STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 18, Number 4, December 2017

Editorial Office **ISSN 1234-7655**

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

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

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 563−566

FROM THE EDITOR

With this issue we conclude the year 2017, during which Statistics in Transition new series published 37 articles written by 73 authors from 7 countries. Relatively most frequently – fourteen articles – were devoted to the sampling methods and estimation, or were classified as research papers. There were also substantial number of 61 reviewers involved in the process of evaluation of, and decision-making on publishing each of these articles. We highly appreciate this collaboration and support since the reviewer's generosity in knowledge sharing makes it possible to qualify papers for publication and contributes to the overall quality of the journal. The acknowledgement of this honorary service is attached to this note.

The Statistics in Transition new series continues to expand also in terms of its visibility and recognition. In addition to such prestigious indexation bases as Scopus, RepEc, Index Copernicus, Central and Eastern European Online Library (CEEOL), Central European Journal of Social Sciences and Humanities (CEJSH), EconPapers, ERIH Plus, Google Scholar, InfoBase Index, IC Journals Master List and BazEkon, which have already included SiTns into their systems, several new started recently to monitor our publications during the last year. These are: BASE/Bielefeld Academic Search Engine, Current Index to Statistics (CIS), JournalGuide, JournalTOCs, Keepers Registry, MIAR, OpenAIRE, ProQuest-Summon and WorldCat. We hope to progress along this line also during the year 2018. One of the technical facilitation which will serve to this aim is an improvement in placing the journal's icon on the portal of the Central Statistical Office (www.stat.gov.pl) – it will appear directly in it as the SiTns window, with no need to further digging to search for it.

Out of nine articles published in this issue, four are concerned with topics related to the sampling and estimation; two are the so-called research papers; the next two are of research communicates, and one is a post-conference paper.

Paper by **Kumari Priyanka** and **Richa Mittal,** ***New approaches using exponential type estimator with cost modelling for population mean on successive waves*** treats on sampling over successive waves taking into account the fact that the ancillary information may also be subjected to the time lag (between two successive waves). For this, the authors propose new approaches to estimate population mean (over two successive waves) using four exponential ratio type estimators. The properties of the proposed estimators have been elaborated theoretically, including the optimum rotation rate and cost models to minimize the total cost of the survey design over two successive waves. The prevalence of using the proposed estimators over well-known existing estimators has been shown, and simulation techniques allowed corroborating the theoretical results.

In the next paper, ***Bayesian estimation of measles vaccination coverage under ranked set sampling*** **Radhakanta Das, Vivek Verma, Dilip C. Nath** discuss the problem of estimating an unknown population proportion p (of a certain population characteristic in a dichotomous population) using the data collected through ranked set sampling (RSS) strategy. Assuming that the proportion p is a random quantity (not fixed) and that the prior density of p belongs to the family of Beta distributions, the authors propose a Bayes estimator of p under squared error loss function. The performance of the proposed RSS-based Bayes estimator is compared with that of the corresponding classical version estimator based on maximum likelihood principle. The proposed procedure is used to estimate measles vaccination coverage probability among the children of age group 12-23 months in India using the real-life epidemiological data from National Family Health Survey-III.

Marcin Szymkowiak, Andrzej Młodak, Łukasz Wawrowski in the paper on ***Mapping poverty at the level of subregions in Poland using indirect estimation*** present the results of estimation of the poverty indicator at the level of subregions in Poland (NUTS 3), using the small area estimation methodology – specifically, the EBLUP estimator based on the Fay-Herriot model – applied to data from the European Survey on Income and Living Conditions (EU-SILC). By optimally choosing covariates derived from sources unaffected by random errors, the authors obtained results with satisfactory precision. However, they are aware of the fact that the efficiency of their approach needs to be verified accounting for specific characteristics of the social and economic situations in the areas of interest.

In the paper ***A new estimator of mean using double sampling,*** **Kalyan Rao Vadlamudi, Stephen A. Sedory, Sarjinder Singh** consider the problem of estimation of population mean of a study variable by making use of first-phase sample mean and first-phase sample median of the auxiliary variable at the estimation stage. The proposed new estimator of the population mean is compared to the sample mean estimator, ratio estimator and the difference type estimator for the fixed cost of the survey by using the concept of two-phase sampling. The magnitude of the relative efficiency of the proposed new estimator has been investigated through simulation study.

The research articles section starts with **Devendra Kumar's** and **Mansoor Rashid Malik's** paper ***Relations for moments of progressively Type-II right censored order statistics from Erlang-truncated exponential distribution***, in which the authors establish some new recurrence relations for the single and product moments of progressively Type-II right censored order statistics from the Erlang-truncated exponential distribution. These relations generalize those established by Aggarwala and Balakrishnan for standard exponential distribution. These recurrence relations enable computation of mean, variance and covariance of all progressively Type-II right censored order statistics for all sample sizes in a simple and efficient manner. By using these relations, the authors tabulate the

means and variances of progressively Type-II right censored order statistics of the Erlang-truncated exponential distribution.

Housila P. Singh, Swarangi M. Gorey in the paper ***A generalized randomized response model*** discuss properties of the suggested generalized version of the Gjestvang and Singh (2006) model. They show that the randomized response models due to Warner (1965), Mangat and Singh (1990), Mangat (1994) and Gjestvang and Singh (2006) are members of the proposed RR model. The conditions are obtained in which the suggested RR model is more efficient than several other models (Warner's model, Mangat's and Singh's model, and Gjestvang's and Singh's model). The results of the study are supported by a numerical illustration.

The next section is opened by the paper by **Małgorzata Kalbarczyk, Agata Miazga** and **Anna Nicińska, *The inter-country comparison of the cost of children maintenance using housing expenditure*** devoted to a comparative study of the cost of maintenance of children between selected countries. Specifically, an analysis of the equivalence scales in Austria, Italy, Poland and France was conducted using data from the European Income and Living Condition (EU-SILC) for mono- and duo-parental households, for the first and second child. The four countries share common European cultural context, yet differ with respect to social environment, in particular to family policy. The results are consistent with other studies, which also showed that the cost of the first child is higher than that of a later child. The scale values are not the same across all the countries involved, with the highest cost observed in Italy and the lowest in Poland.

Andrzej Szymański's and **Agnieszka Rossa's** paper ***Improvement of fuzzy mortality models by means of algebraic methods*** starts with an overview of models of forecasting of mortality, including their historical development, while distinguishing between the so-called static or stationary models and dynamic models. The authors propose a new class of fuzzy mortality models based on a fuzzy version of the Lee-Carter model, the essential idea of which is to focus on representing a membership function of a fuzzy number as an element of C*-Banach algebra, combined with Ishikawa (1997) proposed foundations of the fuzzy measurement theory and termed C*-measurement. The authors use the Hilbert space of quaternion algebra as an introduction to the mortality models. This approach is still under further research.

This issue concludes with an article based on a presentation at the ***Multivariate Statistical Analysis 2016 – Łódź.*** The paper by **Alina Jędrzejczak** and **Jan Kubacki*, Estimation of small area characteristics using multivariate Rao-Yu model*** discusses advantages of the EBLUP estimation based on multivariate Rao-Yu model, involving both autocorrelated random effects between areas and sampling errors. The authors demonstrate them next by providing the estimation of incomes and expenditures of Polish households using data from the Polish Household Budget Survey and administrative registers. The

calculations were performed using the packages for R-project environment. Direct estimates were performed using the WesVAR software, and the precision of the direct estimates was determined using a balanced repeated replication (BRR) method. However, the authors consider their analysis as requiring further comparisons between the Rao-Yu method and dynamic models, panel econometric models and nonlinear models.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 567

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, December 2017
Vol. 18, No. 4, pp. 569–587, DOI 10.21307/stattrans-2017-001

NEW APPROACHES USING EXPONENTIAL TYPE ESTIMATOR WITH COST MODELLING FOR POPULATION MEAN ON SUCCESSIVE WAVES

Kumari Priyanka[1], **Richa Mittal**[2]

ABSTRACT

The key and fundamental purpose of sampling over successive waves lies in the varying nature of study character, it so may happen with ancillary information if the time lag between two successive waves is sufficiently large. Keeping the varying nature of auxiliary information in consideration, modern approaches have been proposed to estimate population mean over two successive waves. Four exponential ratio type estimators have been designed. The properties of proposed estimators have been elaborated theoretically including the optimum rotation rate.Cost models have also been worked out to minimize the total cost of the survey design over two successive waves. Dominances of the proposed estimators have been shown over well-known existing estimators. Simulation algorithms have been designed and applied to corroborate the theoretical results.

Key words: Successive sampling, Exponential type estimators, Dynamic ancillary information Population mean, Bias, Mean squared error, Optimum rotation rate.
Mathematics Subject Classification: 62D05.

1. Introduction

Real life facts always carry varyingnatures which are time dependent. In such circumstances where facts change over a period of time, one time enquiry may not serve the purpose of investigation since statistics observed previously contain superannuated information which may not be good enough to be used after a long period of time. Therefore surveys are being designed sophistically to make sure no possible error gets a margin to escape at least in terms of design. For this longitudinal surveys are considered to be best since in longitudinal surveys, facts are investigated more than once i.e. over the successive waves, Also a frame is

[1] Department of Mathematics, Shivaji College University of Delhi, India-110027.
E-mail: priyanka.ism@gmail.com.
[2] Department of Mathematics, Shivaji College University of Delhi, India-110027.
E-mail: sovereignricha@gmail.com.

provided for reducing the cost of survey by a partial replacement of sample units in sampling over successive waves.

Jessen (1942) is considered to be the pioneer for observing dynamics of facts over a long period of time through partial replacement of sample units over successive waves. The approach of sampling over successive waves has been made more fruitful by using twisted and novel ways to consider extra information along with the study character. Enhanced literature has been made availableby Patterson (1950), Narain (1953), Eckler (1955), Sen(1971, 1972, 1973), Gordon (1983), Singh et al. (1991), Arnab and Okafor (1992), Feng and Zou (1997), Biradar and Singh (2001), Singh and Singh (2001), Singh (2005), Singh and Priyanka (2006, 2007, 2008),Singh and Karna (2009), Singh and Prasad (2010), Singh et al. (2011), Singh et al. (2013), Bandyopadhyay and Singh(2014), Priyanka and Mittal (2014), Priyanka et al. (2015), Priyanka and Mittal (2015a, 2015b)etc.

It has been theoretically established that, in general, the linear regression estimator is more efficient than the ratio estimator except when the regression line y on x passes through the neighbourhood of the origin; in this case the efficiencies of these estimators are almost equal. Also in many practical situations where the regression line does not pass through the neighbourhood of the origin, in such cases the ratio estimator does not perform as good as the linear regression estimator. Here exponential type estimators play a vital role in increasing the precision of the estimates.Motivated with this idea we are aspired to develop unexampled estimators for estimating population mean over two successive waves applying the concept of exponential type ratio estimators. In this line of work, an attempt has been made to consider the dynamic nature of ancillary information also because as the time passes by, not only the nature of study variable changes but the nature of ancillary information also varies with respect to time in many real life phenomenon where time lag is very large between two successive waves.

For example, in a social survey one may desire to observe the number of females human trafficked from a particular region, the number of girls child birth may serve as ancillary information which is completely dynamic over a period of 8-10 years of time span. Similarly in a medicinal survey one may be interested to record the number of survivors from a cancerous disease, here the number of successfully tested drugs for the disease may not sustain to be stable over a period of 10 or 20 years. Likewise, in an economic survey the government may like to record the labor force, the total number of graduates in country may serve as an ancillary character to the study character but it surely inherent dynamic nature over a period of 5 or 10 years. Also in a tourism related survey, one may seek to record the total income (profit) from tourism in a particular country or state. In this kind of survey, total number of tourists visiting to the concerned place may be considered as the auxiliary information as communications and transportations services have emerged drastically to enhance the commutation of people from one place to another.

So such situations cannot be tackled considering the ancillary character to be stable since doing so will affect the final findings of the survey. Keeping the drawback of such flaws in consideration, this work deals in bringing modern approaches for estimating population mean over two successive waves.Four estimators have been habituated with a fine amalgamation of completely known dynamic ancillary information with exponential ratio type estimators.Their properties including optimum rotation rate and a model for optimum total cost have been proposed and discussed. Also detailed empirical illustrations have been done by doing a comparison of proposed estimators with well-known existing estimators in the literature of successive sampling. Simulation algorithms have been devised to make the proposed estimators work in practical environment efficiently.

2. Survey Design and Analysis

2.1. Sample Structure and Notations

Let $U = (U_1, U_2, \ldots, U_N)$ be the finite population of N units, which has been sampled over two successive waves. It is assumed that size of the population remains unchanged but values of units change over two successive waves. The character under study be denoted by x (y) on the first (second) waves respectively. It is assumed that information on an ancillary variable $z_1 (z_2)$ dynamic in nature over the successive waves with completely known population mean $\bar{Z}_1 (\bar{Z}_2)$ is readily available on both the successive waves and positively correlated to x and y respectively.Simple random sample (without replacement) of n units is taken at the first wave. A random subsample of $m = n\lambda$ units is retained for use at the second wave. Now at the current wave a simple random sample (without replacement) of u= (n-m) = $n\mu$ units is drawn afresh from the remaining (N-n) units of the population so that the sample size on the second wave remains the same. Let μ and $\lambda (\mu + \lambda = 1)$ are the fractions of fresh and matched samples respectively at the second (current) successive wave. The following notations are considered here after:

$\bar{X}, \bar{Y}, \bar{Z}_1, \bar{Z}_2$: Population means of the variables x, y, z_1 and z_2 respectively.

$\bar{y}_u, \bar{z}_u, \bar{x}_m, \bar{y}_m, \bar{z}_1(m), \bar{z}_2(m), \bar{x}_n, \bar{z}_1(n), \bar{z}_2(n)$: Sample mean of respective variate based on the sample sizes shown in suffice.

$\rho_{yx}, \rho_{xz_1}, \rho_{xz_2}, \rho_{yz_1}, \rho_{yz_2}, \rho_{z_1z_2}$: Correlation coefficient between the variables shown in suffices.

$S_x^2, S_y^2, S_{z_1}^2, S_{z_2}^2$: Population mean squared of variables x, y, z_1 and z_2 respectively.

2.2 Design of the Proposed Estimators $\check{T}_{ij}\,(i, j=1, 2)$

For estimating the population mean $\bar{Y}$ at the current wave, two sets of estimators have been proposed. The first set of estimators is based on sample of size u drawn afresh at current occasion and is given by

$$\check{T}_u = \{t_{1u}, t_{2u}\}, \tag{1}$$

where $$t_{1u} = \bar{y}_u \left(\frac{\bar{Z}_2}{\bar{z}_2(u)} \right) \tag{2}$$

$$t_{2u} = \bar{y}_u \exp\left(\frac{\bar{Z}_2 - \bar{z}_2(u)}{\bar{Z}_2 + \bar{z}_2(u)} \right) \tag{3}$$

The second set of estimators is based on sample of size m common to both occasion and is

$$\check{T}_m = \{t_{1m}, t_{2m}\}, \tag{4}$$

where $$t_{1m} = \bar{y}_m \left(\frac{\bar{x}_n}{\bar{x}_m} \right) \exp\left(\frac{\bar{Z}_2 - \bar{z}_2(m)}{\bar{Z}_2 + \bar{z}_2(m)} \right) \tag{5}$$

$$t_{2m} = \bar{y}_m^* \left(\frac{\bar{x}_n^*}{\bar{x}_m^*} \right) \tag{6}$$

where $\bar{y}_m^* = \bar{y}_m \exp\left(\frac{\bar{Z}_2 - \bar{z}_2(m)}{\bar{Z}_2 + \bar{z}_2(m)} \right)$, $\bar{x}_m^* = \bar{x}_m \exp\left(\frac{\bar{Z}_1 - \bar{z}_1(m)}{\bar{Z}_1 + \bar{z}_1(m)} \right)$ and

$\bar{x}_n^* = \bar{x}_n \exp\left(\frac{\bar{Z}_1 - \bar{z}_1(n)}{\bar{Z}_1 + \bar{z}_1(n)} \right)$.

Hence, considering the convex combination of the two sets $\check{T}_u$ and $\check{T}_m$, we have the final estimators of the population mean $\bar{Y}$ on the current occasion as

$$\check{T}_{ij} = \varpi_{ij} t_{iu} + (1 - \varpi_{ij}) t_{jm}; (i, j=1, 2) \tag{7}$$

where $(t_{iu}, t_{jm}) \in \check{T}_u \times \check{T}_m$ and ϖ_{ij} are suitably chosen weights so as to minimize the mean squared error of the estimators $\check{T}_{ij}\,(i, j=1, 2)$.

2.3. Analysis of the estimators $\check{T}_{ij}\left(i, j=1, 2\right)$

2.3.1. Bias and Mean Squared Errors of the Proposed Estimators $\check{T}_{ij}\left(i, j=1, 2\right)$

The properties of the proposed estimators $\check{T}_{ij}\left(i, j=1, 2\right)$ are derived under the following large sample approximations

$\bar{y}_u = \bar{Y}\left(1 + e_0\right)$, $\bar{y}_m = \bar{Y}\left(1 + e_1\right)$, $\bar{x}_m = \bar{X}\left(1 + e_2\right)$, $\bar{x}_n = \bar{X}\left(1 + e_3\right)$, $\bar{z}_2\left(u\right) = \bar{Z}_2\left(1 + e_4\right)$,
$\bar{z}_2\left(m\right) = \bar{Z}_2\left(1 + e_5\right)$, $\bar{z}_1\left(m\right) = \bar{Z}_1\left(1 + e_6\right)$ and $\bar{z}_1\left(n\right) = \bar{Z}_1\left(1 + e_7\right)$ such that $|e_i| < 1\ \forall\ i = 0,...,7$.

