right)\left(C_y^2 + \frac{1}{4}C_x^2 + \rho_{yx}C_yC_x\right)\bar{Y}^2 \tag{11}$$

4. Suggested Method and their Estimator

Following the Prasad (2016 & 2017), a product exponential method of imputation and their corresponding estimator has suggested for estimating the population mean $\bar{Y}$ in sample surveys.
The suggested imputation method,
After imputation, the data take the form:

$$y_{.i} = \begin{cases} \phi y_i \exp\left(\frac{(\bar{x}_r - \bar{X})S_x}{(\bar{X}+\bar{x}_r)S_x + 2C_x}\right) & \text{if } i\varepsilon R \\ \phi \frac{\bar{y}_r}{\bar{x}_r}\left(x_i - \frac{n}{n-r}(\bar{x}_n - \bar{x}_r)\right)\exp\left(\frac{(\bar{x}_r - \bar{X})S_x}{(\bar{X}+\bar{x}_r)S_x + 2C_x}\right) & \text{if } i\varepsilon R^c \end{cases} \tag{12}$$

Under above suggested imputation method, the resultant point estimator (2) of the population mean $\bar{Y}$ becomes

$$\zeta = \phi\bar{y}_r \exp\left[\frac{(\bar{x}_r - \bar{X})S_x}{(\bar{X}+\bar{x}_r)S_x + 2C_x}\right] \tag{13}$$

where ϕ is suitably chosen constant, such that the MSE of the resultant point estimator is minimum. It has been assumed that S_x and C_x are known.

5. Properties of the suggested estimator ζ

The bias and their mean square error (MSE) of the suggested estimator ζ are derived up to the first order of large sample approximations under the following transformations:
$\bar{y}_r = \bar{Y}(1+e_y)$ and $\bar{x}_r = \bar{X}(1+e_x)$ such that $E(e_i) = 0$, $|e_i| < 1 \forall\ i = y, x$.
Using the above transformations, the estimator ζ take the following form:

$$\zeta = \phi\bar{Y}(1+e_y)exp\left[\frac{1}{2}\theta e_x\left(1 + \frac{1}{2}\theta e_x\right)^{-1}\right] \tag{14}$$

where $\theta = \frac{\bar{X}S_x}{\bar{X}S_x + C_x}$.
Neglecting the higher power terms of $e's$, the equation(14) can be written as

$$\zeta - \bar{Y} \cong \bar{Y}\left[(\phi - 1) + \phi\left(e_y + \frac{1}{2}\theta e_x + \frac{1}{2}\theta e_y e_x - \frac{1}{8}\theta^2 e_x^2\right)\right] \tag{15}$$

Taking expectation of (15), we obtained the bias of the suggested estimator, is

given as

$$Bias(\zeta) = \bar{Y}\left[(\phi-1) - \frac{1}{8}\theta\phi\left(\frac{1}{r}-\frac{1}{N}\right)\left(\theta C_x^2 - 4\rho_{yx}C_yC_x\right)\right] \tag{16}$$

Now, after squaring of (15) and neglecting the higher power terms of $e's$, we have

$$(\zeta-\bar{Y})^2 \cong \bar{Y}^2\left[(\phi-1)+\phi\left(e_y+\frac{1}{2}\theta e_x+\frac{1}{2}\theta e_y e_x-\frac{1}{8}\theta^2 e_x^2\right)\right]^2 \tag{17}$$

Taking expectation of (17), we get the MSE of the suggested estimator ζ as

$$MSE(\zeta) = \bar{Y}^2\left[(\phi-1)^2+\phi^2 A+2(\phi^2-\phi)B\right] \tag{18}$$

where $A = \left(\frac{1}{r}-\frac{1}{N}\right)\left(C_y^2+\frac{1}{4}\theta^2C_x^2+\theta\rho_{yx}C_yC_x\right)$, $B = \left(\frac{1}{r}-\frac{1}{N}\right)\left(-\frac{1}{8}\theta^2C_x^2+\frac{1}{2}\theta\rho_{yx}C_yC_x\right)$. Differentiating (18) with respect to ϕ, and its equating to zero respectively, we get the optimum value of ϕ, is given by

$$\phi_{opt} = \frac{1+\left(\frac{1}{r}-\frac{1}{N}\right)\left(-\frac{1}{8}\theta^2C_x^2+\frac{1}{2}\theta\rho_{yx}C_yC_x\right)}{1+\left(\frac{1}{r}-\frac{1}{N}\right)\left(C_y^2+2\theta\rho_{yx}C_yC_x\right)} \tag{19}$$