The estimators belonging to the sets $\check{T}_u$ and $\check{T}_m\left(i, j=1, 2\right)$ are ratio, exponential ratio, ratio to exponential ratio and chain type ratio to exponential ratio type in nature respectively. Hence they are biased for population mean $\bar{Y}$. Therefore, the final estimators $\check{T}_{ij}\left(i, j=1, 2\right)$ defined in equation (7) are also biased estimators of $\bar{Y}$. The bias $B\left(.\right)$ and mean squared errors $M\left(.\right)$ of the proposed estimators $\check{T}_{ij}\left(i, j=1, 2\right)$ are obtained up to first order of approximations and thus we have following theorems:

Theorem 2.3.1. Bias of the estimators $\check{T}_{ij}\left(i, j=1, 2\right)$ to the first order of approximations are obtained as

$$B\left(\check{T}_{ij}\right) = \varpi_{ij}\, B\left(t_{iu}\right) + \left(1 - \varpi_{ij}\right) B\left(t_{jm}\right);\ (i, j=1,2), \tag{8}$$

where $$B\left(t_{1u}\right) = \frac{1}{u}\bar{Y}\left(\frac{C_{0002}}{\bar{Z}_2^2} - \frac{C_{0101}}{\bar{Y}\,\bar{Z}_2}\right), \tag{9}$$

$$B\left(t_{2u}\right) = \frac{1}{u}\bar{Y}\left(\frac{3}{8}\frac{C_{0002}}{\bar{Z}_2^2} - \frac{1}{2}\frac{C_{0101}}{\bar{Y}\,\bar{Z}_2}\right), \tag{10}$$

$$B\left(t_{1m}\right) = \bar{Y}\left(\frac{1}{m}\left(\frac{C_{2000}}{\bar{X}^2} + \frac{3}{8}\frac{C_{0002}}{\bar{Z}_2^2} - \frac{C_{1100}}{\bar{X}\bar{Y}} - \frac{1}{2}\frac{C_{0101}}{\bar{Y}\bar{Z}_2} + \frac{1}{2}\frac{C_{1001}}{\bar{X}\bar{Z}_2}\right) + \frac{1}{n}\left(\frac{C_{1100}}{\bar{X}\bar{Y}} - \frac{C_{2000}}{\bar{X}^2} - \frac{1}{2}\frac{C_{1001}}{\bar{X}\bar{Z}_2}\right)\right), \tag{11}$$

and

$$B\left(t_{2m}\right) = \bar{Y}\left(\frac{1}{m}\left(\frac{C_{2000}}{\bar{X}^2} - \frac{1}{8}\frac{C_{0020}}{\bar{Z}_1^2} + \frac{3}{8}\frac{C_{0002}}{\bar{Z}_2^2} - \frac{C_{1100}}{\bar{X}\bar{Y}} - \frac{1}{2}\frac{C_{1010}}{\bar{X}\bar{Z}_1} + \frac{1}{2}\frac{C_{1001}}{\bar{X}\bar{Z}_2} + \frac{1}{2}\frac{C_{0110}}{\bar{Y}\bar{Z}_1} - \frac{1}{2}\frac{C_{0101}}{\bar{Y}\bar{Z}_2} - \frac{1}{4}\frac{C_{0011}}{\bar{Z}_1\bar{Z}_2}\right)\right.$$
$$\left. + \frac{1}{n}\left(\frac{1}{8}\frac{C_{002}}{\bar{Z}_1^2} - \frac{C_{2000}}{\bar{X}^2} + \frac{C_{1100}}{\bar{X}\bar{Y}} + \frac{1}{2}\frac{C_{1010}}{\bar{X}\bar{Z}_1} - \frac{1}{2}\frac{C_{1001}}{\bar{X}\bar{Z}_2} - \frac{1}{2}\frac{C_{0101}}{\bar{Y}\bar{Z}_2} + \frac{1}{4}\frac{C_{0011}}{\bar{Z}_1\bar{Z}_2}\right)\right) \tag{12}$$

where $C_{rstq} = E\left[\left(x_i - \bar{X}\right)^r \left(y_i - \bar{Y}\right)^s \left(z_{1i} - \bar{Z}_1\right)^t \left(z_{2i} - \bar{Z}_2\right)^q\right]; \left(r, s, t, q\right) \geq 0$.

Theorem 2.3.2.Mean squared errors of the estimators $\check{T}_{ij}\left(i, j=1, 2\right)$ to the first order of approximations are obtained as

$$M\left(\check{T}_{ij}\right) = \varpi_{ij}^2 M\left(t_{iu}\right) + \left(1 - \varpi_{ij}\right)^2 M\left(t_{jm}\right) + 2\,\varpi_{ij}\left(1 - \varpi_{ij}\right)\mathrm{Cov}\left(t_{iu}, t_{jm}\right); \quad (i,j=1,2) \tag{13}$$

$$\text{where } M\left(t_{1u}\right) = \frac{1}{u} A_1 S_y^2 \tag{14}$$

$$M\left(t_{2u}\right) = \frac{1}{u} A_2 S_y^2 \tag{15}$$

$$M\left(t_{1m}\right) = \left(\frac{1}{m} A_3 + \frac{1}{n} A_4\right) S_y^2 \tag{16}$$

$$M\left(t_{2m}\right) = \left(\frac{1}{m} A_5 + \frac{1}{n} A_6\right) S_y^2 \tag{17}$$

$\mathrm{Cov}\left(t_{iu}, t_{jm}\right)=0$, $A_1 = 2\left(1 - \rho_{yz_2}\right)$, $A_2 = \frac{5}{4} - \rho_{yz_2}$, $A_3 = \frac{9}{4} - 2\rho_{yx} - \rho_{yz_2} + \rho_{xz_2}$,

$A_4 = 2\rho_{yx} - \rho_{xz_2} - 1$,

$A_5 = \frac{5}{2} - 2\rho_{yx} - \rho_{xz_1} + \rho_{xz_2} + \rho_{yz_1} - \rho_{yz_2} - \frac{1}{2}\rho_{z_1z_2}$ and $A_6 = 2\rho_{yx} + \rho_{xz_1} - \rho_{xz_2} - \rho_{yz_1} + \frac{1}{2}\rho_{z_1z_2} - \frac{5}{4}$.

2.3.2. Minimum Mean Squared Errors of the Proposed Estimators $\check{T}_{ij}\left(i, j=1, 2\right)$

Since the mean squared errors of the estimators $\check{T}_{ij}\left(i, j=1, 2\right)$ given in equation (13) are the functions of unknown constants $\varpi_{ij}\left(i, j = 1, 2\right)$, therefore, they are minimized with respect to ϖ_{ij} and subsequently the optimum values of $\varpi_{ij}\left(i, j = 1, 2\right)$ and $M\left(\check{T}_{ij}\right)_{opt.}\left(i, j=1, 2\right)$ are obtained as

$$\varpi_{ij_{opt.}} = \frac{M\left(t_{jm}\right)}{M\left(t_{iu}\right) + M\left(t_{jm}\right)}; (i, j = 1, 2) \tag{18}$$

$$M\left(\check{T}_{ij}\right)_{opt.} = \frac{M\left(t_{iu}\right) . M\left(t_{jm}\right)}{M\left(t_{iu}\right) + M\left(t_{jm}\right)}; \quad \left(i, j = 1, 2\right) \tag{19}$$

Further, substituting the values of the mean squared errors of the estimators defined in equations (14)-(17) in equation (18)-(19), the simplified values of $\varpi_{i j_{opt.}}$ and $M\left(\check{T}_{i j}\right)_{opt.}$ are obtained as

$$\varpi_{11_{opt.}} = \frac{\mu_{11}\left[\mu_{11}\, A_4 - \left(A_3 + A_4\right)\right]}{\left[\mu_{11}^2\, A_4 - \mu_{11}\left(A_3 + A_4 - A_1\right) - A_1\right]} \tag{20}$$

$$\varpi_{12_{opt.}} = \frac{\mu_{12}\left[\mu_{12}\, A_6 - \left(A_5 + A_6\right)\right]}{\left[\mu_{12}^2\, A_6 - \mu_{12}\left(A_5 + A_6 - A_1\right) - A_1\right]} \tag{21}$$

$$\varpi_{21_{opt.}} = \frac{\mu_{21}\left[\mu_{21}\, A_4 - \left(A_3 + A_4\right)\right]}{\left[\mu_{21}^2\, A_4 - \mu_{21}\left(A_3 + A_4 - A_2\right) - A_2\right]} \tag{22}$$

$$\varpi_{22_{opt.}} = \frac{\mu_{22}\left[\mu_{22}\, A_6 - \left(A_5 + A_6\right)\right]}{\left[\mu_{22}^2\, A_6 - \mu_{22}\left(A_5 + A_6 - A_2\right) - A_2\right]} \tag{23}$$

$$M\left(\check{T}_{11}\right)_{opt.} = \frac{1}{n}\frac{\left[\mu_{11}\, B_1 - B_2\right]S_y^2}{\left[\mu_{11}^2\, A_4 - \mu_{11}\, B_3 - A_1\right]} \tag{24}$$

$$M\left(\check{T}_{12}\right)_{opt.} = \frac{1}{n}\frac{\left[\mu_{12}\, B_4 - B_5\right]S_y^2}{\left[\mu_{12}^2\, A_6 - \mu_{12}\, B_6 - A_1\right]} \tag{25}$$

$$M\left(\check{T}_{21}\right)_{opt.} = \frac{1}{n}\frac{\left[\mu_{21}\, B_7 - B_8\right]S_y^2}{\left[\mu_{21}^2\, A_4 - \mu_{21}\, B_9 - A_2\right]} \tag{26}$$

$$M\left(\check{T}_{22}\right)_{opt.} = \frac{1}{n}\frac{\left[\mu_{22}\, B_{10} - B_{11}\right]S_y^2}{\left[\mu_{22}^2\, A_6 - \mu_{22}\, B_{12} - A_2\right]} \tag{27}$$

where

$B_1 = A_1A_4$, $B_2 = A_1A_3 + A_1A_4$, $B_3 = A_3 + A_4 - A_1$, $B_4 = A_1A_6$, $B_5 = A_1A_5 + A_1A_6$, $B_6 = A_5 + A_6 - A_1$, $B_7 = A_2A_4$, $B_8 = A_2A_3 + A_2A_4$, $B_9 = A_3 + A_4 - A_2$, $B_{10} = A_2A_6$ $B_{11} = A_2A_5 + A_2A_6$, $B_{12} = A_5 + A_6 - A_2$ and $\mu_{i j}\left(i, j = 1, 2\right)$ are the fractions of the sample drawn afresh at the current(second) wave.

2.3.3. Optimum Rotation Rate for the Estimators $\check{T}_{i j}\left(i, j=1, 2\right)$

Since the mean squared errors of the proposed estimators $\check{T}_{i j}\left(i, j=1, 2\right)$ are the function of the $\mu_{i j}\left(i, j = 1, 2\right)$, hence to estimate population mean $\bar{Y}$ with maximum precision and minimum cost,an amicable fraction of sample drawn afresh is required at the current wave. For this the mean squared errors of the

estimators $\check{T}_{ij}(i, j=1, 2)$ in equations (24) – (27) have been minimized with respect to $\mu_{ij}(i, j = 1, 2)$. Hence an optimum rotation rate has been obtained for each of the estimators $\check{T}_{ij}(i, j=1, 2)$ and given as:

$$\mu_{11} = \frac{C_2 \pm \sqrt{C_2^2 - C_1 C_3}}{C_1} \tag{28}$$

$$\mu_{12} = \frac{C_5 \pm \sqrt{C_5^2 - C_4 C_6}}{C_4} \tag{29}$$

$$\mu_{21} = \frac{C_8 \pm \sqrt{C_8^2 - C_7 C_9}}{C_7} \tag{30}$$

$$\mu_{22} = \frac{C_{11} \pm \sqrt{C_{11}^2 - C_{10} C_{12}}}{C_{10}} \tag{31}$$

where

$C_1 = A_4B_1,\ C_2 = A_4B_2,\ C_3 = A_1B_1 + B_2B_3,\ C_4 = A_6B_4,\ C_5 = A_6B_5,\ C_6 = A_1B_4 + B_5B_6$
$C_7 = A_4B_7,\ C_8 = A_4B_8,\ C_9 = A_2B_7 + B_8B_9,\ C_{10} = A_6B_{10},\ C_{11} = A_6B_{11}$ and $C_{12} = A_2B_{10} + B_{11}B_{12}$.

The real values of $\mu_{ij}(i, j = 1, 2)$ exist, iff $C_2^2 - C_1 C_3 \geq 0$, $C_5^2 - C_4 C_6 \geq 0$, $C_8^2 - C_7 C_9 \geq 0$, and $C_{11}^2 - C_{10} C_{12} \geq 0$ respectively. For any situation, which satisfies these conditions, two real values of $\mu_{ij}(i, j = 1, 2)$ may be possible, hence to choose a value of $\mu_{ij}(i, j = 1, 2)$, it should be taken care of that $0 \leq \mu_{ij} \leq 1$, all other values of $\mu_{ij}(i, j = 1, 2)$ are inadmissible. If both the real values of $\mu_{ij}(i, j = 1, 2)$ are admissible, the lowest one will be the best choice as it reduces the total cost of the survey. Substituting the admissible value of μ_{ij} say $\mu_{ij}^{(0)}(i, j = 1, 2)$ from equation (28) - (31) in equation (24) - (27) respectively, we get the optimum values of the mean squared errors of the estimators $\check{T}_{ij}(i, j = 1, 2)$ with respect to ϖ_{ij} as well as $\mu_{ij}(i, j = 1, 2)$ which are given as

$$M(\check{T}_{11})^*_{opt.} = \frac{\left[\mu_{11}^{(0)} B_1 - B_2\right] S_y^2}{n\left[\mu_{11}^{(0)2} A_4 - \mu_{11}^{(0)} B_3 - A_1\right]} \tag{32}$$

$$M\left(\check{T}_{12}\right)^{*}_{opt.} = \frac{\left[\mu_{12}^{(0)}\, B_4 - B_5\right] S_y^2}{n\left[\mu_{12}^{(0)2}\, A_6 - \mu_{12}^{(0)}\, B_6 - A_1\right]} \tag{33}$$

$$M\left(\check{T}_{21}\right)^{*}_{opt.} = \frac{\left[\mu_{21}^{(0)}\, B_7 - B_8\right] S_y^2}{n\left[\mu_{21}^{(0)2}\, A_4 - \mu_{21}^{(0)}\, B_9 - A_2\right]} \tag{34}$$

$$M\left(\check{T}_{22}\right)^{*}_{opt.} = \frac{\left[\mu_{22}^{(0)}\, B_{10} - B_{11}\right] S_y^2}{n\left[\mu_{22}^{(0)2}\, A_6 - \mu_{22}^{(0)}\, B_{12} - A_2\right]} \tag{35}$$

3. Cost Analysis

The total cost of survey design and analysis over two successive waves is modelled as:

$$C_T = nc_f + mc_r + uc_s \tag{36}$$

where c_f : The average per unit cost of investigating and processing data at previous (first)wave,

c_r : The average per unit cost of investigating and processing retained data at current wave,

c_s : The average per unit cost of investigating and processing freshly drawndata at current wave.

Remark 3.1: $c_f < c_r < c_s$, when there is a large gap between two successive waves, the cost of investigating a single unit involved in the survey sample should be greater than before (at previous occasion) since as time passes by different commodities (software) and services (human resources, daily wages and conveyance) become expensive so the cost incurring at second occasion increases in a considerable amount. Also the average cost of investigating a retained unit from previous wave should be lesser than investigating a freshly drawn sample unit since survey investigator as well as respondent has some experiences from the previous wave.

Theorem 3.1.1: The optimum total cost for the proposed estimators $\check{T}_{ij}\,(i, j=1, 2)$ is derived as

$$C_T\left(\check{T}_{ij}\right)_{opt.} = n\left\{c_f + c_s + \left(1 - \mu_{ij}^{(0)}\right)\left(c_r - c_s\right)\right\} \forall\ i, j=1, 2 \tag{37}$$

Remark 3.2: The optimum total costsobtained in equation (37) are dependent on the value of n. Therefore, if a suitable guess of n is available, it can be used for obtaining optimum total cost of the survey by above equation. However, in the absence of suitable guess of n, it may be estimated by following Cochran (1977).

4. Efficiency Comparison

To evaluate the performance of the proposed estimators, the estimators $\check{T}_{ij}\,(i, j=1, 2)$ at optimum conditions, are compared with the sample mean estimator $\bar{y}_n$, when there is no matching from previous wave and the estimator $\hat{\bar{Y}}$ due to Jessen (1942) given by

$$\hat{\bar{Y}} = \psi\,\bar{y}_u + (1 - \psi)\,\bar{y}_m^{'}, \tag{38}$$

where $\bar{y}_m^{'} = \bar{y}_m + \beta_{yx}(\bar{x}_n - \bar{x}_m)$, β_{yx} is the population regression coefficient of y on x and ψ is an unknown constant to be determined so as to minimize the mean squared error of the estimator $\hat{\bar{Y}}$. The estimators $\bar{y}_n$ and $\hat{\bar{Y}}$ are unbiased for population mean, therefore variance of the estimators $\bar{y}_n$ and $\hat{\bar{Y}}$ at optimum conditions are given as

$$V(\bar{y}_n) = \frac{1}{n}S_y^2, \tag{39}$$

$$V\left(\hat{\bar{Y}}\right)^*_{opt.} = \left(\frac{1}{2}\left(1 + \sqrt{1 - \rho_{yx}^2}\right)\right)\frac{S_y^2}{n}, \tag{40}$$

and the fraction of sample to be drawn afresh for the estimator $\hat{\bar{Y}}$

$$\mu_J = \frac{1}{1 + \sqrt{1 - \rho_{yx}^2}} \tag{41}$$

The percent relative efficiencies $E_{ij}(M)$ and $E_{ij}(J)$ of the estimator $\check{T}_{ij}\,(i, j=1, 2)$ (under optimum conditions) with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ are respectively given by

$$E_{ij}(M) = \frac{V(\bar{y}_n)}{M\left(\check{T}_{ij}\right)^*_{opt.}} \times 100 \text{ and } E_{ij}(J) = \frac{V\left(\hat{\bar{Y}}\right)^*_{opt.}}{M\left(\check{T}_{ij}\right)^*_{opt.}} \times 100\,(i, j=1, 2). \tag{42}$$

5. Numerical Illustrations and Simulation

5.1. Generalization of empirical study

A generalized study has been done to show the impact of varying ancillary information in enhancing the performance of the proposed estimators $\check{T}_{ij}\,(i, j=1, 2)$. To elaborate the scenario, various choices of correlation coefficients of study and auxiliary variables have been considered. The results obtained have been shown in Table 1.

Table 1. Generalized empirical results while the proposed estimators $\check{T}_{ij}\,(i, j=1, 2)$ have been compared to the estimators $\bar{y}_n$ and $\hat{\bar{Y}}$ for $\rho_{yz_1} = \rho_{yz_2} = \rho_1$ and $\rho_{xz_1} = \rho_{xz_2} = \rho_2$.

$\rho_{z_1z_2} = \rho_{yx} = 0.5$

ρ_2	ρ_1	μ_J	$\mu_{11}^{(0)}$	$\mu_{12}^{(0)}$	$\mu_{21}^{(0)}$	$\mu_{22}^{(0)}$	$E_{11}(M)$	$E_{12}(M)$	$E_{21}(M)$	$E_{22}(M)$	$E_{11}(j)$	$E_{12}(j)$	$E_{21}(j)$	$E_{22}(j)$
0.4	**0.6**	0.53	0.66	0.58	0.44	0.41	119.69	114.58	135.48	128.91	111.67	106.90	126.41	120.28
	0.8	0.53	0.33	0.32	0.42	0.37	197.61	176.31	187.18	166.66	184.38	164.50	174.64	155.50
0.5	**0.6**	0.53	0.61	0.58	0.42	0.41	117.08	114.58	132.04	128.91	109.24	106.90	123.20	120.28
	0.8	0.53	0.33	0.32	0.40	0.37	191.44	176.31	181.18	166.66	178.61	164.50	169.04	155.50
0.6	**0.6**	0.53	0.58	0.58	0.41	0.41	114.58	114.58	128.91	128.91	106.90	106.99	120.28	120.28
	0.8	0.53	0.33	0.32	0.39	0.37	185.89	176.31	175.84	166.66	173.44	164.50	164.06	155.50
0.7	**0.6**	0.53	0.55	0.58	0.40	0.41	112.22	114.58	126.04	128.91	104.70	106.90	117.60	120.28
	0.8	0.53	0.32	0.32	0.38	0.37	180.88	176.31	171.03	166.66	168.76	164.50	159.57	155.50

$\rho_{z_1z_2} = \rho_{yx} = 0.6$

ρ_2	ρ_1	μ_J	$\mu_{11}^{(0)}$	$\mu_{12}^{(0)}$	$\mu_{21}^{(0)}$	$\mu_{22}^{(0)}$	$E_{11}(M)$	$E_{12}(M)$	$E_{21}(M)$	$E_{22}(M)$	$E_{11}(j)$	$E_{12}(j)$	$E_{21}(j)$	$E_{22}(j)$
0.4	**0.6**	0.55	0.87	0.69	0.46	0.44	124.52	121.01	143.54	137.34	112.07	108.91	129.18	123.60
	0.8	0.55	0.29	0.33	0.45	0.40	212.25	188.59	201.85	178.44	191.02	169.73	181.58	160.54
0.5	**0.6**	0.55	0.73	0.69	0.45	0.44	122.31	121.01	139.24	137.34	110.08	108.91	125.36	123.60
	0.8	0.55	0.32	0.33	0.43	0.40	204.53	188.59	193.99	178.44	184.08	169.73	174.59	160.59
0.6	**0.6**	0.55	0.66	0.69	0.44	0.44	119.69	121.01	135.48	137.34	107.72	108.91	121.94	123.60
	0.8	0.55	0.33	0.33	0.42	0.40	197.61	188.59	187.18	178.44	177.85	169.73	168.46	160.54
0.7	**0.6**	0.55	0.61	0.69	0.42	0.44	117.08	121.01	132.04	137.34	105.37	108.91	118.84	123.60
	0.8	0.55	0.33	0.33	0.40	0.40	191.44	188.59	181.18	178.44	172.29	169.73	163.06	160.59

Note: The values for μ_j, $\mu_{11}^{(0)}$, $\mu_{12}^{(0)}$, $\mu_{21}^{(0)}$ and $\mu_{22}^{(0)}$ have been rounded off up to two places of decimal for presentation.

5.2. Generalized study based on correlation coefficients andoptimum total cost model

To validate the proposed cost model, a hypothetical survey design has been assumed in which various choices of correlation coefficient and different input costs have been considered over two successive waves.

Table 2. Optimum total cost of the survey design at the current wave of the proposed estimators $\check{T}_{ij}\,(i, j=1, 2)$

ρ_{yx} =0.5, n=30, C_f = ₹ 50.00, C_r = ₹ 75.00 and C_s = ₹ 80.00

ρ_2	ρ_1	μ_J	$\mu_{11}^{(0)}$	$\mu_{12}^{(0)}$	$\mu_{21}^{(0)}$	$\mu_{22}^{(0)}$	$C_T(J)$	$C_T(11)$	$C_T(12)$	$C_T(21)$	$C_T(22)$
0.5	**0.6**	0.53	0.61	0.58	0.42	0.41	3830.4	3842.7	3837.5	3814.4	3812.8
	0.8	0.53	0.33	0.32	0.40	0.37	3830.4	3799.9	3798.2	3811.2	3806.3
0.6	**0.6**	0.53	0.58	0.58	0.41	0.41	3830.4	3837.5	3837.5	3812.8	3812.8
	0.8	0.53	0.33	0.32	0.39	0.37	3830.4	3799.5	3798.2	3809.3	3806.3
0.7	**0.6**	0.53	0.55	0.58	0.40	0.41	3830.4	3833.4	3837.5	3811.4	3812.8
	0.8	0.53	0.32	0.32	0.38	0.37	3830.4	3798.9	3798.2	3807.7	3806.3

ρ_{yx} =0.6, n=30, C_f = ₹ 50.00, C_r = ₹ 75.00 and C_s = ₹ 80.00

ρ_2	ρ_1	μ_J	$\mu_{11}^{(0)}$	$\mu_{12}^{(0)}$	$\mu_{21}^{(0)}$	$\mu_{22}^{(0)}$	$C_T(J)$	$C_T(11)$	$C_T(12)$	$C_T(21)$	$C_T(22)$
0.5	**0.6**	0.55	0.73	0.69	0.45	0.44	3833.3	3860.9	3854.8	3817.9	3817.0
	0.8	0.55	0.32	0.33	0.43	0.40	3833.3	3798.9	3799.8	3815.5	3810.2
0.6	**0.6**	0.55	0.66	0.69	0.44	0.44	3833.3	3849.9	3854.4	3816.1	3817.0
	0.8	0.55	0.33	0.33	0.42	0.40	3833.3	3833.3	3799.8	3813.2	3810.2
0.7	**0.6**	0.55	0.61	0.69	0.42	0.44	3833.3	3842.7	3854.8	3814.4	3817.0
	0.8	0.55	0.33	0.33	0.40	0.40	3833.3	3833.3	3799.9	3811.2	3810.2

ρ_{yx} =0.5, n=40, C_f = ₹ 50.00, C_r = ₹ 75.00 and C_s = ₹ 80.00

ρ_2	ρ_1	μ_J	$\mu_{11}^{(0)}$	$\mu_{12}^{(0)}$	$\mu_{21}^{(0)}$	$\mu_{22}^{(0)}$	$C_T(J)$	$C_T(11)$	$C_T(12)$	$C_T(21)$	$C_T(22)$
0.5	**0.6**	0.53	0.61	0.58	0.42	0.41	5107.2	5123.7	5116.7	5085.8	5083.8
	0.8	0.53	0.33	0.32	0.40	0.37	5107.2	5066.6	5064.3	5081.5	5075.0
0.6	**0.6**	0.53	0.58	0.58	0.41	0.41	5107.2	5116.7	5116.7	5083.8	5083.8
	0.8	0.53	0.33	0.32	0.39	0.37	5107.2	5066.0	5064.3	5079.1	5075.0
0.7	**0.6**	0.53	0.55	0.58	0.40	0.41	5107.2	5111.2	5116.7	5081.9	5083.8
	0.8	0.53	0.32	0.32	0.38	0.37	5107.2	5062.2	5064.3	5077.0	5075.0

ρ_{yx} =0.6, n=40, C_f = ₹ 50.00, C_r = ₹ 75.00 and C_s = ₹ 80.00

ρ_2	ρ_1	μ_J	$\mu_{11}^{(0)}$	$\mu_{12}^{(0)}$	$\mu_{21}^{(0)}$	$\mu_{22}^{(0)}$	$C_T(J)$	$C_T(11)$	$C_T(12)$	$C_T(21)$	$C_T(22)$
0.5	**0.6**	0.55	0.73	0.69	0.45	0.44	5111.1	5147.9	5139.7	5090.5	5089.3
	0.8	0.55	0.32	0.33	0.43	0.40	5111.1	5065.1	5066.3	5087.3	5080.3
0.6	**0.6**	0.55	0.66	0.69	0.44	0.44	5111.1	5133.3	5139.7	5088.1	5089.3
	0.8	0.55	0.33	0.33	0.42	0.40	5111.1	5066.5	5066.3	5084.2	5080.3
0.7	**0.6**	0.55	0.61	0.69	0.42	0.44	5111.1	5123.7	5139.7	5085.8	5089.3
	0.8	0.55	0.33	0.33	0.40	0.40	5111.1	5066.6	5066.3	5081.5	5080.3

5.3. Monte Carlo Simulation

Monte Carlo simulation has been performed to get an overview of the proposed estimators in practical scenario through considering different choices of n and μ for better analysis.

Following set has been considered for the simulation study

Set I: n = 20, μ = 0.15, (m = 17,u = 3).

5.3.1. Simulation Algorithm

(i) Choose 5000 samples of size n=20 using simple random sampling without replacement on first wave for both the study and auxiliary variable.

(ii) Calculate sample mean $\bar{x}_{n|k}$ and $\bar{z}_{1|k}(n)$ for k =1, 2, - - -, 5000.

(iii) Retain m=17 units out of each $n=20$ sample units of the study and auxiliary variables at the first wave.

(iv) Calculate sample mean $\bar{x}_{m|k}$ and $\bar{z}_{1|k}(m)$ for k= 1, 2, - - -, 5000.

(v) Select u=3 units using simple random sampling without replacement from N-n=31 units of the population for study and auxiliary variables at second (current) wave.

(vi) Calculate sample mean $\bar{y}_{u|k}$ and $\bar{z}_{2|k}(m)$ for k = 1, 2, - - -, 5000.

(vii) Iterate the parameter ϖ from 0.1 to 0.9 with a step of 0.2.

(viii) Iterate ψ from 0.1 to 0.9 with a step of 0.1 within (ix).

(ix) Calculate the percent relative efficiencies of the proposed estimator $\check{T}_{ij}(i, j=1, 2)$ with respect to estimator to $\bar{y}_n$ and $\hat{\bar{Y}}$ as

$$E(\check{T}_{ij}, M) = \frac{\sum_{k=1}^{5000}\left[\check{T}_{ij|k} - \bar{y}_{n|k}\right]^2}{\sum_{k=1}^{5000}\left[\check{T}_{ij|k}\right]^2} \times 100 \text{ and } E(\check{T}_{ij}, J) = \frac{\sum_{k=1}^{5000}\left[\check{T}_{ij|k} - \hat{\bar{Y}}_{|k}\right]^2}{\sum_{k=1}^{5000}\left[\check{T}_{ij|k}\right]^2} \times 100 \; ;(i, j=1, 2), k=1, 2, ..., 5000.$$

Table 3. Simulation Results when proposed estimator $\check{T}_{ij}\,(i, j=1, 2)$ have been compared to $\bar{y}_n$

SET	ϖ_{ij}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
I	$E(\check{T}_{11}, M)$	218.44	244.19	271.66	303.33	333.45	364.67	393.36	419.85	439.71
	$E(\check{T}_{12}, M)$	461.69	514.46	566.60	619.47	665.56	703.63	731.26	746.47	743.76
	$E(\check{T}_{21}, M)$	231.69	260.16	283.97	304.00	306.75	299.67	281.22	256.46	228.66
	$E(\check{T}_{22}, M)$	505.81	562.87	585.67	578.89	529.97	467.34	397.98	334.78	280.42

Table 4. Simulation results when the proposed estimator $\check{T}_{11}$ is compared with the estimator $\hat{\bar{Y}}$

ψ \ ϖ_{11}	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	182.55	153.16	137.82	159.12	205.37	293.91	390.68	499.10	660.27
0.2	209.72	166.19	153.72	179.11	229.31	301.97	423.52	550.37	732.32
0.3	229.76	184.83	170.77	196.28	255.02	336.89	470.10	618.90	818.38
0.4	252.68	205.47	188.91	216.26	278.49	376.42	523.01	679.76	908.90
0.5	278.45	227.24	209.50	239.44	304.72	411.04	574.21	748.40	994.88
0.6	303.72	249.08	229.98	261.05	333.80	449.52	625.62	813.86	1085.2
0.7	327.71	270.01	249.29	281.45	362.61	482.77	674.92	882.95	1164.0
0.8	350.68	287.18	267.02	300.12	385.81	515.96	718.68	947.91	1240.0
0.9	366.09	300.07	280.72	315.76	404.82	541.84	752.31	995.79	1298.8

Table 5. Simulation results when the proposed estimator $\check{T}_{12}$ is compared with the estimator $\hat{\bar{Y}}$

ψ \ ϖ_{12}	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	427.90	343.13	312.59	375.43	472.41	634.04	887.31	1166.2	1504.7
0.2	475.65	373.99	347.93	407.95	521.89	684.91	962.09	1278.4	1656.4
0.3	516.76	407.22	383.11	444.72	572.07	757.85	1047.3	1398.8	1833.6
0.4	556.92	445.77	416.14	480.80	614.70	829.45	1138.9	1507.8	1992.2
0.5	596.58	480.22	449.92	516.79	655.86	884.66	1221.4	1616.5	2127.7
0.6	627.38	510.46	477.32	544.38	696.06	935.55	1286.6	1703.9	2248.6
0.7	650.44	531.63	496.40	562.32	724.39	961.78	1335.1	1767.5	2313.1
0.8	663.15	538.25	507.13	569.47	733.17	977.60	1353.2	1808.1	2343.7
0.9	654.95	532.58	504.31	567.16	727.56	972.56	1343.2	1799.6	2322.1

Table 6: Simulation results when the proposed estimator $\check{T}_{21}$ is compared with the estimator $\hat{\bar{Y}}$

ψ \ ϖ_{21}	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	194.16	162.27	147.18	169.77	218.56	290.84	417.32	531.93	700.09
0.2	226.59	179.69	165.54	192.53	246.92	327.07	458.49	592.66	787.86
0.3	244.65	197.87	182.29	208.92	270.87	359.27	503.59	657.30	870.32
0.4	259.87	210.16	193.93	220.93	284.73	384.16	536.82	700.23	931.24
0.5	266.55	215.90	199.93	226.91	291.38	391.08	548.24	717.8	951.67
0.6	261.83	212.32	196.50	222.57	286.35	384.76	536.17	701.89	930.06
0.7	244.56	200.80	185.50	209.60	269.14	363.34	504.47	663.99	879.04
0.8	224.81	183.61	170.18	191.41	246.68	331.87	458.85	606.02	800.17
0.9	200.96	162.94	151.25	171.04	220.55	296.36	409.10	542.09	714.32

Table 7: Simulation results when the proposed estimator $\check{T}_{22}$ is compared with the estimator $\hat{\bar{Y}}$

ψ \ ϖ_{22}	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	471.41	378.45	347.07	417.54	523.23	700.91	985.84	1294.1	1660.1
0.2	528.24	417.18	384.72	451.31	575.64	760.29	1077.0	1413.8	1830.2
0.3	341.14	430.65	402.61	465.7	597.23	788.46	1112.2	1462.5	1926.8
0.4	527.18	420.54	394.10	451.06	576.66	772.73	1081.0	1434.4	1888.2
0.5	485.30	388.39	363.79	412.40	530.19	709.26	990.21	1314.4	1732.8
0.6	425.04	341.89	318.71	361.92	465.0	622.42	863.99	1149.0	1510.5
0.7	355.08	291.67	267.62	306.67	392.04	532.08	733.74	974.43	1290.4
0.8	299.61	244.69	227.33	255.76	329.85	444.27	610.04	811.62	1072.3
0.9	250.56	202.66	188.41	213.37	275.45	370.39	508.81	677.67	893.24

7. Interpretations of Results

7.1. Results from Generalized Empirical Study

a) The optimum values $\mu_{11}^{(0)}, \mu_{12}^{(0)}, \mu_{21}^{(0)}$ and $\mu_{22}^{(0)}$ exist for almost each combination of correlation coefficients. For increasing values of correlation of study and ancillary information, the values $\mu_{11}^{(0)}, \mu_{12}^{(0)}, \mu_{21}^{(0)}$ and $\mu_{22}^{(0)}$ decrease, which in accordance with Sukhatme et al (1984.)

b) As the correlation between study and ancillary information is increased, the percent relative efficiencies increase and the proposed estimators perform better than $\bar{y}_n$ and $\hat{\bar{Y}}$.

c) The proposed estimators provide a lesser fraction of fresh sample drawn afresh as compared to the estimator $\hat{\bar{Y}}$ for almost every considered choice of correlation coefficients.

d) The estimator $\check{T}_{21}$ performs best in terms of percent relative efficiency and the estimator $\check{T}_{22}$ performs best in terms of sample drawn afresh at current occasion.

e) As a result, it is also observed that the proposed estimators are working efficiently even for low and moderate correlation values of study and dynamic auxiliary variable on both the occasions.

7.2. Results based on Cost Analysis

a) Theoretically, it is expected that if auxiliary and study variable possess high correlation then this should contribute in reducing the total cost of survey. It is quite evident from the cost analysis that the optimum total cost of the survey decreases for increasing correlation between study and ancillary character.

b) The estimator $\check{T}_{21}$ and $\check{T}_{22}$ requires the least total cost for the survey at the current occasion and they both are good in terms of efficiency as well.

7.3 Simulation Results

a) From Table 3to Table 7, it can be seen that the proposed estimators $\check{T}_{ij}\,(i, j=1, 2)$ are efficient over $\bar{y}_n$ and $\hat{\bar{Y}}$ for the considered set.

b) Also in simulation study, it is observed that the estimator $\check{T}_{22}$ is most efficient over the estimators $\bar{y}_n$ and $\hat{\bar{Y}}$ for the considered set.

8. Ratiocination

The entire detailed generalized and simulation studies attest that accompanying dynamic ancillary character with the study character certainly serves the purpose in long lag of two successive waves. The proposed estimators $\check{T}_{ij}\,(i, j=1, 2)$ prove to be worthy in terms of precisionas compared to the estimators $\bar{y}_n$ and estimator due to Jessen (1942). The minute observation suggest that the estimators $\check{T}_{21}$ and $\check{T}_{22}$ are providing approximately same fraction of sample to be drawn afresh at the current occasion but the total cost of survey is least for the estimator $\check{T}_{22}$ and $\check{T}_{21}$ is best in terms of efficiency . Since both the

estimators $\check{T}_{21}$ and $\check{T}_{22}$ are better than the sample mean estimator and the estimator due to Jessen (1942) but for little amount of precision, the cost of survey cannot be put on stake, therefore $\check{T}_{22}$ may be regarded as best in terms cost and $\check{T}_{21}$ may be regarded best in terms of precision. Hence according to the requirement of survey, one is free to choose any of the estimators out of $\check{T}_{21}$ and $\check{T}_{22}$. Hence the proposed estimators are recommended to the survey statisticians for their practical use.

Acknowledgement

Authors highly appreciate the valuable remarks made by honourable reviewers for enhancing the quality of the article. Authors are also thankful to UGC, New Delhi, India for providing the financial assistance to carry out the present work. Authors also acknowledge the free access of the data from Statistical Abstracts of the United States that is available on internet.

REFERENCES

BANDYOPADHYAY, A., SINGH, G. N., (2014). On the use of two auxiliary variables toimprove the precision of estimate in two-occasion successive sampling, Int. J. Math. Stat., 15, pp.73–88.

ECKLER, A. R., (1955). Rotation Sampling, Ann. Mathe. Stat. 26, pp. 664–685.

SEN, A. R., (1971). Successive sampling with two auxiliary variables. Sankhya, B 33, 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). Theory and application of sampling on repeated occasions with several auxiliary variables. Biometrics, 29, pp. pp. 381–385.

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., PRIYANKA, K., (2006). On the use of chain-type ratio to difference estimator in successive sampling. I. J. App. Math. Statis., 5, pp. 41–49.

SINGH, G. N., PRIYANKA, K., (2007). Effect of non-response on current occasion in search of good rotation patterns on successive occasions, Statist. Trans. New Ser. 8, pp. 273–292.

SINGH, G. N., PRIYANKA, K., (2008). Search of good rotation patterns to improve theprecision of estimates at current occasion, Comm. Stat. – Theo. Meth. 37, pp. 337–348.

SINGH, G. N., KARNA, J. P., (2009). Search of efficient rotation patterns in presence of auxiliary information in successive sampling over two occasions. Statis. Tran. N. Ser. 10, pp. 59–73.

SINGH, G. N., PRASAD, S., (2010). Some estimators of population mean in two-occasion rotation patterns. Asso. Adva. Mode. Sim. Tech. Enter., 12, pp. 25–44.

SINGH, G. N., PRIYANKA, K., PRASAD, S., SINGH, S., KIM, J., M., (2013). A class of estimators for estimating the population variance in two occasion rotation patterns,Comm. Statis. App. Meth., 20, pp. 247–257.

PATTERSON, H. D., (1950). Sampling on successive occasions with partial replacement of units, J. Royal Statis. Soci., 12, 241–255.

SINGH, H. P., KUMAR, S., BHOUGAL, S., (2011). Multivariate ratio estimation in presence of non-response in successive sampling. J. Statis. Theo. Prac., 5, pp. 591–611.

GORDON, L., (1983). Successive sampling in finite populations, The Ann. Stat., 11, pp. 702–706.

PRIYANKA, K., MITTAL, R., (2014). Effective rotation patterns for median estimation in successive sampling. Statis. Trans., 15, pp. 197–220.

PRIYANKA, K., MITTAL, R., KIM, J., M., (2015). Multivariate Rotation Design for Population Mean in sampling on Successive Occasions.Comm. Statis. Appli. Meth., 22, pp. 445–462.