After substituting the optimum value of ϕ, i. e., ϕ_{opt} in equation (18), we obtain the minimum MSE of the suggested estimator ζ, is given as

$$MSE(\zeta)_{opt} = \left[1-\frac{\left(1+\left(\frac{1}{r}-\frac{1}{N}\right)\left(-\frac{1}{8}\theta^2C_x^2+\frac{1}{2}\theta\rho_{yx}C_yC_x\right)\right)^2}{1+\left(\frac{1}{r}-\frac{1}{N}\right)\left(C_y^2+2\theta\rho_{yx}C_yC_x\right)}\right]\bar{Y}^2 \tag{20}$$

6. Simulation Study

We have considered the four data sets for the sample population between 25% to 50%, response rate between 60% to 94% with different correlation coefficient. The percent relative efficiency of the suggested estimator engaged in simulation study. The PREs of the suggested estimator ζ with respect to the mean imputation method, Singh and Deo (2003) estimator and Adapted estimator (Bahl and Tuteja (1991)) are obtained as

$$PRE_1 = \frac{V(\bar{y}_r)}{MSE(\zeta)_{opt}} \times 100 \tag{21}$$

$$PRE_2 = \frac{MSE(\bar{y}_{SD})}{MSE(\zeta)_{opt}} \times 100 \tag{22}$$

$$PRE_3 = \frac{MSE(\bar{y}_{AE})}{MSE(\zeta)_{opt}} \times 100 \tag{23}$$

Table 1: Description of Data sets

Parameters	Data set 1 Maddala (1977)	Data set 2 Pandey and Dubey (1988)	Data set 3 Singh, S. (2003)	Data set 4 Swain (2013)
N	16	20	30	50
n	4	8	10	15
r	3	(6, 7)	(6,7,8,9)	(10, 12, 14)
$\bar{Y}$	7.6375	19.55	384.2	6.5945
$\bar{X}$	75.4343	18.8	67.267	55
C_y	0.2278	0.3552	0.1559	0.2134
C_x	0.0986	0.3943	0.1371	0.2968
ρ_{yx}	-0.6823	-0.9199	-0.8552	-0.8628

Table 2: Percent relative efficiency of the suggested estimator ζ over the estimators $\bar{y}_r$, $\bar{y}_{SD}$ and $\bar{y}_{AE}$ respectively under the four different Data sets.

$Dataset$	N	n	r	PRE_1	PRE_2	PRE_3
1	16	4	3	133.898	117.282	100.626
2	20	8	6	349.680	248.514	100.332
			7	349.152	294.760	100.180
3	30	10	6	226.579	143.789	99.9824
			7	226.579	161.787	99.9823
			8	226.579	181.421	99.9822
			9	226.579	202.924	99.9821
4	50	15	10	353.940	285.272	100.376
			12	353.641	310.309	100.291
			14	353.429	338.191	100.231

7. Analysis of Simulation Study

From Tables (1 - 2), the following interpretation can be read out:
(1) From Table 1 presents the parameters of the four data sets for different correlation coefficient. We are taking different values for n and r.
(2) From Table 2 it is observed that

(a) For a 25% sample population with response rate is 75%, the PRE of the suggested estimator ζ with respect to the other existing estimators like as the mean imputation method remains 133.898%, Singh and Deo ($\bar{y}_{SD}$) estimator remains 117.282 % and Adapted estimator remains 100.626%.

(b) For a 40% sample population with response rate between 75% to 87%, the PRE of the suggested estimator ζ with respect to the other existing estimators like as the mean imputation method remains 349.152% to 349.680%, Singh and Deo ($\bar{y}_{SD}$) estimator remains 248.514% to 294.760% and Adapted estimator remains 100.180% to 100.332%.

(c) For a 33% sample population with response rate between 60% to 90%, the PRE of the suggested estimator ζ with respect to the other existing estimators like as the mean imputation method remains 226.579% to 226.579%, Singh and Deo ($\bar{y}_{SD}$) estimator remains 143.789% to 202.924% and Adapted estimator remains 99.9821% to 99.9824 %.