PRIYANKA, K., MITTAL, R., (2015a). Estimation of Population Median in Two-OccasionRotation. J. Stat. App. Prob. Lett., 2, pp. 205–219.

PRIYANKA, K., MITTAL, R., (2015 b). A Class of Estimators for Population Median in two Occasion Rotation Sampling. HJMS, 44, pp. 189–202.

SUKHATME, P. V., SUKHATME, B. V., SUKHATME, S., ASOK, C., (1984). Sampling Theory of Surveys with applications. Ames, Iowa State University Press and New Delhi India, Indian Society of Agricultural Statistics.

ARNAB, R., OKAFOR, F. C., (1992). A note on double sampling over two occasions, Paki. J. Stat., 8, pp. 9–18.

NARAIN, R. D., (1953). On the recurrence formula in sampling on successive occasions, J. Indi.Soci. Agri. Stat., 5, pp. 96–99.

JESSEN, R. J., (1942). Statistical investigation of a sample survey for obtaining farm facts, IowaAgri. Exp. Stat.Road Bull., 304, pp. 1–104.

BIRADAR, R. S., SINGH, H. P., (2001). Successive sampling using auxiliary information on both occasions, Cal. Statist. Assoc. Bull., 51, pp. 234–251.

SINGH, R., SINGH, N., (1991). Imputation methods in two-dimensional survey.Recent Advances in Agricultural Statistics Research, Wiley Eastern Ltd, New Delhi,.

SINGH, V. K., SINGH, G. N., SHUKLA, D., (1991). An efficient family of ratio cum difference type estimators in successive sampling over two occasions. J. Sci. Res., 41, pp. 149–159.

FENG, S., ZOU, G., (1997). Sample rotation method with auxiliary variable, Comm. Stat.- Theo. Meth., 26, pp. 1497–1509.

COCHRAN, W., G., (1977). Sampling Techniques, John Wiley & Sons, New Delhi.

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 589–608, DOI 10.21307/stattrans-2017-002

BAYESIAN ESTIMATION OF MEASLES VACCINATION COVERAGE UNDER RANKED SET SAMPLING

Radhakanta Das[1], **Vivek Verma**[2], **Dilip C. Nath**[3]

ABSTRACT

The present article is concerned with the problem of estimating an unknown population proportion p, say, of a certain population characteristic in a dichotomous population using the data collected through ranked set sampling (RSS) strategy. Here, it is assumed that the proportion p is not fixed but a random quantity. A Bayes estimator of p is proposed under squared error loss function assuming that the prior density of p belongs to the family of Beta distributions. The performance of the proposed RSS-based Bayes estimator is compared with that of the corresponding classical version estimator based on maximum likelihood principle. The proposed procedure is used to estimate measles vaccination coverage probability among the children of age group 12-23 months in India using the real-life epidemiological data from National Family Health Survey-III.

Key words: Bayes estimator, maximum likelihood principle, square error loss, risk function and immunization coverage.

1. Introduction

Ranked Set Sampling (RSS) was first introduced by McIntyre (1952). This is an alternative method of sampling procedure that is used to achieve the greater efficiency in estimating the population characteristics. Generally the most appropriate situation for employing RSS is where the exact measurement of sampling units is expansive in time or effort; but the sample units can be readily ranked either through subjective judgement or *via* the use of relevant concomitant variables. The most basic version of RSS is balanced RSS where the same number of observations is drawn corresponding to each judgement order statistic. In order to draw a balanced ranked set sample of size n, first an integer s is chosen such that $n = ms$, for some positive

[1]Department of Statistics, Presidency University, 86/1 College Street Kolkata, 700073, West Bengal, India. E-mail: radhakanta.stat@presiuniv.ac.in

[2]Corresponding Author. Department of Statistics, Gauhati University, Guwahati, 781014, Assam, India. E-mail: viv_verma456@yahoo.com

[3]Department of Statistics, Gauhati University, Guwahati, 781014, Assam, India. E-mail: dilipc.nath@gmail.com

integer m. Then we select s^2 units from the population at random and the units are divided into sets of s units each. Within each set, s units are ranked according to the characteristic of interest by judgement or with the help of one or more auxiliary variables. From the i^{th} set $(i = 1,2,\cdots,s)$ we observe the actual measurement corresponding to only the i^{th} ordered unit in that set. This entire procedure, which may be called a *cycle*, is repeated m times independently to obtain a ranked set sample of size $n = ms$.

Let $X_{[i]j}$ denote the quantified i^{th} judgement order statistic from the j^{th} cycle. Thus, the sampling scheme yields the following ranked set sample of size n.

$$\begin{array}{c} X_{[1]1},\cdots,X_{[1]j},\cdots,X_{[1]m} \\ \vdots \qquad\quad \vdots \qquad\quad \vdots \\ X_{[i]1},\cdots,X_{[i]j},\cdots,X_{[i]m} \\ \vdots \qquad\quad \vdots \qquad\quad \vdots \\ X_{[s]1},\cdots,X_{[s]j},\cdots,X_{[s]m} \end{array} \tag{1.1}$$

It is obvious that the observations within each row of above observation matrix are independently and identically distributed (iid), and the observations within any column are independently but not identically distributed. To acquire depth in theories and logistics of RSS methodology one can go through the book by Chen et al.(2004).

In the present investigation we assume that the variable of interest is binary; that is, there are only two possible outcomes, generally called success (denoted as 1) and failure (denoted as 0). Thus, the study variable is supposed to follow Bernoulli distribution with success probability p $(0 < p < 1)$, say. Here, the ranking of s binary observations in each set, where there are only 0 and 1 runs in the series, is done systematically as discussed by Terpstra and Nelson (2005). For instance, suppose $s = 4$ and the observations are, say, $X_1 = 1; X_2 = 0; X_3 = 1$, and $X_4 = 0$. Then, a possible ordered arrangement of the observations might be (X_2, X_4, X_1, X_3), or (X_4, X_2, X_1, X_3) or (X_2, X_4, X_3, X_1) or (X_4, X_2, X_3, X_1). But for the sake of uniqueness we take the arrangement (X_2, X_4, X_1, X_3) where the suffix of X in each run is in increasing order and hence we get the ordered statistics as $X_{(1)} = X_2$; $X_{(2)} = X_4$; $X_{(3)} = X_1$ and $X_{(4)} = X_3$. The same systematic rule can easily be extended in the ranking of a polytomous variable also. For a binary population the success probability p can be viewed as a proportion of individuals possessing certain known characteristic in the population. In classical inference on a population proportion, the ranked set sampling with binary data has already been introduced and used by many researchers like, among others, Lacayo et al. (2002), Kvam (2003), Terpstra (2004), Terpstra

and Liudahl (2004), Chen et al. (2005, 2006, 2007, 2009), Terpstra and Nelson (2005), Terpstra and Miller (2006), Chen (2008), Gemayel etal.,(2012) used RSS for auditing purpose, Wolfe (2010, 2012) and Zamanzade and Mahdizadeh (2017, 2017) discussed application of RSS to air quality monitoring. Jozani and Mirkamali (2010, 2011) used ranked set sample for binary data in the context of control charts for attributes. In earlier works, estimation of p using ranked set samples is based on the assumption that the parameter p is an unknown but a fixed quantity. But there may be situations where some prior knowledge on p may be available in terms of its changing pattern over time or with respect to other factors, which amounts to treat p as a random quantity. In this article a Bayesian estimation of p in the domain of ranked set sample is considered.

We organize the paper in the following way. In section 2, a Bayes estimator of the population proportion is proposed. As a natural competitor of the proposed estimator, a classical version estimator based on maximum likelihood principle is discussed in section 3. Section 4 provides an efficiency comparison of the estimators in terms of risk under square error loss function. In section 5 the proposed procedure is used to estimate measles vaccination coverage probability among the children of age group 12-23 months in India using the real-life epidemiological data from National Family Health Survey-III. Lastly, section 6 gives a brief concluding remark.

2. Bayes Estimators of p

Let X be the variable of interest assumed to follow Bernoulli (p) distribution with p being the success probability. It has been found in the literature (e.g. Stokes (1977)) that the use of a single concomitant variable for ranking is effective regardless of whether the association of the concomitant variable of interest is positive or negative. Suppose, after applying judgement ranking made on the basis of a readily available auxiliary variable, say Y, we have a ranked set sample $\{X_{[i]j}, i = 1(1)s, j = 1(1)m\}$ of size $n = ms$, where $X_{[i]j}$ denotes the quantified i^{th} judgement order statistics in the j^{th} cycle. It can easily be justified that, for each $i = 1(1)s$, the observations in the i^{th} ranking group $X_{[i]1}, \cdots, X_{[i]j}, \cdots, X_{[i]m}$ constitute a simple random sample (SRS) of size m from Bernoulli distribution with success probability denoted by $p_{[i]}$, say. So, for each $i = 1(1)s$, $p_{[i]}$ represents the probability of assuming the value 1 (which corresponds to success) for the i^{th} judgement order statistic $X_{[i]1}$. Immediately we get the following result.

Result 2.1: *Suppose an observation with a higher judgement order is more likely to*

be a 'success'. Then, we have

$$p_{[i]} = I_p(s-i+1,i), \ \textit{for each } i = 1,2,\cdots,s \tag{2.1}$$

and

$$\frac{1}{s}\sum_{i=1}^{s} p_{[i]} = p, \tag{2.2}$$

where $I_x(a,b)$, $x \in (0,1)$, *is the standard incomplete beta integral given by*

$$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1}dt.$$

The above result is standard (see Tepstra (2004)) and hence omitted.

Note 2.1: If an observation with a lower judgement order is more likely to be a success, then, for every $i = 1,2,\cdots,s$, that

$$p_{[i]} = I_p(i, s-i+1) \tag{2.3}$$

and (2.2) also holds in this case.

In this section the proportion parameter p is assumed to be a random variable and the randomness is quantified in terms of suitable prior density, say $\tau(p)$ of p over the interval $[0,1]$. Here, we derive a Bayesian estimator of p by incorporating the available prior information on p along with the information provided by the ranked set sample data. By the virtue of ranked set sampling all the observations $X_{[i]j}, \forall i = 1(1)s, \ \forall j = 1(1)m$ are independent. Let us define the variables

$$Z_i = \sum_{j=1}^{m} X_{[i]j}, \ \forall i = 1(1)s.$$

Obviously, the variables $Z_1, Z_2, \cdots, Z_s$ are independently distributed as $Z_i \sim$ Binomial $(m, p_{[i]})$. For each i, $1 \le i \le s$, $p_{[i]}$ is a function of the basic parameter p, so we denote it as $p_{[i]}(p)$. With this notation one can easily write the likelihood function of p, given the ranked set sample data $\mathbf{z} = (z_1, z_2, \cdots, z_s)$ as

$$\begin{aligned} L(p|\mathbf{z}) &= \prod_{i=1}^{s} P[Z_i = z_i | p_{[i]}(p)] \\ &= \prod_{i=1}^{s} \binom{m}{z_i} [p_{[i]}(p)]^{z_i} \ [1 - p_{[i]}(p)]^{(m-z_i)} \end{aligned} \tag{2.4}$$

where $\mathbf{z}$ is a particular realization of the random vector $Z = (Z_1, Z_2, \cdots, Z_s)$. Thus, the posterior density of p, given $\mathbf{Z} = \mathbf{z}$, with respect to the prior $\tau(p)$ for p is given by

$$\begin{aligned} h(p|\mathbf{z}) &= \frac{L(p|\mathbf{z})\tau(p)}{\int_0^1 L(p|\mathbf{z})\tau(p)dp} \\ \Leftrightarrow h(p|\mathbf{z}) &\propto L(p|\mathbf{z})\tau(p) \\ \Leftrightarrow h(p|\mathbf{z}) &\propto \prod_{i=1}^{s}[p_{[i]}(p)]^{z_i}\,[1-p_{[i]}(p)]^{(m-z_i)}\tau(p). \end{aligned} \tag{2.5}$$

The information regarding unknown parameter is upgraded in the light of the observed data and is quantified through the posterior distribution $h(p|\mathbf{z})$ w.r.t. the prior $\tau(p)$ and hence any statistical inference regarding p is made on the basis of its posterior distribution given the ranked set sample data. Here, the posterior distribution does not have any standard form. In such a situation, to make any statistical inference on p one should use Monte Carlo simulation technique which provides a great deal of computational facilities. According to this method, a sufficiently large number, say N of observations are drawn at random independently from the posterior distribution $h(p|\mathbf{z})$ and let it be denoted as $p^{(1)}, p^{(2)}, \cdots, p^{(N)}$. Then the posterior mean and variance of p can be approximated as

$$\begin{aligned} E(p|\mathbf{z}) &= \int_0^1 p h(p|\mathbf{z})dp \\ &\simeq \frac{1}{N}\sum_{j=1}^{N} p^{(j)} \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} V(p|\mathbf{z}) &= \int_0^1 \{p - E(p|\mathbf{z})\}^2 h(p|\mathbf{z})dp \\ &\simeq \frac{1}{N}\sum_{j=1}^{N}[p^{(j)}]^2 - \left[\frac{1}{N}\sum_{j=1}^{N} p^{(j)}\right]^2 \end{aligned} \tag{2.7}$$

Thus, under square error loss function the Bayes estimate ($\hat{p}_B$) of p w.r.t. the prior $\tau(p)$ is given by the mean of the posterior distribution $h(p|\mathbf{z})$, that is, for sufficiently

large N,

$$\hat{p}_B \simeq \frac{1}{N}\sum_{j=1}^{N} p^{(j)}. \tag{2.8}$$

Alternative Bayes Esimator of p:

An alternative Bayes estimator of p can easily be constructed as the average of the Bayes estimators of $p_{[i]}$'s by assuming that the probabilities $p_{[1]}, p_{[2]}, \cdots, p_{[s]}$ are all unknown parameters, although all of them are the functions of the basic parameter p, satisfying the relation (2.2). Suppose $p_{[1]}, \cdots, p_{[s]}$ are independently and identically distributed with common prior density given below.

$$\tau(\theta) = \frac{1}{\mathscr{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\ ,\ 0<\theta<1,\ \alpha>0,\ \beta>0. \tag{2.9}$$

Then, under squared error loss function the Bayes estimator of $p_{[i]}$, based on Z_i, can easily be obtained as (see Ferguson (2014))

$$\hat{p}^*_B[i] = \frac{Z_i+\alpha}{m+\alpha+\beta},\ \text{ for } i=1,..s. \tag{2.10}$$

After having the estimators $\hat{p}^*_B[i]; i=1,2,\cdots,s$, we are in a position to construct, by virtue of the relation (2.2), a Bayesian estimator of p as

$$\hat{p}^*_B = \frac{1}{s}\sum_{i=1}^{s}\hat{p}^*_B[i] \tag{2.11}$$

which, by using (2.10), takes the form

$$\hat{p}^*_B = \frac{m\bar{X}+\alpha}{m+\alpha+\beta}, \tag{2.12}$$

where $\bar{X}\left(=\frac{1}{ms}\sum_{i=1}^{s}\sum_{j=1}^{m}X_{[i]j}\right)$ is the grand mean of ms sample observations. Note that, under square error loss, the risk function of $\hat{p}_B^*$ is given as

$$\begin{aligned}
R_{\hat{p}_B^*}(p) &= E(\hat{p}_B^* - p)^2 \\
&= E\left(\frac{m\bar{X}+\alpha}{m+\alpha+\beta} - p\right)^2 \\
&= (m+\alpha+\beta)^{-2} E\left\{(m\bar{X} - mp) + \alpha - p(\alpha+\beta)\right\}^2 \\
&= (m+\alpha+\beta)^{-2}\left[\frac{1}{s^2}\sum_{i=1}^{s} mp_{[i]}(1-p_{[i]}) + \{\alpha - p(\alpha+\beta)\}^2\right] \\
&= \frac{m}{(m+\alpha+\beta)^2 s^2}\sum_{i=1}^{s} p_{[i]}(1-p_{[i]}) + \frac{\{\alpha - p(\alpha+\beta)\}^2}{(m+\alpha+\beta)^2}. \qquad (2.13)
\end{aligned}$$

3. Estimator of p based on Maximum Likelihood Principle

Here, we briefly describe a classical version estimator of p based on the maximum likelihood (ML) principle. For this we first assume that the proportion parameter p is an unknown fixed number lying between 0 and 1. Now, all the observations in the ranked set sample are independently distributed, the likelihood function of p based on the given ranked set sample $\mathbf{X} = \mathbf{x}$ can be expressed as

$$\begin{aligned}
L_1(p|\mathbf{x}) &= \prod_{i=1}^{s}\prod_{j=1}^{m}[p_{[i]}(p)]^{x_{[i]j}}[1-p_{[i]}(p)]^{1-x_{[i]j}} \\
&= \prod_{i=1}^{s}[p_{[i]}(p)]^{z_i}\,[1-p_{[i]}(p)]^{m-z_i} \\
&= \prod_{i=1}^{s}\{I_p(s-i+1,i)\}^{z_i}\{I_{1-p}(i,s-i+1)\}^{m-z_i}. \qquad (3.1)
\end{aligned}$$

Equivalently, the log-likelihood function of p is given as

$$l(p|\mathbf{x}) = \log_e L_1(p|\mathbf{x}) = \sum_{i=1}^{s} z_i \log_e I_p(s-i+1,i) + \sum_{i=1}^{s}(m-z_i)\log_e I_{1-p}(i,s-i+1),$$

and hence the likelihood equation for determining MLE of p is obtained as

$$\begin{aligned} & \frac{d}{dp} l(p|\mathbf{x}) = 0 \\ \Leftrightarrow \quad & \sum_{i=1}^{s} \frac{z_i b(p|s-i+1,i)}{I_p(s-i+1,i)} = \sum_{i=1}^{s} \frac{(m-z_i) b(1-p|i,s-i+1)}{I_{1-p}(i,s-i+1)} \end{aligned} \tag{3.2}$$

where $b(x|\alpha,\beta)$ represents the probability density function of $Beta(\alpha,\beta)$ distribution. Due to the complicated nature of the above likelihood equation it is difficult to get an explicit solution for p and hence the RSS-based MLE of p does not have any closed form.

As an alternative way out we can obtain an estimate of p by indirectly using the maximum likelihood principle. For this we first consider the probabilities $p_{[1]}, p_{[2]}, \cdots, p_{[s]}$ as the unknown parameters, although all of them are the functions of the basic parameter p. Then we determine the maximum likelihood estimates of those parameters separately and substitute these estimates in the relation (2.2) of Result 2.1 to get an estimate of p. Given the ranked set sample $\mathbf{X} = \mathbf{x}$, the likelihood function of the parameters $p_{[1]}, p_{[2]}, \cdots, p_{[s]}$ is written as

$$\begin{aligned} L_1(p_{[1]}, p_{[2]}, \cdots, p_{[s]}|\mathbf{x}) &= \prod_{i=1}^{s} \prod_{j=1}^{m} [p_{[i]}]^{x_{[i]j}} [1-p_{[i]}]^{1-x_{[i]j}} \\ &= \prod_{i=1}^{s} [p_{[i]}]^{z_i} \, [1-p_{[i]}]^{m-z_i} \end{aligned} \tag{3.3}$$

and the corresponding log-likelihood function is given by

$$l(p_{[1]}, p_{[2]}, \cdots, p_{[s]}|\mathbf{x}) = \sum_{i=1}^{s} z_i \log_e p_{[i]} + \sum_{i=1}^{s} (m-z_i) \log_e (1-p_{[i]}).$$

Thus, by solving s maximum likelihood equations, $\frac{\partial}{\partial p_{[i]}} l(p_{[1]}, p_{[2]}, \cdots, p_{[s]}|\mathbf{x}) = 0$, for $i = 1, 2, \cdots, s$, we easily get the MLE of $p_{[i]}$ as

$$\hat{p}_{[i]} = \frac{Z_i}{m}, \;\; i = 1, 2, \cdots, s.$$

Then, after replacing $p_{[i]}$'s by $\hat{p}_{[i]}$'s in the relation (2.2) we get an estimate of p as

$$\hat{p}_M = \frac{1}{s} \sum_{i=1}^{s} \hat{p}_{[i]}. \tag{3.4}$$

Note 3.1: It is easy to argue that the ML estimates $\hat{p}_{[1]}, \hat{p}_{[2]}, \cdots, \hat{p}_{[s]}$ are statistically independent as the variables Z_i's are independently distributed. Again, substituting the value $\frac{Z_i}{m}$ of $\hat{p}_{[i]}$ in the equation (3.4), the estimate $\hat{p}_M$ can be shown to be identical with the overall mean of the given ranked set sample. It is also readily verified that $\hat{p}_M$ is an unbiased estimator of p.