(d) For a 30% sample population with response rate between 67% to 94%, the PRE of the suggested estimator ζ with respect to the other existing estimators like as the mean imputation method remains 353.429% to 353.940%, Singh and Deo ($\bar{y}_{SD}$) estimator remains 285.272% to 338.191% and Adapted estimator remains 100.231% to 100.376 %.

8. Conclusions

From the above analysis, it is observed that the suggested estimator is more efficient than the mean imputation method, Singh and Deo (2003) estimator and Adapted estimator (Bahl and Tuteja (1991)). Hence, it can be recommended for future use.

Acknowledgements

The author is grateful to the reviewers for their constructive comments and valuable suggestions regarding improvement of the article.

REFERENCES

BAHL, S., TUTEJA, R. K., (1991). Ratio and Product type exponential estimator. Journal of Information and Optimization Sciences, 12 (1), pp. 159–164.

MADDALA, G. S., (1977). Econometrics. McGraw Hills Pub. Co., New York.

MURTHY, M. N., (1967). Sampling theory and methods. Statistical Publishing Society, Calcutta, India.

PANDEY, B. N., DUBEY, V., (1988). Modified product estimator using coefficient of variation of auxiliary variable. Assam Stat. Review, 2, pp. 64–66.

PRASAD, S., (2016). A study on new methods of ratio exponential type imputation in sample surveys. Hacettepe Journal of Mathematics and Statistics, DOI:10.15672/HJMS.2016.392.

PRASAD, S., (2017). Ratio exponential type imputation in sample surveys. Model Assisted Statistics and Application, 12 (2), pp. 95–106.

SINGH, S., DEO, B., (2003). Imputation by power transformation. Statistical Papers, 44, pp. 555–579.

SINGH, S., (2003). Advanced sampling theory with applications. Klumer Academic Publishers, the Netherlands, Vol. 1.

SWAIN, A. K. P. C., (2013). On some modified ratio and product type estimators-Revisited. Revista Investigacion Operacional, 34 (1), pp. 35–57.

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 167

GUS-SCORUS 2018 conference
NEW SCALE – NEW NEEDS – NEW STATISTICS

On behalf of the co-organizers of the Warsaw Conference SCORUS 2018 – the Statistics Poland (GUS) and Standing Committee of Regional and Urban Statistics (SCORUS) – we are pleased to inform that the event will be held on 6-8 June 2018 in Warsaw, Poland.

"The SCORUS 2018 Conference is the opportunity to share knowledge on most recent developments and challenges of regional and urban statistics. Statistics Poland, Eurostat and SCORUS have a pleasure to invite all those working on development of urban and regional statistics to join this event."

The Scientific Programme Committee of the SCORUS 2018 Conference has selected three themes: "New Scale – New Needs – New Statistics".

http://scorus.org/index.php/home/

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 169

On behalf of the co-organizers of the Q2018 – the Statistics Poland (CSO) and Eurostat – we are pleased to inform that the 2018 European Conference on Quality in Official Statistics will take place 26-29 June in Kraków, Poland.

"The Q2018 Conference will be one of the series of scientific gatherings covering methodological and quality-related issues that are relevant to the development of the European Statistical System".

http://ec.europa.eu/eurostat/web/ess/-/q2018-conference

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 171

The 2nd Congress of Polish Statistics organised on the occasion of the 100th anniversary of the establishment of the Statistics Poland will be held on July 10-12, 2018 in Warsaw.

"The Congress will last three days. The framework program of the event contains a series of thematic sessions, including a jubilee panel on the history of Polish statistics, as well as sessions devoted to Polish statistics on the international arena, methodology of statistical surveys, mathematical statistics, regional statistics, population statistics, social and economic statistics, statistical data issues and, also sports and tourism statistics.

In the Congress, which will emphasise the contribution of Poles to the global treasury of statistical knowledge, representatives of foreign institutions and scientific units will participate.
We are convinced that the Congress will constitute a unique opportunity for the representatives of official statistics, research centres and other partners involved in the study of social, economic and environmental processes to meet and exchange their knowledge, views and experiences."

https://kongres.stat.gov.pl/en/

STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 173–176

ABOUT THE AUTHORS

Agiwal Varun is a research scholar in the Department of Statistics, Central University of Rajasthan. His main area of interest includes time series and Bayesian inference. Currently, he is working on the problem of structural changes in time series.

Bouchahed Lahsen is a faculty member at the Department of Mathematics at The University of Badji-Mokhtar, Annaba-Algeria. He received his magister degree in mathematics from Badji Mokhtar University. His research areas are in: applied statistics, dynamics systems and probability.

Hämäläinen Auli is a Junior Researcher at the University of Helsinki and a Systems Analyst at National Land Survey of Finland.

Karna Jaishree Prabha is currently a postdoctoral fellow of the Department of Statistics, Gauhati University, Guwahati. She received PhD in Applied Mathematics from the Indian School of Mines, Dhanbad, India, in 2010. Her main research interests are survey sampling, handling non response problems and related areas.

Kordos Jan graduated from Jagiellonian University and Wroclaw University (in mathematical statistics, 1955); PhD in Econometrics from the Academy of Economics, Katowice, Poland (1965), and Professorship (1990). He served as the FAO Adviser in Agricultural Statistics in Ethiopia (1974-80). He was the Vice President of the CSO Poland (1992-96). He was lecturing and training on agricultural Statistics in China in late 80s, and also in Kathmandu, Nepal (1991). During 1994-96 he served as the World Bank Consultant in Household Budget Surveys in Latvia and Lithuania. He was President of the Polish Statistical Association (1985-94). He was founder and the Editor-in-Chief of the Statistics in Transition (1993-2007). Now, he is Professor of statistics at the Warsaw Management University. His publications include four books and over three hundreds articles and other papers.

Krzyśko Mirosław is a Full Professor and Professor Emeritus at the Department of Probability and Mathematical Statistics in Adam Mickiewicz University, Poznań, Poland. His research interests are multivariate statistical analysis, analysis of multivariate functional data, statistical inference and data analysis in particular. Professor Krzyśko has published over 150 research papers in international/national journals and conferences. He has also published five books/monographs. Professor Krzyśko is an active member of many scientific professional bodies.

Kumar Jitendra is working as an Associate Professor in the Department of Statistics, Central University of Rajasthan, Bandersindri, Ajmer, India. Before joining Central University of Rajasthan he served CUB, Patna, IDRBT, Hyderabad

and SHIAU, Allahabad. His interest include time series, outlier, crime statistics, policy process reengineering and big data.

Laaksonen Seppo is Professor of statistics at the University of Helsinki, who has worked in various survey institutes including Statistics Finland, Eurostat and the Finnish Centre for Social and Health Research. He was the former Vice President of the International Association of Survey Statisticians from 2007 to 2009. Laaksonen was a member of the sampling expert team of the European Social Survey, 2001-2018. He has also been involved in a number of European research projects and has been a consultant for surveys in Moldova, Ethiopia, Slovenia, the United Kingdom and Hungary. His main research area is survey methodology in its broad meaning, that is, it includes empirical microeconometrics (incomes, wages, employment, etc.). He has published 210 research articles or monographs.

Longford Nicholas T. is a Senior Statistician in the Neonatal Data Analysis Unit, Department of Medicine, Imperial College London (since 2015). His interests include methods for dealing with missing data, interpreted broadly to cover causal analysis, model selection and small-area estimation. He is an Associate Editor of Statistical Methods in Medical Research and Journal of Educational and Behavioral Statistics. His past appointments include Educational Testing Service, USA and De Montfort University, UK. During 2004-2015 he directed the statistical consulting company SNTL (website: www.sntl.co.uk).

Łukaszonek Wojciech is working as an assistant at President Wojciechowski Higher Vocational State School in Kalisz, Poland. In 2002 he received his Master's Degree in Mathematics from Technical University of Łódź. Currently he is working on PhD in the field of the analysis of higher educational market at Poznań University of Economics and Business.

Nath Dilip C. is currently a Vice-Chancellor, Assam University, Silchar, India. Previously, he was Professor of Statistics in Gauhati University, Guwahati. He was associated with different national and international universities, viz. Delhi University, Banaras Hindu University, Duke University, USA, University of Washington, USA and Max Plank Institute for Demographic Research, Germany. His areas of research include fertility, reproductive health, aging and health of elderly population, disease burden. He was awarded Rockefeller Foundation Fellowship (1991-93) and Andrew Mellon Fellowship (1997-98).