4. Comparison Between $\hat{p}_B$, $\hat{p}_B^*$ and $\hat{p}_M$

The goal of this section is to compare the estimators of p derived in sections 2 and 3. Since the posterior mean, by definition, minimizes the Bayes risk under squared error loss function, it is not surprising that a Bayes estimator of an unknown parameter is often superior to the corresponding MLE in respect of mean squared error (MSE). However, MLE neither requires any specification of prior distribution for the parameter nor it involves any particular loss function. Thus, the comparison should be made on the basis of a criterion which does not bother about the particular nature of prior information regarding unknown parameter. However, as MSE of an estimator can be regarded as risk under squared error loss, one can use risk function for comparison purpose. The expressions for risk functions of the estimators are described below. Under square error loss the risk of $\hat{p}_B$ is given as

$$R_{\hat{p}_B}(p) = E(\hat{p}_B - p)^2, \tag{4.1}$$

which cannot be further simplified analytically. On the other hand, the risk of $\hat{p}_M$ has a theoretical expression obtained as

$$\begin{aligned} R_{\hat{p}_M}(p) &= E(\hat{p}_M - p)^2 \\ &= V\left(\frac{1}{s}\sum_{i=1}^{s}\hat{p}_{[i]}\right) \\ &= \frac{1}{ms^2}\sum_{i=1}^{s} p_{[i]}(1-p_{[i]}), \end{aligned} \tag{4.2}$$

after using the fact that $\hat{p}_{[i]}$'s are independent with $V(\hat{p}_{[i]}) = \frac{1}{m}p_{[i]}(1-p_{[i]})$.

Result 4.1: The risk function $R_{\hat{p}_M}(p)$ of the estimator $\hat{p}_M$ is symmetric around '$p = \frac{1}{2}$'.

Proof: By (2.1) we rewrite $R_{\hat{p}_M}(p)$ as

$$\begin{aligned} R_{\hat{p}_M}(p) &= \frac{1}{ms^2}\sum_{i=1}^{s} I_p(s-i+1,i)[1-I_p(s-i+1,i)] \\ &= \frac{1}{ms^2}\sum_{i=1}^{s} I_p(s-i+1,i)I_{1-p}(i,s-i+1), \text{ for } p\in[0,1]. \end{aligned}$$

Now, for any $\xi \in [0,1]$, we see that

$$\begin{aligned} ms^2 R_{\hat{p}_M}\left(\frac{1}{2}+\xi\right) &= \sum_{i=1}^{s} I_{\frac{1}{2}+\xi}(s-i+1,i)\, I_{\frac{1}{2}-\xi}(i,s-i+1) \\ &= \sum_{j=1}^{s} I_{\frac{1}{2}+\xi}(j,s-j+1)\, I_{\frac{1}{2}-\xi}(s-j+1,j), \quad \text{putting } j=s-i+1 \\ &= \sum_{j=1}^{s} I_{\frac{1}{2}-\xi}(s-j+1,j)\, I_{\frac{1}{2}+\xi}(j,s-j+1) \\ &= ms^2 R_{\hat{p}_M}\left(\frac{1}{2}-\xi\right), \end{aligned}$$

and hence the required proof follows.

In the present situation we compare the performances of the estimators $\hat{p}_B$, $\hat{p}_B^*$ and $\hat{p}_M$ by plotting their risk functions in the same co-ordinate axes. The estimator $\hat{p}_B$ performs uniformly better than the estimator $\hat{p}_M$ if

$$R_{\hat{p}_B}(p) \le R_{\hat{p}_M}(p),$$

for all $p \in [0,1]$ with strict inequality for at least one value of p. Here, we conveniently choose Beta(α,β) distribution as a prior for p, that is,

$$\tau(p) = \{\mathscr{B}(\alpha,\beta)\}^{-1} p^{\alpha-1}(1-p)^{\beta-1}, \;\; 0<p<1, \alpha>0, \beta>0.$$

In the numerical computation we take, in particular, $(s,m) = (3,50), (5,30), (3,100), (5,60)$ and $(\alpha,\beta) = (\frac{1}{2},\frac{1}{2}), (2,2), (\frac{1}{2},3), (3,\frac{1}{2})$. Here, we compute the risk values for the Bayes estimator $\hat{p}_B$ by simulation technique with the help of Metropolis-Hasting's algorithm (given in Appendix) and then plot them over the whole range of p. The plotted risk functions are shown in Figures 1-4 given in Appendix. These figures show that the risk curves corresponding to Bayes estimators $\hat{p}_B$ and $\hat{p}_B^*$ completely lie below the risk curve of $\hat{p}_M$ implying that the proposed Bayes estimators are uniformly better than the estimator based on ML principle so far as the given parametric combinations are concerned. Again it is observed that the risk curves

corresponding to Bayes estimators $\hat{p}_B$ and $\hat{p}_B^*$ are not significantly distinct and hence we can conclude that these two estimators are more or less equally good.

5. Estimation of Measles Vaccination Coverage Probability

In public health related studies, the virus of measles is regarded as highly epidemic and is responsible for severe diseases. According to the Medical Dictionary, the virus of measles infects the lungs at childhood which may cause pneumonia and in older children it can lead to inflammation of the brain, called encephalitis, which can cause seizures and brain damage (Perry and Halsey, 2004). As precautionary measures the proper vaccination is introduced from the very beginning of the childhood to acquire the immunity against measles viruses. According to the Integrated Child Development Services (ICDS) program in India, a child should have received basic vaccinations (BCG, polio, DPT and measles) in the 12-23 months of their age.

Here, our objective is to illustrate the proposed procedures for estimating the vaccination coverage of the measles among the children of age group 12-23 months (the age by which children should have received all basic vaccinations) in India 2005-06. The study data has been taken from the website of the Measure DHS-Demographic and Health Surveys (DHS) (`http://www.measuredhs.com`). DHS provides national and state estimates of fertility, child mortality, the practice of family planning, attention to mother and child and access to services for mothers and children. For this study, data set of National Family Health Survey-III (NFHS-III,2005-2006) for the year 2005-06 of India is considered. Here, the samples of DHS are treated as our population of interest and those children who are in the 12-23 months of their age considered as our study population.

The event of receiving vaccination for a child usually depends on awareness of the child's mother regarding vaccination. The higher the educational qualification of a mother during child bearing period, the higher would be the awareness as expected. Therefore, mother's educational qualification is used as auxiliary variable for ranking purpose in ranked set sampling. The observations are obtained through the following steps.

1. A simple random sample of s^2 units is drawn from the target population and is randomly partitioned into s sets, each having s units.

2. In each of s sets the units are ranked according to the mother's qualification {1 = "No education", 2 = "Primary", 3 = "Secondary", 4 = "Higher"}. The

ranking process could also be based on the individuals' duration (in terms of years) of study. Here, samples in different sets are ranked based on mother's qualification denoted as (1, 2, 3, 4) and duration (in terms of years) of study. Obviously, there is a high chance of having ties. Then in that situation the observations are ordered systematically in the sequence, as discussed by Terpstra and Nelson (2005).

3. From the first set, the unit corresponding to the mother with lowest qualification (or duration of study) is selected. From the second set, the unit corresponding to the mother with the second lowest qualification is selected and so on. Finally, from the s^{th} set, the unit corresponding to the mother with the highest qualification is selected. The remaining $s(s-1)$ sampled units are discarded from the data set.

4. The Steps 1 - 3, called a cycle, are repeated m times to obtain a ranked set sample of size $n = ms$.

Here, in particular, we take $(s,m) = (4,100)$. Corresponding to each selected mother, information regarding whether her child is administrated with measles vaccination or not is collected. Suppose X is the binary response that takes value '1' if the child is vaccinated and '0' otherwise. With this notation we have the sample $\{X_{[1]1}, X_{[1]2}, \cdots, X_{[4]100}\}$ of size 400, where $X_{[r]j}$ takes the values '1' or '0' accordingly as the j^{th} child in the r^{th} ranking class is vaccinated or not. Obviously, $p_{[r]}$ is the proportion, in r^{th} class, of children who received the vaccination and p is the overall proportion of children receiving the vaccine in entire target population. The implementation of the proposed Bayes approach requires assuming the prior distributions of $p_{[r]}$'s. Here we use Beta (α,β) priors with $(\alpha,\beta) = (0.5,0.5),(1,1),(2,2),(5,5),(1,2),(2,1)$ and $(5,3)$. With these parametric combinations we compute the estimates $\hat{p}_B$, $\hat{p}_B^*$ and $\hat{p}_M$. We also calculate the estimated relative risk of Bayes estimators w. r. t. $\hat{p}_M$ defined by

$$\hat{\rho}_{\hat{p}_B} = \frac{\hat{R}(\hat{p}_M)}{\hat{R}(\hat{p}_B)}$$

$$\hat{\rho}_{\hat{p}_B^*} = \frac{\hat{R}(\hat{p}_M)}{\hat{R}(\hat{p}_B^*)}$$

and all computed results are summarized in Table 5.1. From the table, it is observed that the Bayes estimate of the proportion of children receiving measles vaccine is very close to that based on ML approach. Also, both the estimates are very close to the value 58.8%, which is the estimated value of p reported by NFHS-III(2005-

Table 5.1: Estimates of the proportion and Relative Efficiency

Estimate	Bayesian approach							ML
	Symmetric prior				Asymmetric prior			
	$\alpha=0.5$ $\beta=0.5$	$\alpha=1$ $\beta=1$	$\alpha=2$ $\beta=2$	$\alpha=5$ $\beta=5$	$\alpha=1$ $\beta=2$	$\alpha=2$ $\beta=1$	$\alpha=5$ $\beta=3$	
$\hat{p}_B$	0.562	0.562	0.562	0.562	0.562	0.562	0.564	0.578
$\hat{p}_B^*$	0.577	0.576	0.575	0.570	0.570	0.580	0.581	0.578
$\hat{\rho}_{\hat{p}_B}$	2.13	2.02	2.14	2.25	2.37	2.02	2.15	
$\hat{\rho}_{\hat{p}_B^*}$	1.02	1.03	1.05	1.10	1.03	1.04	1.09	

06). It is also clear that the proposed Bayes procedure, especially the estimator $\hat{p}_B$, shows greater efficiency than the corresponding ML based procedure.

6. Concluding Remarks

The present work is concerned with the problem of estimating unknown population proportion p based on ranked set sample (RSS) drawn from a binary population. Since the RSS-based likelihood function of p is complicated, the direct application of Bayes principle (in the context of Bayesian paradigm) or maximum likelihood principle (in connection with classical framework) is not straightforward for estimating p. In Bayesian framework the RSS-based Bayes estimator does not have a simple explicit form even if we choose the simplest distribution, i.e. Uniform$(0,1)$ $(\equiv \text{Beta}(1,1))$ as a possible prior for p. The RSS-based likelihood function of p can easily be expressed in the form of polynomial in p. Thus, under the assumption of Beta(α,β) prior for p, the posterior distribution can be shown, through a routine calculation, to be a mixture of several Beta distributions. Also, the explicit form of the Bayes estimator is not so convenient from the computational point of view. Obviously, the posterior distribution does not belong to the Beta-family and hence Beta(α,β) prior is not a conjugate prior in this case. In fact, there does not exist any conjugate prior in standard form due to complexity in the functional form of RSS-based likelihood of p.

As a natural competitor of the Bayes estimator of p, we have used here a very common estimator indirectly based on maximum likelihood principle used in finding MLEs of intermediate parameters $p_{[1]}, p_{[2]}, \cdots, p_{[s]}$. On the other hand, one can directly use the MLE of p as considered by Tepstra (2004) for comparison purpose. However, this estimator does not exist in a closed form but can be computed through

numerical methods. Since our main focus lies in the Bayesian approach of estimation by incorporating available prior information regarding the parameter of interest. We here consider a commonly used estimator for comparison purpose only.

In Bayesian statistics the selection of the prior distribution is crucial to the analysis of data because the final conclusion depends on this particular choice. In our proposed procedure we have considered the Beta prior due to its important features, *viz.,* proper interpretability according to the model (see Paolino (2001)), less computational complexity of posterior distribution (see Gupta and Nadarajah (2004)), having reasonable reflection of prior uncertainty (see Ferrari and Cribari-Neto (2004)), capability to extend to higher dimensions (see Pham-Gia (1994)), etc. However, the Beta-Binomial conjugate analysis may not be adequately robust. Thus, the precision of the prior is important and the sensitivity analysis regarding the prior is necessary. Keeping these in mind one can carry out a robust Bayesian analysis using non-conjugate priors. One such way out might be the use of Cauchy priors after expressing the likelihood of binary data in terms of its exponentially family form, and this Cauchy-Binomial model for binary data might be more robust (see Fuquene et al. (2008)). Several other robust approaches are also discussed in, among others, Berger et al. (1994) and Wang and Blei (2015). The consideration of robust Bayesian approach is beyond the scope of the present work and will be considered in a separate issue.

Acknowledgements

The first author is thankful to the authority of the Presidency University, Kolkata, for providing the grant under FRPDF scheme. The second author would like to express his deepest gratitude and sincere thanks to the Department of Science & Technology, India (Grant No. - IF130365) for funding. The third author acknowledges the financial support provided by Indian Council of Medical Research (ICMR), New Delhi (Grant No-69/40/2008 ECD-II). The authors would like to thank both of the anonymous reviewers for their valuable comments and suggestions to improve the quality of this manuscript.

REFERENCES

BERGER, J. O., MORENO, E., PERICCHI, L. R., BAYARRI, M. J., BERNARDO, J. M., CANO, J. A., ... and DASGUPTA, A., (1994). An overview of robust Bayesian analysis. *Test*, 3 (1), pp. 5–124.

CHEN, H., (2008). Alternative Ranked set Sample Estimators for the Variance of a Sample Proportion ,*Applied Statistics Research Progress; Nova Publishers*, pp. 35–38.

CHEN, Z., BAI, Z. D., SINHA, B. K., (2004). Ranked Set Sampling: Theory and Applications, Lecture Notes in Statistics, 176, Springer-Verlag, New York.

CHEN, H., STASNY, E. A., WOLFE, D. A., (2005). Ranked Set Sampling for Efficient Estimation of a Population Proportion *Statistics in Medicine*, 24, pp. 3319–3329.

CHEN, H., STASNY, E. A., WOLFE, D. A., (2006). Unbalanced Ranked Set Sampling for Estimating a Population Proportion, *Biometrics*, 62, pp. 150–158.

CHEN, H., STASNY, E. A., WOLFE, D. A., (2007). Improved Procedures for Estimation of Disease Prevalence Using Ranked Set Sampling, *Biometrical Journal*, 49 (4), pp. 530–538.

CHEN H. , STASNY A. E., WOLFE A. D., MACEACHERN N. S., (2009). Unbalanced Ranked Set Sampling for Estimating A Population Proportion Under Imperfect Rankings, *Communications in Statistics - Theory and Methods*, 38 (12), pp. 2116–2125.

FERRARI, S., CRIBARI-NETO, F., (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31 (7), pp. 799–815.

FERGUSON, T. S., (2014). Mathematical statistics: A decision theoretic approach (Vol. 1) *Academic press.*

FUQUENE P.J.A., COOK, J.D., PERICCHI, L. R., (2008). A Case for Robust Bayesian priors with Applications to Binary Clinical Trials, *http://biostats.bepress.com/mdandersonbiostat/paper44*

GEMAYEL, N. M., STASNY, E. A., TACKETT, J. A., WOLFE, D. A., (2012). Ranked set sampling: an auditing application, *Review of Quantitative Finance and Accounting*, 39 (4), pp. 413–422.

GUPTA, A. K., NADARAJAH, S. (Eds.), (2004). Handbook of beta distribution and its applications. *CRC press.*

JOZANI, M.J., MIRKAMALI, S. J., (2010). Improved attribute acceptance sampling plans based on maxima nomination sampling, *Journal of Statistical Planning and Inference*, 140 (9), pp. 2448–2460.

JOZANI, M. J., MIRKAMALI, S. J., (2011). Control charts for attributes with maxima nominated samples, *Journal of Statistical Planning and Inference*, 141, pp. 2386–2398.

KVAM, H. P., (2003). Ranked Set Sampling Based on Binary Water Quality Data With Covariates, *Journal of Agricltural, Biological and Environmental Statistics*, 8 (3), pp. 271–279.

LACAYO, H., NEERCHAL, N. K., SINHA, B. K., (2002). Ranked Set Sampling from a Dichotomous Population, *Journal of Applied Statistical Science*, 11(1), pp. 83–90.

MCINTYRE, G. A., (1952). A method for unbiased selective sampling, using ranked sets, *Australian Journal of Agricultural Research*, 3, pp. 385–390.

NFHS-III., (2005-2006). National Family Health Survey", *International Institute for Population Sciences*, Bombay.

PAOLINO, P., (2001). Maximum likelihood estimation of models with beta-distributed dependent variables, *Political Analysis*, pp. 325–346.

PERRY, T. R., HALSEY, A. N., (2004). The Clinical Significance of Measles: A Review" *The Journal of Infectious Diseases,* 189, Suppl 1: S4–16.

PHAM-GIA, T., (1994). Value of the beta prior information, *Communications in Statistics - Theory and Methods*, 23 (8), pp. 2175–2195.

STOKES, L. S., (1977). Ranked Set Sampling with Concomitant variables, *Communications in Statistics - Theory and Methods*, 6 (12), pp. 1207–1211.

TERPSTRA, J. T., (2004). On Estimating a Population Proportion via Ranked Set Sampling, *Biometrical Journal*, 2, pp. 264–272.

TERPSTRA, J. T., LIUDAHL, L. A., (2004). Concomitant-based Rank Set Sampling Proportion Estimates, *Statistics in Medicine*, 23, pp. 2061–2070.

TERPSTRA, J. T., MILLER, Z. A., (2006). Exact Inference for a Population Proportion Based on a Ranked Set Sample, *Communications in Statistics: Simulation and Computation*, 35 (1), pp. 19–27.

TERPSTRA, J. T., NELSON, E. J., (2005). Optimal Rank Set Sampling Estimates for a Population Proportion, *Journal of Statistical Planning and Inference*, 127, pp. 309–321.

WANG, C., BLEI, D. M., (2015). A general method for robust Bayesian modeling, arXiv preprint arXiv:1510.05078.

WOLFE, D. A., (2010). Ranked set sampling, *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (4), pp. 460–466.

WOLFE, D. A., (2012). Ranked set sampling: its relevance and impact on statistical inference, *ISRN Probability and Statistics*.

ZAMANZADE, E., MAHDIZADEH, M., (2017). A more efficient proportion estimator in ranked set sampling, *Statistics & Probability Letters*, pp. 28–33.

ZAMANZADE, E., MAHDIZADEH, M., (2017). Estimating the population proportion in pair ranked set sampling with application to air quality monitoring, *Journal of Applied Statistics*, pp. 1–12.

APPENDIX

Algorithm: Metropolis-Hastings algorithm

The purpose of the Metropolis-Hastings (MH) algorithm is to simulate samples from a probability distribution by utilizing the full joint density function and (independent) proposals distributions corresponding to each variable of interest. The steps of Algorithm mainly consist of three components and are given below:

Initialize $x^{(0)} \uparrow q(x)$
Initialize the sample value for each random variable (this value is often sampled from the variable's prior distribution).
for iteration $i = 1, 2, \cdots$ do

Propose: $x^{cand} \uparrow q(x^{(i)}|x^{(i-1)})$
Generate a proposal (or a candidate) sample x^{cand} *from the proposal distribution* $q(x^{(i)}|x^{(i-1)})$

Acceptance Probability:

$$\alpha\ (x^{cand}|x^{(i-1)}) = Min\ \left\{1,\ \frac{q(x^{(i)}|x^{cand})\pi(x^{cand})}{q(x^{cand}|x^{(i-1)})\pi(x^{(i-1)})}\right\}$$

$u \sim$ Uniform $(u; 0, 1)$
Compute the acceptance probability via the acceptance function $\alpha\ (x^{cand}|x^{(i-1)})$ *based on the proposal distribution and the full joint density* $\pi(.)$

if $u < \alpha$ **then**
Accept the proposal: $x^{(i)} \leftarrow x^{cand}$
else
Reject the proposal: $x^{(i)} \leftarrow x^{(i-1)}$
end if
Accept the candidate sample with probability α*, the acceptance probability, or reject it with probability* $1 - \alpha$

end for

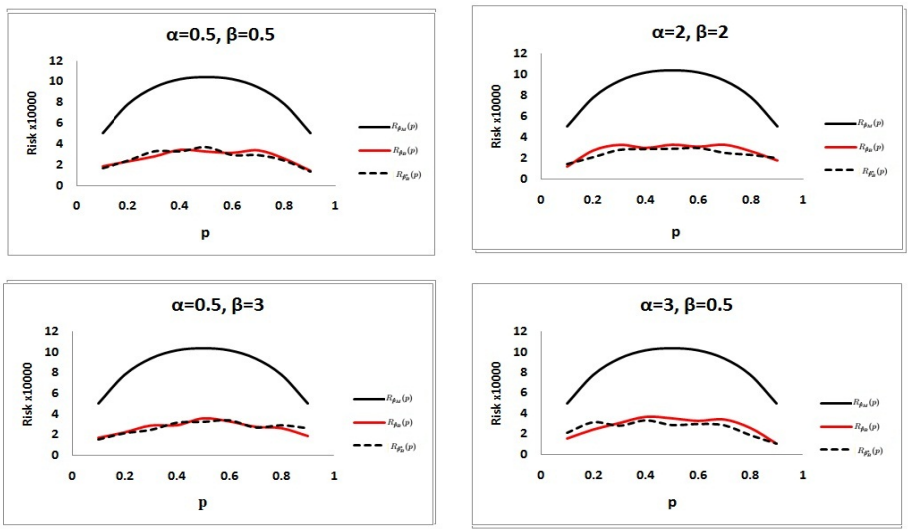


Figure 1: Risk curves for the estimators $\hat{p}_M$, $\hat{p}_B$ and $\hat{p}_B^*$ when $s = 3, m = 50$ at different choices of α and β

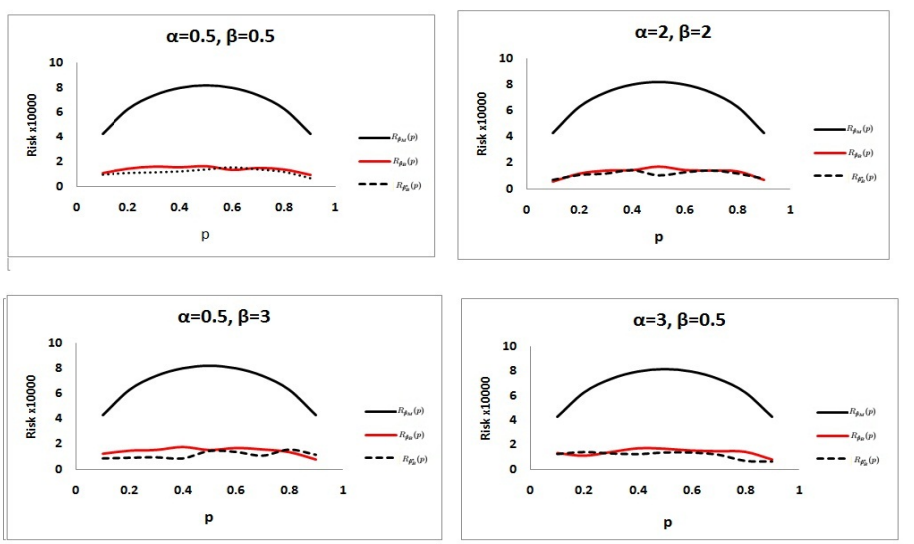


Figure 2: Risk curves for the estimators $\hat{p}_M$, $\hat{p}_B$ and $\hat{p}_B^*$ when $s = 5, m = 30$ at different choices of α and β

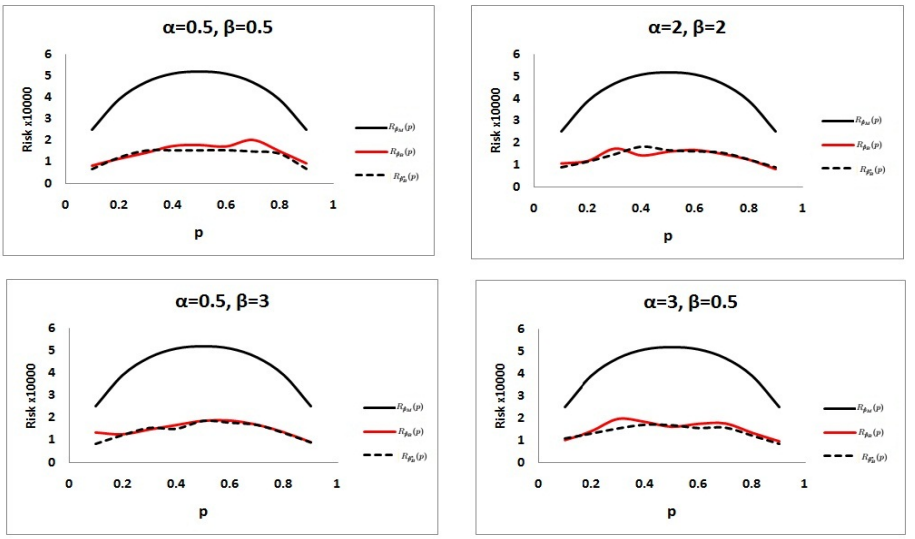


Figure 3: Risk curves for the estimators $\hat{p}_M$, $\hat{p}_B$ and $\hat{p}_B^*$ when $s = 3, m = 100$ at different choices of α and β

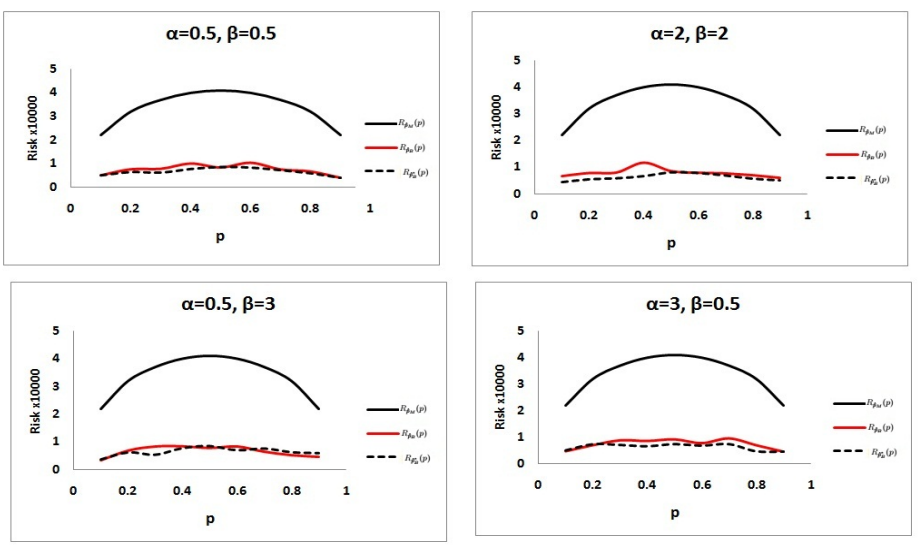


Figure 4: Risk curves for the estimators $\hat{p}_M$, $\hat{p}_B$ and $\hat{p}_B^*$ when $s = 5, m = 60$ at different choices of α and β

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 609–635, DOI 10.21307/stattrans-2017-003

MAPPING POVERTY AT THE LEVEL OF SUBREGIONS IN POLAND USING INDIRECT ESTIMATION

Marcin Szymkowiak[1], **Andrzej Młodak**[2], **Łukasz Wawrowski**[3]

ABSTRACT

The European Survey on Income and Living Conditions (EU–SILC) is the basic source of information published by CSO (the Central Statistical Office of Poland) about the relative poverty indicator, both for the country as a whole and at the regional level (NUTS 1). Estimates at lower levels of the territorial division than regions (NUTS 1) or provinces (NUTS 2, also called 'voivodships') have not been published so far. These estimates can be calculated by means of indirect estimation methods, which rely on information from outside the subpopulation of interest, which usually increases estimation precision. The main aim of this paper is to show results of estimation of the poverty indicator at a lower level of spatial aggregation than the one used so far, that is at the level of subregions in Poland (NUTS 3) using the small area estimation methodology (SAE), i.e. a model–based technique – the EBLUP estimator based on the Fay–Herriot model. By optimally choosing covariates derived from sources unaffected by random errors we can obtain results with adequate precision. A territorial analysis of the scope of poverty in Poland at NUTS 3 level will be also presented in detail[4]. The article extends the approach presented by Wawrowski (2014).

Key words: EU–SILC, poverty, direct estimation, indirect estimation, EBLUP, Fay–Herriot model.

1. Introduction

In modern statistics there is a growing demand for information concerning the quality of life, especially poverty. This demand is necessitated by a number of social policy strategies aimed at reducing social and economic disparities. Such activities

[1]Department of Statistics, Poznań University of Economics and Business, Center for Small Area Estimation, Statistical Office in Poznań. E-mail: m.szymkowiak@ue.poznan.pl

[2]Center for Small Area Estimation, Statistical Office in Poznań, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz. E-mail: a.mlodak@stat.gov.pl

[3]Department of Statistics, Poznań University of Economics and Business, Center for Small Area Estimation, Statistical Office in Poznań. E-mail: lukasz.wawrowski@ue.poznan.pl

[4]The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of the Central Statistical Office of Poland.

are more and more often initiated and implemented by agencies of local government. To achieve these objectives, they require relevant statistical data characterising the diversification of the population in terms of living conditions at a relatively low level or for smaller subpopulations (for instance lower level units of territorial division) and functional areas. Because most regular statistical surveys are based on relatively small samples of the population and do not guarantee the required quality when processed using traditional methods of estimation, more advanced methods of small area estimation should be applied.

The literature devoted to the analysis of poverty using small area estimation techniques is very rich. One comprehensive source of information regarding the use of SAE methods for poverty measurement is the book by Pratesi et al. (2016). This monograph provides a review of SAE methods for poverty mapping and demonstrates many applications of SAE techniques in real-life case studies. In particular, the authors pay special attention to advanced methods and techniques which have been developed recently in the survey data analysis literature devoted to SAE. This includes, for instance, issues related to small area estimation modelling and robustness, spatio-temporal modelling of poverty and small area estimation of the distribution function of income and inequalities. A comprehensive description of different small area estimation techniques for poverty can also be found in many recently published articles, see for instance, Molina and Rao (2010), Molina et al. (2014), Guadarrama et al. (2016). Poverty has also been at the center of interest at many conferences devoted to small area estimation methodology (Jyväskylä 2005, Pisa 2007, Elche 2009, Trier 2011, Bangkok 2013, Poznan 2014, Santiago 2015, Maastricht 2016 and Paris 2017)[5]. All of this indicates the importance of the problem of poverty and its status as one of the main trends in small area estimation methodology.

The article describes an experimental study aimed at exploring the possible use of SAE tools to obtain efficient estimates of the poverty indicator for Polish regions for the purpose of a regular production of reliable poverty maps, which would provide an important source of knowledge about the spatial variation of poverty and inform decision making in cohesion policies, see. Bedi et al. (2007).

Poverty mapping in Poland has been developing intensively in recent years. Apart from the study described in this paper, Polish statistics was engaged in exploring possibilities of estimating some of the Laeken indicators of poverty in the period 2005–2012, included in the Europe 2020 strategy:

- at–risk–of–poverty and social exclusion (AROPE),

[5]A full list of conferences on SAE can be found on the website `http://sae2017.ensai.fr/useful-links-2/`.

- at–risk–of–poverty threshold (ARPT),
- indicator of low work intensity in households (LWI),
- indicator of severe material deprivation (SMD).

Work in this field was conducted as part of subprojects created under the Operational Programme – Technical Assistance financed by the European Commission. These experimental studies were aimed at estimating these indicators at various territorial levels (NUTS1, NUTS 2 and NUTS 3). A variety of statistical methods were tested: direct and indirect estimators, the Fay–Herriot model, a synthetic taxonomy-based measure used as an auxiliary variable in estimation, etc. The results of such studies indicate that relatively efficient estimation is possible at NUTS 1 and NUTS 2 levels but for NUTS 3 units it is much more problematic due to the lower quality of the most efficient model with optimally chosen auxiliary variables, which are strongly correlated with the target indicator and as weakly as possible with one another. One possible cause of this problem is the fact that an increase in the number of observations severely affects the linear correlation, i.e. one can observe a decrease in the correlation coefficient as the number of observations increases. Some attempts were made to assess to what extent increasing the EU-SILC sample would improve estimation precision.

This paper presents an attempt to estimate the poverty rate[6]. It is defined as the percentage of people whose equivalised disposable income (after social transfer) is below the at–risk–of–poverty threshold set at 60% of the national median equivalised disposable income, CSO (2012). This definition is used in the European Survey on Income and Living Conditions. Data collected in the EU–SILC are the basic source of information published by CSO about this indicator both for the country as a whole and at the macro-regional level (NUTS 1). However, nowadays users of statistical data expect reliable estimates of this indicator for lower levels of spatial units. To meet this demand, CSO's Department of Social Surveys and Living Conditions started cooperation with the World Bank and the Centre for Small Area Estimation in order to test various techniques of small area estimation for creating poverty maps at the level of subregions (NUTS 3). The main purpose of this cooperation was to address issues concerning the selection of covariates for the appropriate model to estimate the poverty rate.

The present paper presents results of an analysis and calculations from the methodological, experimental, study. Estimates at lower levels of territorial division than regions (NUTS 1) or provinces (NUTS 2, in Poland called "voivodships") can be

[6]Throughout this paper the term 'scope of poverty' is used interchangeably with the poverty indicator and at–risk–of–poverty rate.

calculated by means of indirect estimation methods. They are based on information from outside the subpopulation of interest, which usually increases estimation precision. Since the estimation process used in these techniques is model–based, the indirect estimation methodology poses a challenge for official statistics in many countries. This paper tries to address this challenge by presenting an attempt to estimate the poverty indicator at a lower level of spatial aggregation than the one used so far, that is at the level of subregions in Poland (NUTS 3).

In the literature devoted to small area estimation methodology different poverty indicators can be estimated at area level under the design-based, model-assisted or model-based approach, see. Pratesi et al. (2016). In the simplest case, direct estimates are produced only on the basis of information from one sample survey, while in the model-assisted or model-based approach the quality and accuracy of survey estimates can be improved by using auxiliary variables and appropriate models. In most cases, auxiliary information comes from censuses, administrative registers or from other surveys. There are many SAE methods which can be applied to estimate different indicators of poverty. They include direct estimation, the EBLUP based on the Fay-Herriot area-level model (Fay and Herriot, 1979), the method of Elbers et al. (2003), the empirical Best/Bayes (EB) method of Molina and Rao (2010), the hierarchical Bayes (HB) method of Molina et al. (2014) and the M-Quantile approach of Chambers and Tzavidis (2006). A comprehensive review of most of these methods can be found in Guadarrama et al. (2016). In this paper, we discuss advantages and disadvantages of each technique from a practical point of view.

In this article the authors focused on the Fay–Herriot regression model (Fay and Herriot (1979)), whose parameters and area effects are estimated using the feasible generalized least squares (FGLS) method (Greene (2003)) and, apart from proposing special models relevant to the Polish statistical reality, extend the methodological approach suggested by Quintaes et al. (2011) by analysing the gain in precision between the direct and indirect estimator.

In Section 2, we describe our estimation model, which is based on the Fay–Herriot regression, and ways of assessing its precision. Section 3 contains a description of data sources which can be used to obtain relevant variables required for an efficient estimation of poverty. In Section 4 we present properties of the final model, including their quality assessment in terms of spatial variability of residuals and variation of covariates. Finally, Section 6 includes some concluding remarks.

2. The model

The task of estimating poverty was conducted using a model–based approach. Considering the form of available data[7], we chose the Fay–Herriot model on account of its good empirical properties and inherent simplicity. The choice of variables for the model was motivated by qualitative considerations and was based on the relationship between the poverty indicator and selected independent variables, using regression. Whether or not a given variable was to be included in the model depended on model validity, that is on the degree to which the model reflected the relationship. After selecting a variable, a comprehensive analysis was conducted to determine whether the significance level and the sign of the coefficient present next to a given variable matched the reality. In the course of the study references were made to publications concerning the labour market and living conditions. Variables were also considered in terms of their potential to increase the coefficient of determination R^2. In other words, we analysed how much the coefficient of determination increased if a given variable was added to the model. The second – but even more important – criterion of selection was a strong link between the target variables and poverty. Because the assessment of the strength of these relationships was largely subjective, our decision regarding the final form of the model was based on the opinion of experts. More details concerning the justification of our choice are given in Section 4.

As mentioned above, we used the Fay-Herriot model to estimate the at-risk-of-poverty rate in Poland. This model is constructed in two stages, see Pratesi et al. (2016). In the first stage, the so-called sampling model is used to represent the sampling error of the direct estimator. Assuming that μ_d is the variable of interest in the d-th area and $\hat{y}_d$ is a direct estimator of μ_d, the sampling model can be expressed as follows:

$$\hat{y}_d = \mu_d + \varepsilon_d, \quad d = 1, \ldots, D, \tag{1}$$

where D is the number of areas/domains, ε_d are sampling errors, which, given μ_d, are independent and normally distributed with known variances: $\varepsilon_d|\mu_d \overset{iid}{\sim} N(0, \psi_d)$. We assume that ψ_d is known, design-based variance of direct estimator $\hat{y}_d$, $d = 1, \ldots, D$. In the second stage, we assume that the true area characteristics μ_d vary linearly with p area-level auxiliary variables as follows:

$$\mu_d = \mathbf{x}_d^T \beta + u_d, \quad d = 1, \ldots, D, \tag{2}$$

where $\mathbf{x}_d^T$ denotes a vector containing the aggregated (population) values of p auxiliary variables for area d, β is a vector of regression coefficients, u_d are model errors,

[7] Owing to statistical confidentiality, unit-level data could not be used.

which are assumed to be independent and identically distributed, $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ and the vector u_d is independent of the vector ε_d, i.e. $u_d \perp \varepsilon_d$, $d = 1, \ldots, D$. Combining (1) and (2) we obtain a linear mixed model as follows:

$$\hat{y}_d = \mathbf{x}_d^T \beta + u_d + \varepsilon_d, \quad d = 1, \ldots, D. \tag{3}$$

The model is mostly used if only data for a given subpopulation are available (Pfeffermann (2013)).