Osaulenko Oleksandr is the Rector of the National Academy of Statistics, Accounting and Audit, Doctor of Public Administration, Professor, Corresponding Member of the National Academy of Sciences of Ukraine, Honoured Economist of Ukraine. During 1996-2014, he headed the national statistical office of Ukraine. He is an author of over 200 scientific works, including 20 monographs and handbooks on the problems of statistics, public administration and information security.

Panchenko Volodymyr is a doctoral student of Mariupol State University in Ukraine. He is an expert on economic and industrial policies, a Director of Alex Pol Institute. He held the position of a Director of the International Centre for Policy Studies, a Director of Investments Department and an Advisor to the

Minister in the Ministry of Economic Development and Trade of Ukraine, led the Department of the National Space Agency of Ukraine. Mr. Panchenko acted as a Deputy Chair of National Associations of Commodity Producers Council in Ukraine. His research interests today are in the field of international economic policy coordination, economic nationalism and neo-protectionism policy. He is an author of over 80 articles and books, and a developer of strategies in definite sectors.

Prasad Shakti is an Assistant Professor at the Department of Basic & Applied Science in National Institute of Technology, Arunachal Pradesh, India. His research interests are sample surveys and statistical inference. He has published over twenty research papers in the reputed international /national journals and conferences.

Reznikova Nataliia is an Academy Professor at the Department of World Economy and International Economic Relations, Institute of International Relations of Kyiv Taras Shevchenko National University. She contributed to expert analyses, reviewing, making proposals on draft laws at the Parliament of Ukraine, the Cabinet of Ministers of Ukraine, the National Security and Defense Counsel of Ukraine, Ministry of Foreign Affairs of Ukraine, Ministry of Economic Development and Trade of Ukraine. Her research interests today are in the field of world economy (problems of countries' neo-dependence; global problems of international economic relations). She is an author of over 120 research publications.

Shangodoyin Dahud Kehinde is a Professor at the Department of Statistics, University of Botswan. His area of interest is Data mining in Health, Population, Education and Agriculture, Econometrics, Bayesian Modelling, Multivariate Analysis and Time Series Analysis.

Shanker Rama has completed his Bachelor, Master and PhD in Statistics from Department of Statistics, Patna University, India. Presently, he is working as Professor and Head, Department of Statistics, College of Science, Eritrea Institute of Technology, Eritrea. He is the founding head of Department of Statistics under College of Science at Eritrea Institute of Technology, Asmara, Eritrea. He was also the founding Editor-in-Chief of "Eritrean Journal of Science and Engineering (EJSE)", a biannual Science and Engineering Journal, published from Eritrea Institute of Technology, Eritrea. His research interests include Distribution Theory, Modeling of Lifetime Data, Statistical Inference, Mathematical Demography, Transportation and Assignment Problems. He has 105 research papers published in national and international journals of Statistics, Mathematics and Operations research. He is a member of the international advisory board of many research journals.

Shukla Kamlesh Kumar is currently working as an Associate Professor at Department of Statistics, College of Science, Eritrea Institute of Technology, Asmara, Eritrea since February, 2016. He was awarded PhD in Statistics from Banaras Hindu University, Varanasi, India, and Masters Degree in Statistics (Gold Medalist) in 1997. He has worked as Assistant Professor in Adama Science and Technology University, Ethiopia and has over 14 years of teaching experience in colleges/universities including international experience. He has worked on six

projects in international and national organizations, viz. IIPS, WHO, DST, New Delhi and presented many papers in international and national conferences. He has published over 45 research papers in international and national journals. His research fields of interest are distribution theory, mathematical modelling and migration (demography).

Zeghdoudi Halim is a faculty member at the Department of Mathematics at the University of Badji-Mokhtar, Annaba-Algeria. He received his PhD degree in Mathematics and the highest academic degree (HDR) specializing in Probability and Statistics from Badji-Mokhtar University, Annaba-Algeria. He also did his Post-Doc at Waterford Institute of Technology-Cork Rd, Waterford, Ireland. His research areas are in Actuarial Science, Particles Systems, Dynamics Systems, and Applied Statistics. He has published over 50 research papers in international journals and conferences. Currently he is a member of two editorial boards: Asian Journal of Probability and Statistics and Journal of Advanced Statistics and Probability

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, *Key words* (in bold italics) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).