Since the sample size in subpopulations (subregions) varies, u_d is frequently heteroskedastic. In such cases, the feasible generalized least squares (FGLS), which uses an estimated variance–covariance matrix, is more effective than the classic method of least squares. Under the classic approach to the estimation of regression model parameters, the random error is assumed to be homoskedastic.

After estimating the vector of regression coefficients using FGLS, the estimator based on the Fay–Herriot model, given by (3), is the best linear unbiased predictor (BLUP), which is a weighted mean of the direct and synthetic regression estimator:

$$\hat{\mu}_d = \gamma_d \hat{y}_d + (1 - \gamma_d) \mathbf{x}_d \tilde{\beta}, \tag{4}$$

where

$$\tilde{\beta}(\sigma_u^2) = \left(\sum_d \mathbf{x}_d \mathbf{x}_d^T / (\psi_d + \sigma_u^2) \right)^{-1} \left(\sum_d \mathbf{x}_d \hat{y}_d / (\psi_d + \sigma_u^2) \right), \tag{5}$$

$$\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \psi_d}. \tag{6}$$

After replacing σ_u^2 by its estimate — $\hat{\sigma}_u^2$ in formulas (5) and (6), we obtain an empirical best linear unbiased predictor (EBLUP):

$$\hat{\mu}_d = \hat{\gamma}_d \hat{y}_d + (1 - \hat{\gamma}_d) \mathbf{x}_d \hat{\beta}, \tag{7}$$

where

$$\hat{\beta} = \tilde{\beta}(\hat{\sigma}_u^2) = \left(\sum_d \mathbf{x}_d \mathbf{x}_d^T / (\psi_d + \hat{\sigma}_u^2) \right)^{-1} \left(\sum_d \mathbf{x}_d \hat{y}_d / (\psi_d + \hat{\sigma}_u^2) \right), \tag{8}$$

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_d} \tag{9}$$

and $\hat{\sigma}_u^2$ is the variance of the random error of the model and ψ_d is the direct estimator

variance for a specific area. As mentioned above, ψ_d is assumed to be known, but, in practice, they are estimated from the data. Datta et al. (2005) show that the method of ψ_d estimation can affect the mean square error and its bias. The variance $\hat{\sigma}_u^2$ can be estimated using maximum likelihood (ML) and restricted maximum likelihood (REML). We used the Bayes approach, which is designed to estimate the uncertainty of parameters of the between-area variance by integrating over the posterior density for $\frac{\sigma_u^2}{\psi_d}$ in the case of an area-level model (Rao (2003)) and it is implemented in *hbsae* R package, see Boonstra (2012).

In equation (9), it can be seen that the direct estimator component has a larger weight when ψ_d is small. It means that the EBLUP is (approximately) equal to the direct estimator when it has desirable precision, or is equal to the synthetic component otherwise, see Boonstra and Buelens (2011). It is well known from the literature (Rao (2003)) that this linear combination provides better results than each of its components on its own.

EBLUP can be also expressed as:

$$\hat{\mu}_d = \mathbf{x}_d\hat{\beta} + \hat{u}_d, \tag{10}$$

where: $\hat{u}_d = \hat{\gamma}_d(\hat{y}_d - \mathbf{x}_d\hat{\beta})$. In equation (10) it can be seen that for unrepresented subpopulations, estimates of the target variable are obtained only from the regression model:

$$\hat{\mu}_d = \mathbf{x}_d\hat{\beta}.$$

To calculate the Mean Square Error (MSE) of the EBLUP, we set the following regularity conditions:

(i) ψ_d are uniformly bounded,
(ii) $\sup_{1\leq d\leq D}\mathbf{x}_d^T\left(\sum_{d=1}^D \mathbf{x}_d\mathbf{x}_d^T\right)^{-1}\mathbf{x}_d = O\left(D^{-1}\right)$.

Under normality of the errors u_d and ε_d associated with model (3) and the above regularity conditions, a second order approximation to the MSE is given by:

$$\mathrm{MSE}(\hat{\mu}_d) = g_{1,d}(\sigma_{u^2}) + g_{2,d}(\sigma_{u^2}) + g_{3,d}(\sigma_{u^2}) + O\left(D^{-1}\right), \tag{11}$$

where:

$$g_{1,d}(\sigma_u^2) = \sigma_u^2\psi_d/(\sigma_u^2+\psi_d) = \gamma_d\psi_d \tag{12}$$

is the random error component,

$$g_{2,d}(\sigma_u^2) = (1-\gamma_d)^2\mathbf{x}_d^T\left(\sum_d \mathbf{x}_d\mathbf{x}_d^T/(\sigma_u^2+\psi_d)\right)^{-1}\mathbf{x}_d \tag{13}$$

is the component accounting for the variation of the vector of regression coefficients in the Fay–Herriot model and

$$g_{3,d}(\sigma_u^2) = \psi_d^2(\psi_d + \sigma_u^2)^{-3}\bar{V}(\hat{\sigma}_u^2) \tag{14}$$

is the random-effect component where $\bar{V}(\hat{\sigma}_u^2)$ is the asymptotic variance of the estimator $\hat{\sigma}_u^2$ of σ_u^2.

After replacing σ_u^2 by its estimate – $\hat{\sigma}_u^2$ and γ_d by $\hat{\gamma}_d$ in formulas (12)–(14) the estimator of MSE given by (11) can be calculated using equation (15):

$$\text{mse}(\hat{\mu}_d) = g_{1,d}(\hat{\sigma}_u^2) + g_{2,d}(\hat{\sigma}_u^2) + 2g_{3,d}(\hat{\sigma}_u^2). \tag{15}$$

The standard error of the EBLUP is, of course, represented by the square root of mse, given by (15).

A gain-in-precision index (GPI) was also calculated from equation (16):

$$\text{GPI}_d = \frac{\sqrt{\psi_d}}{\sqrt{mse(\hat{\mu}_d)}}, \tag{16}$$

where: $\sqrt{\psi_d}$ is the direct estimator error, $\sqrt{mse(\hat{\mu}_d)}$ – the error of the estimator based on the Fay–Herriot model. This index shows how much the estimation error could be reduced after applying EBLUP in relation to the direct estimator.

In addition to assessing the precision of Fay-Herriot poverty estimates, we calculated empirical bias using the following bootstrap algorithm:

1. Use model (10) to obtain estimates of $\hat{\sigma}_u^2$ and $\hat{\beta}$.

2. Generate a vector $\omega_1^* \sim N(0,1)$ containing the number of values equal to the number of domains. Calculate $u^* = \hat{\sigma}_u^2\omega_1^*$ and $\theta^* = \mathbf{x}^T\hat{\beta} + u^*$, where $\mathbf{x}^T$ denotes a vector containing the aggregated (population) values of p auxiliary variables.

3. Generate a vector $\omega_2^* \sim N(0,1)$ containing the number of values equal to the number of domains, independently of the ω_1^*. Calculate $e^* = \sqrt{\psi_d}\omega_2^*$.

4. Construct bootstrap data $\hat{\theta}^* = \theta^* + e^* = \mathbf{x}^T\hat{\beta} + u^* + e^*$.

5. Fit model (10) to the new independent variable $\hat{\theta}^*$ and obtain new bootstrap estimates of $\hat{\sigma}_u^{2*}$ and $\hat{\beta}^*$.

6. Calculate EBLUP as $\hat{\theta}^{E*} = \mathbf{x}^T\hat{\beta}^* + \frac{\hat{\sigma}_u^{2*}}{\hat{\sigma}_u^{2*}+\psi_d}(\hat{\theta}^* - \mathbf{x}^T\hat{\beta})$.

7. Repeat steps 2–6 B times. Let $\hat{\theta}^{E*(b)}$ be the bootstrap EBLUP and $\theta^{*(b)}$ be the bootstrap true value obtained in b-th bootstrap replication.

8. Estimated bootstrap bias is given by:

$$\text{BIAS} = B^{-1} \sum_{b=1}^{B} (\hat{\theta}^{E*(b)} - \theta^{*(b)}). \tag{17}$$

Values obtained from equation (17) will be compared with the empirical bias of the direct estimator.

It is worth noting that in the literature many different extensions of the model (3) have been proposed. These include a multivariate generalization studied by González-Manteiga et al. (2008) and models where time and spatial correlations play a crucial role. The problem of borrowing strength over time was considered by Choudry and Rao (1989), who extended the basic Fay-Herriot model by taking into account the impact of time and considering an autocorrelated structure for sampling errors, which for each domain are assumed to follow an autoregressive process AR(1). More precisely they considered the following model:

$$\hat{y}_{dt} = \mathbf{x}_{dt}^T \beta + u_d + \varepsilon_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, T \tag{18}$$

where

$$\varepsilon_{dt} = \rho \varepsilon_{d,t-1} + \varepsilon_{dt}, \quad |\rho| < 1, \quad \varepsilon_{dt} \overset{iid}{\sim} N(0, \psi_d). \tag{19}$$

Esteban et al. (2011) considered a very similar model as in (18):

$$\hat{y}_{dt} = \mathbf{x}_{dt}^T \beta + u_{dt} + \varepsilon_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, T \tag{20}$$

but the authors assumed that random effects u_{dt} follow an AR(1) stochastic process.

The problem of spatial correlation in the data was considered by Singh et al. (2005), Petrucci and Salvati (2006) and Pratesi and Salvati (2008), who extended the basic Fay-Herriot model by assuming that area effects u_d follow a spatial autoregressive process of order 1 or SAR(1). In general, the authors demonstrated that if there is unexplained spatial correlation in the data, then it is possible to improve model efficiency by taking into account the fact that data from neighbouring areas are correlated.

The problem of borrowing strength simultaneously across areas and over time was considered by Rao and Yu (1994). They proposed the following model:

$$\hat{y}_{dt} = \mathbf{x}_{dt}^T \beta + u_{1d} + u_{2dt} + \varepsilon_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, T, \tag{21}$$

where area effects u_{1d} are constant over time and follow the usual assumptions adopted in the basic Fay-Herriot model and u_{2dt} are time-varying effects that follow an AR(1) process and are independent across areas. This model was extended, for instance, by Marhuenda et al. (2013) by considering spatial correlation in domain random effects u_{1d}, which follow a SAR(1) process. A very comprehensive review of different extensions of the basic Fay-Herriot model is also provided in Pratesi et al. (2016), where a new modification of the model (21) with moving average MA(1) of correlated random effects is also proposed.

Extensions of the Fay-Herriot model which allow for spatial correlation assume spatial stationarity, i.e. parameters of the associated regression model for the small area characteristic of interest do not vary spatially. Chandra et al. (2015) proposed an extension of the Fay-Herriot model, which accounts for the presence of spatial nonstationarity, i.e., where parameters of this regression model vary spatially.

It is worth noting that in the basic Fay-Herriot model, it is assumed that direct survey estimators are a linear function of covariates, an assumption which, in practice, may not hold. As a consequence, this may lead to biased estimators of the small area parameters. A remedy for this inconvenience may be a semiparametric specification of the Fay-Herriot model proposed by Giusti et al. (2012), which allows nonlinearity in the relationship between the response variable and auxiliary variables by using penalized splines.

The basic Fay-Herrtiot model discussed in this article has both good and bad properties. According to Guadarrama et al. (2016), the advantages (a–d) and disadvantages (e–i) of using the Fay-Herriot model, also in the context of poverty, include:

a) the Fay-Herriot estimator automatically borrows strength for areas where it is necessary,

b) if parameter $\gamma_d > 0$, then it makes use of the sampling weights through the direct estimator $\hat{y}_d$, thus it is design-consistent (as $n_d \to \infty$),

c) because it relies on aggregated data, it is not very much affected by isolated unit-level outliers,

d) it only requires area-level auxiliary information and therefore avoids confidentiality issues associated with micro-data,

e) the sampling variances ψ_d are assumed to be known, but in practice they have to be estimated,

f) the number of observations used to fit the Fay-Herriot model is equal to the number of areas, which, in most cases, is relatively small; as a result, model

parameters are estimated with less efficiency compared to unit-level models, where the number of observations is much greater than the number of areas,

g) it requires normality of μ_d and ε_d for MSE estimation; it is very difficult to fulfil this assumption for complex poverty indicators,

h) in order to estimate several indicators that depend on a common continuous variable, it is necessary to fit a different model and search for good covariates for each indicator,

i) after fitting the model at the area level, small area estimates $\hat{\mu}_d$ cannot be further disaggregated for subareas/subdomains within the areas unless a new good model is found at that subarea level.

Finally, although there are many extensions of the classical Fay-Herriot area-level model in the literature, we decided to apply the basic one. For one thing, we could only use area-level data. Another important factor was the relative simplicity of this model, which is especially important in the context of official statistics, where the use of complex models is still limited (Brakel and Bethlehem, 2008). This is also true of official statistics in Poland, which generally relies on more traditional design-based approaches. Finally, the basic assumptions of the Fay-Herriot model were fulfilled, which prompted us to apply it to the estimation of poverty rate in Poland across subregions.

3. Basic sources of data required for estimation

The model constructed in the study was based on data from a few statistical sources. The only variable (response variable) taken from the EU–SILC survey was the poverty indicator, since the use of other variables as independent variables would only have contributed to a higher random error and generated biased estimates of β parameters in the model. For this reason, explanatory variables came from the 2011 National Census of Population and Housing and the Local Data Bank data from 2005–2011.

The amount of random error depends on the sample size, the amount of variability associated with a given variable and the sampling scheme used. In the case of a full census, there is no random error. However, as a mixed-mode census, the 2011 National Census of Population and Housing included a 20% sample of Poland's population, i.e. about 8 million people were surveyed. In contrast, the sample size in the 2011 EU-SILC survey was only 28,305 respondents, corresponding to 0.075% of the total. The level of random error in both cases is incomparable, and with respect to the survey part of the 2011 National Census of Population and Housing, for

general cross classifications with large sample sizes, can be regarded as negligible. The sample size in subregions in the 2011 National Census of Population and Housing ranged from 41,014 respondents (the city of Szczecin) to 216,923 (the region of Ostrołęka–Siedlce). For example, for the variable *the percentage of single people aged over 25 by subregion*, the coefficient of variation ranged from 0.66% to 1.48%, which is a very low value. However, in the estimation theory a variety of models that account for estimation errors in auxiliary variables have been discussed. For instance, Buonaccorsi (1995) considered a modification of the estimation models and discussed the question when to correct the model by the measurement error and what method of estimating the standard deviation of the prediction to use in this situation. He justified the necessity of such corrections especially when the MSE of estimates of covariates varies. Moreover, Ybarra and Lohr (2005) proposed the best measurement error estimator and discussed some of its asymptotic properties. They concluded that if MSEs of auxiliary variables are larger, the target indicator is underestimated. The application of the measurement error estimation can improve the quality of final estimates expressed in terms of their MSE, but – on the other hand – the estimator of the variance of the model error is often greater. Their results imply that if estimators for auxiliary variables are unbiased, have the least possible variance and are based on relatively large samples, then their errors have no significant impact on the final quality. This fact and properties of our data justify the assumption of negligibility of such errors for covariates.

In order to build the final model at the level of subregions, we considered the following variables:

- demographic information, including population structure in terms of age, sex, education level and marital status;
- division into urban and rural areas;
- economic activity status: the number of economically active, employed, unemployed in a given population;
- housing infrastructure: dwelling size per person, access to electricity, sewerage system, central heating, gas, shower, bathtub;
- household indicators: number of employed persons, unemployed, economically active (aged 15–64), number of people in a household, number of rooms per person, the level of education of household members;
- budgets of territorial units;
- road infrastructure;

- environmental conservation: gas and particle pollution;
- health care and social welfare, including pre–school care;
- migration balance for specific years;
- territorial division: peripheral subregions; metropolitan cities, former provincial capitals, area of subregions, cities with populations exceeding 100,000.

A total of over 200 independent variables were considered and analysed. The variables were then used to construct the model, with special emphasis on determinants presented in the publication by CSO (2013). Apart from data collected during the 2011 National Population and Housing Census in Poland, we investigated covariates with similar properties from many other sources, e.g. Local Data Bank of the Central Statistical Office(`https://bdl.stat.gov.pl`), containing data at various local levels compiled from such primary sources as the Head Office of Land Surveying and Cartography, the Polish Population Register (PESEL), Polish Tax Register (POLTAX), reports provided by register offices and provincial courts, etc. During the selection of variables various models of verification of regressors were applied: a multiple linear model with the control of the coefficient of determination and Student's t–test as well as stepwise regression based on the forward and backward approach, where choice/elimination of variables is made in subsequent steps according to optimization of the F test. To achieve this objective, we relied on the approach developed by the World Bank in cooperation with national statistical institutes of numerous European countries to estimate poverty at the NUTS 3 level or lower (Bedi et al. (2007)).

4. The model and its properties

If the dependent variable is subjected to arcsin square-root transformation, estimates will be included in $[0;1]$ interval and variance will be stabilised (Burgard et al. (2015)). After applying this approach to direct estimates of the poverty rate at subregions level in Poland, the distribution of target variable was less skewed (from 0.83 to 0.49). However, the obtained model was only slightly better than the model based on raw data — a very small increase in R^2. It is worth noting that the variance of the direct estimator at subregion level does not vary considerably. Moreover, since this model was intended for official statistics, the use of simple techniques was preferred. The use of raw data has yet another advantage — β coefficients are easy to interpret.

The problem of the optimal selection of correlates of poverty is widely discussed in the literature. For example, Bedi et al. (2007) discuss the per capita consumption

in households. However, they use this variable only at the national level. In contrast, we deal with much lower territorial units, for which such information is unavailable[8]. This problem was in some sense confirmed by Chandra et al. (2016), who investigated possibilities of estimating household consumption in Italian subregions – the final CVs of some estimates exceeded 50% and in extreme cases were even several times greater. A more interesting collection of correlates (for Italian NUTS 3 provinces) was proposed by Quintano et al. (2007) – their model uses several demographic variables, Gross Domestic Product (GDP), Growth Enterprises Rate and four binary variables determining the macro-region which these provinces are located in. It is worth noting that their approach is mainly based on GDP and leaves out some important aspects of living conditions, which cannot be fully reflected by GDP. A much poorer set of covariates was considered by Morales et al. (2015) for estimating poverty in Spanish provinces – it consists of only three variables: age group 50–65, secondary education completed and unemployed persons, but uses an interesting method of estimation based on some natural partitioning of spatial units. Salvati et al. (2014) used some key household characteristics as poverty covariates: mean income, percentage of divorced households, ownership of the dwelling where the household lives and – depending on the region – ratio of widowed to married households and the percentage of households with an employed person. These variables were used only for large regions.

The final model included 6 explanatory variables, which are listed below, together with their sources given in round brackets:

- the percentage of single people aged over 25 (2011 National Population and Housing Census);
- the number of rooms per one household member (2011 National Population and Housing Census);
- the percentage of households with a bathroom or shower (2011 National Population and Housing Census);
- the percentage of households with two persons aged over 25 with no more than vocational education (2011 National Population and Housing Census);
- population density (it is a ratio of the population to the area of a given spatial unit: population data were derived from the 2011 National Population and

[8]These data are collected only during the Household Budget Survey, where the sample size is too small to ensure the sufficient quality of estimates at this level.

Housing Census, and data about the area from the Local Data Bank – the Head Office of Land Surveying and Cartography (as of 31 December 2011))[9];

- the ratio of people deregistered to the number of people registered for permanent residence in the subregion (Local Data Bank: based on the National Census and the PESEL register, reports provided by register offices and provincial courts (as of 31 December 2011)).

It was observed that as the percentage of single people aged 25+ in a subregion increases, the poverty indicator increases as well. On the other hand, with the increasing number of rooms per one household member and the rising percentage of households with a bathroom, the poverty indicator gradually decreases. It was also observed that the poverty indicator was positively correlated with the percentage of households with two persons having no more than vocational education. However, the direction of this correlation is ambiguous, i.e. without a detailed analysis it is hard to determine whether poverty is caused by the low level of education or vice versa (Haughton and Khandker (2009)). Subregions with lower population density and a higher ratio of people deregistered to the number of people registered for permanent residence exhibited a higher poverty rate. More precisely, low population density is a key feature of rural or deindustrialized areas, where poverty is naturally higher. On the other hand, larger than average population density could also contribute to an increase in poverty, but only in overpopulated cities, where sustainable development of population cannot be achieved (it is the classical definition of overpopulated areas – see, e.g. Kamaraj et al. (2014)). However, this problem currently does not exist in Poland. Another phenomenon positively correlated with poverty is intensive emigration from a given area – if there is no other, e.g. political, reason – owing to a high risk of poverty (caused by, e.g. insufficient number of workplaces). Inversely, high immigration indicates that the recipient region is attractive for incoming people who want to improve their standard of living.

The model based on these variables explained 60% of the variation in the poverty indicator. Table 1 shows a summary of estimation results obtained by applying the model based on the regression dependence of the poverty indicator on the explanatory variable and the assessment of its quality. We present the adjusted coefficient of determination R^2, Fisher's test and estimates of particular regression coefficients with relevant standard errors, t–statistics and its ex post significance level (p–value).

[9] More precisely, a binary variable was created, taking the value of 1 if the subregion's population density was below the 33rd percentile of the population density distribution for all subregions, or 0 otherwise. The variable was used to identify subregions with low population density. If a given subregion is in the group of subregions with population density below 33rd percentile of population density distribution in subregions, then it is reasonable to suppose that it will negatively affect the at–risk–of–poverty indicator.

The overall quality of this model seems to be high. The degree of determination is quite satisfactory, as indicated by the high value of the F–test statistic showing that the vector of β coefficients is significant.

Table 1. The final model – diagnostics

Model	σ_u^2 0.0017	F–statistic 16.96		
	Adj.R^2 59.56	DF 59		
	Coefficient	Standard error	t–statistic	p–value	
Intercept	0.7437	0.2239	3.32	0.0015	**
The share of households with bath or shower	-0.7854	0.1606	-4.89	0.0000	***
The percentage of single people (aged 25+)	1.3958	0.5209	2.68	0.0095	**
The number of rooms per one person	-0.1464	0.0768	-1.91	0.0614	.
The share of households with two persons having no more than vocational education	0.3031	0.1903	1.59	0.1166	
The ratio of people deregistered to the number of people registered for permanent residence	0.0199	0.0327	0.61	0.5458	
Population density (lower than 33rd percentile)	0.0187	0.0153	1.22	0.2285	
Significance codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1					

Three covariates – the share of households with two persons with not more than vocational education, the ratio of people deregistered to the number of people registered for permanent residence and population density (lower than 33rd percentile) – are statistically less significant than the others, but their information value from the point of view of the target estimation is high. That is, a low level of education is usually one of the main factors contributing to increasing the risk of poverty. The relevance of the level of migration and population density was presented earlier. Elimination of such covariates would significantly deteriorate the quality of the model. Hence – according to our experience and following the advice of specialists from the World Bank – we retained them in the model. The share of households with a bath or shower and the number of rooms per person are negatively correlated with the poverty indicator (the higher they are, the lower the risk of poverty).

4.1. Model checking

The requirement for applying the Fay-Herriot model is that a number of assumptions, mainly concerning normality, should be satisfied. This part of the paper is dedicated to model checking. Firstly, the results were analysed to check for non-normality of residuals and outliers — Figure 1.

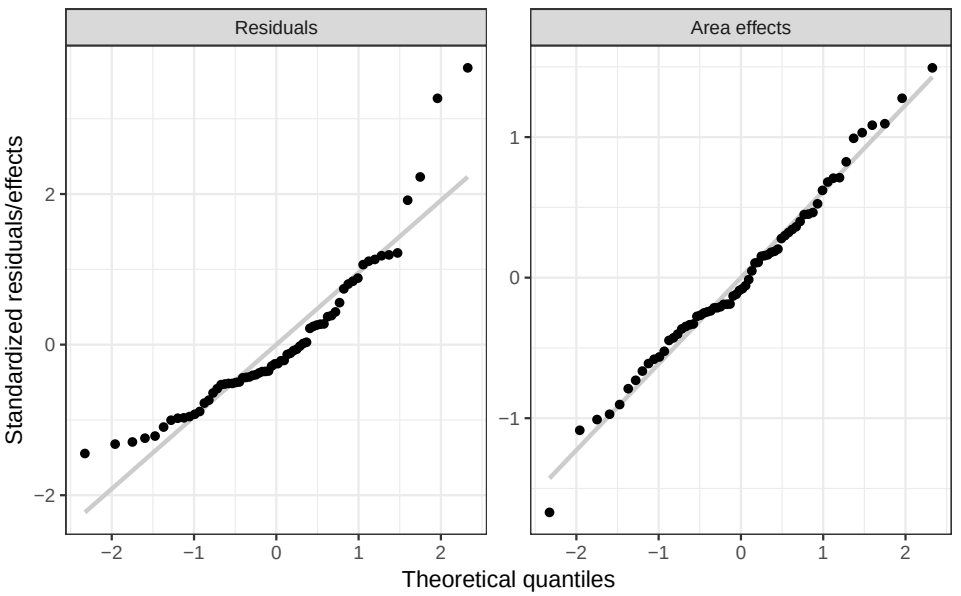


Figure 1. Q-Q plot of Fay-Herriot model residuals and area effects

If the residuals are normally distributed, the dots will be plotted along the line. As can be seen in Figure 1, the dots are very close to the line but there are two evident outliers (standardized residual more than 3). These values are connected with Tarnowski and Gorzowski subregion. Nonetheless, the Pearson correlation coefficient is quite high and equals $\rho = 0.94$. It must be emphasized that each residual represented by one point on the plot has a different variance (ψ_d). The distribution of residuals and area random effects in the Fay-Herriot model seems to be normal. This is confirmed by the Kolmogorow-Smirnov normality test with p-value equal to 0.575 for residuals and 0.438 for area effects, which means that there is no evidence against the null hypothesis.

We also tested multicollinearity and homogeneity of variance. Variance Inflation Factors are less than 2 for all independent variables, so there is no multicollinearity.

5. Results of estimating the at–risk–of–poverty rate

The final estimates of the at–risk–of–poverty rate were calculated using the empirical best linear unbiased predictor EBLUP expressed by the equation (7). Estimates of the at–risk–of–poverty rate based on the Fay-Herriot model and covariates from Table 1 at NUTS 3 level are presented in Figure 2.

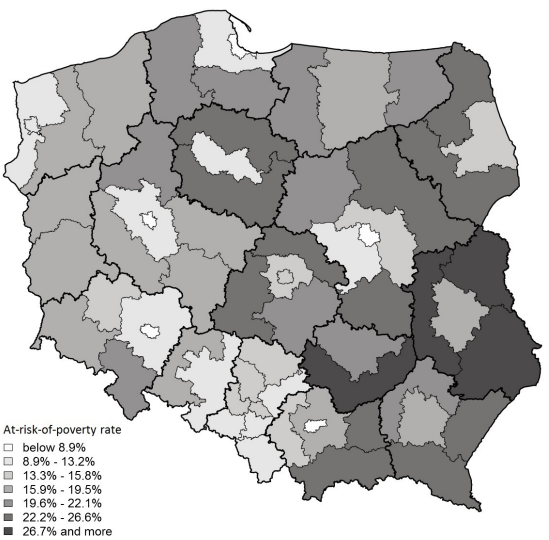


Figure 2. The poverty indicator at the level of subregions based on the final model shown on a 7–class color scale

The results reveal a strong territorial variation in the poverty indicator. The cartogram shows that Poland can be divided into two parts: Central and Eastern Poland on the one hand, and Western Poland, on the other. Western Poland is characterized by a much lower percentage of poor people than Central and Eastern Poland.

According to CSO data, the poverty indicator for Poland based on the EU-SILC survey amounts to 17.7% (CSO (2012)). Estimates presented in this paper provide information about the scope of poverty in Poland at the level of subregions (NUTS 3 - 66 units). So far, poverty statistics have not been published at this level of aggregation. A preliminary analysis of the poverty map reveals a difference between Central and Eastern Poland (with a higher poverty rate) and Western Poland, characterised by a lower scope of poverty. The highest percentage of poor persons in the population (at least 29%) was observed in 4 subregions located in Lubelskie province (3 subregions), and Świętokrzyskie province (1 subregion). On the other

hand, the lowest level of poverty (below 9%) can be observed in 5 major cities (with the exception of Łódź), which constitute separate subregions, i.e. in Warszawa, Kraków, Tri–City (Gdańsk, Gdynia and Sopot), Wrocław and Poznań.

The highest at–risk–of–poverty rate (the poverty indicator over 29%) is estimated for people living in households located in 4 subregions in the province of Lublin (the subregions of Biała Podlaska, Puławy, and Chełm–Zamość) and Świętokrzyskie (the subregion of Sandomierz–Jędrzejów). The lowest values of the poverty indicator can be observed in big cities (with the exception of Łódź at 14.2%). The poverty rate in Warszawa was estimated at the level of 6.3%, followed by the Tri–City (Gdańsk, Gdynia and Sopot) subregion (7.4%) and the subregion of Wrocław (7.5%), Poznań (8.5%) and Kraków (8.7%). It should also be noted that most subregions surrounding big cities exhibit significantly lower levels of poverty (below 13%) than other subregions in the same province.

Detailed information about direct estimates, their standard errors, EBLUP estimates and their standard errors and GPI can be found in Table 2 in the appendix. In the table estimates of the poverty indicator obtained by means of the direct estimator are compared with those generated by the model using equation (7) for particular subregions ordered according to code values used in the territorial units register. In addition, it presents the gain-in-precision index expressed as a ratio of the standard error of the direct estimator to the standard error of the EBLUP estimator given by equation (16), showing the number of times the standard error was reduced by the model–based estimator in comparison with the direct estimator.

It is interesting that the gain-in-precision index is usually smaller for highly–urbanized areas (Warszawa, Kraków, Łódź, etc.) or areas located within functional zones of large cities (e.g. Warszawski wschodni and Warszawski zachodni subregions). Conversely, the largest values of the GPI are achieved for regions with the prevalence of agriculture and a low level of industrialization (Przemyski, Szczeciński, ełcki, etc.), where the poverty indicator is relatively high. This can be associated, firstly, with the obviously greater representation of large cities and their surroundings in the sample and, secondly, with the efficient choice of covariates in the final model.

Figure 3 is a cartogram showing differences between estimates of the at–risk–of–poverty rate obtained by the direct and EBLUP estimator. It was used to analyze the estimation results obtained using different estimators and find possible systematic patterns. As can be seen, the distribution of residuals does not reveal systematic spatial patterns.

Another aspect worth analysing is whether the differences between direct and EBLUB estimates are significant. The result of the classical t–test is 1.63 (p–

value=0.1082). In other words, the hypothesis that the expected value of the difference is zero cannot be rejected. However, the value of the pooled F–test is 1.63 (p–value=0.0509). In other words, for the *ex ante* significance level greater than 0.051, the variances can be regarded as different. Alternatively, we can perform the Satterthwaite's or Cochran's test, but they give similar – and even stronger – results in the classical case (their p–values are slightly greater than 0.48). These tests indicate that comparisons based only on point estimates, like those performed in this case, could not express all important differences, but analysis of variability can exhibit them more clearly.

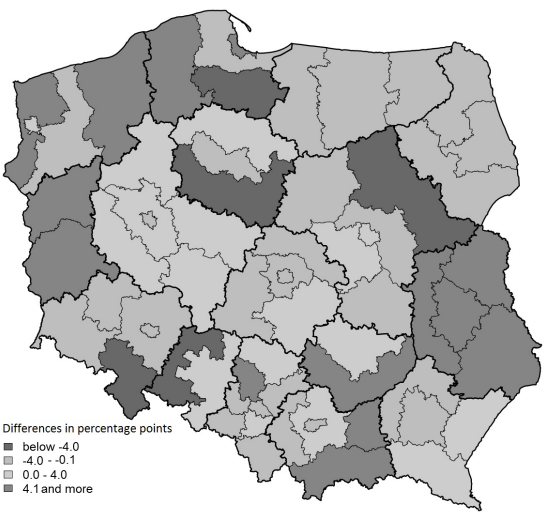


Figure 3. Differences between estimates of the poverty indicator obtained by means of the direct and EBLUP estimators

Further analysis focuses on bias and standard errors of the estimates. Its results are presented in Figure 4. Bias estimation was based on the bootstrap procedure described in Section 2 with $B = 500$ replicates. The left panel of Figure 4 shows the spread of the EBLUP estimator for unplanned domains is larger than that of the direct estimator. Nevertheless, mean bias of direct estimates is equal to 0.01, compared to -0.03 for the Fay-Herriot model. It is clear from the right panel of Figure 4 that the EBLUP estimator is significantly more efficient than the direct one.

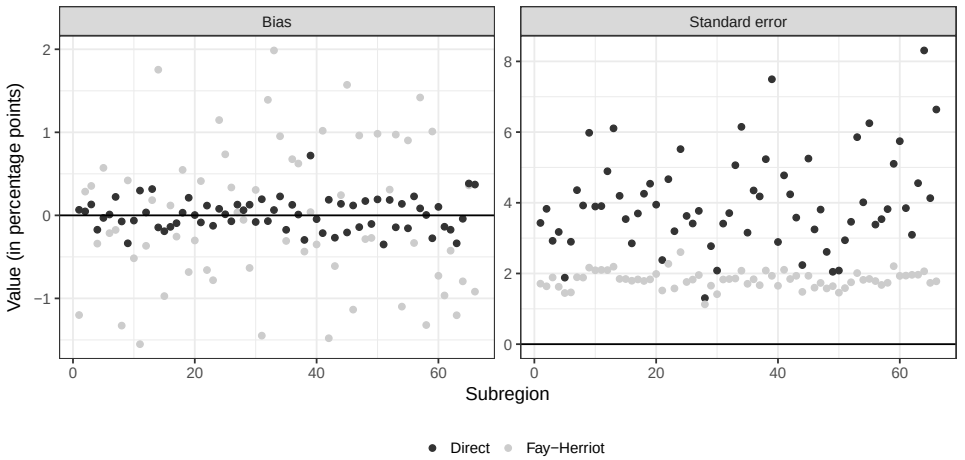


Figure 4. Bias and standard errors of the estimates of the poverty rate.

6. Conclusion

The study described in the article shows that given a carefully selected set of covariates that come from sources which are either not burdened with random error (e.g. administrative registers) or where this kind of error is very low (for instance censuses), it is possible to construct efficient models of the composite EBLUP estimator, which provide reliable estimates at lower levels of aggregation than those currently available. Although there are many possible ways of building such models, the selection of the final model can be optimized on the basis of various important criteria determined by the objectives of the study and properties of the dependent variable and explanatory variables as well as correlations between them. In our case, we considered cause–effect relations, the degree and precision of determination of dependence of the poverty indicator on explanatory variables, their information value and the quality of estimates obtained using EBLUP with the Fay–Herriot model based on these covariates. It is worth noting that in the case of EBLUP, weights associated with the direct estimator were relatively high and the variation in EBLUP estimates was significantly lower than in the case of the direct estimator.

Therefore, our model can be efficiently applied in statistical practice. It gives much more precise estimates of poverty based on covariates, which can be treated as indicators; as a result, not only can they be regarded as high–quality statistical outputs but can also be used as a reliable criterion of assessing comparability of the poverty indicator over time and across areas (Młodak (2013)). However, one should remember that our model – like any econometric model – is a simplification

of reality and hence there is a risk that under some circumstances estimates of the at–risk–of–poverty rate may not reflect the actual status in this respect. Therefore, its efficiency should be verified taking into account specific characteristics of the social and economic situation in areas of interest. Nevertheless, in general, this approach seems to be better than other similar attempts.

Acknowledgements

The authors wish to thank two anonymous reviewers for detailed and helpful comments about the manuscript. Special thanks to Alexandru Cojocaru, Céline Ferré, Ken Simler and Roy van der Weide from the World Bank and collaborators from the Statistical Office in Poznan and the Department of Social Surveys and Living Conditions at the Central Statistical Office. The work of Szymkowiak has been developed under the support of the project entitled 'Indirect estimation of disability in the 2011 census', which was financed by the National Science Centre in Poland in the framework of a grant awarded by virtue of decision no. DEC-2013/11/B/HS4/01472.

REFERENCES

BEDI, T., COUDOUEL, A., SIMLER, K., (2007). More Than a Pretty Picture. Using Poverty Maps to Design Better Policies and Interventions. The World Bank, Washington D.C., U.S.A.

BOONSTRA, H. J., (2012). hbsae: Hierarchical Bayesian Small Area Estimation. R package version 1.0.

BOONSTRA, H. J., BUELENS, B., (2011). Model-Based Estimation. Statistische Methoden (201106), Statistics Netherlands, The Hague/Heerlen, The Netherlands.

BRAKEL, J., BETHLEHEM, J., (2008). Model-Based Estimation for Official Statistics. Discussion paper, Statistics Netherlands, pp. 1–16.

BUONACCORSI, J. P., (1995). Prediction in the presence of measurement error: General discussion and an example predicting defoliation. Biometrics 51, pp. 1562–1569.

BURGARD, J. P., MÜNNICH, R., ZIMMERMANN, T., (2015). Impact of Sampling Designs in Small Area Estimation with Applications to Poverty Measurement. Analysis of Poverty Data by Small Area Estimation, pp. 83–108.

CHAMBERS, R., TZAVIDIS, N., (2006). M-quantile models for small area estimation, Vol. 93 (2), Biometrika, pp. 255–268.

CHANDRA, H., SALVATI, N., CHAMBERS, R., (2015). A spatially nonstationary Fay-Herriot model for small area estimation. Journal of Survey Statistics and Methodology, 3 (2), pp. 109–135.

CHANDRA, H., SALVATI, N., CHAMBERS, R., (2016). Model-based Direct Estimation of a Small Area Distribution Function. [in:] Pratesi, M. (ed.) Analysis of Poverty Data by Small Area Estimation, Series: Wiley Series in Survey Methodology, John Wiley & Sons Ltd., The Atrium, Southern Gate, Chichester, West Sussex, United Kingdom.

CHOUDRY, G.H., RAO, J.N.K., (1989). Small area estimation using models that combine time series and cross sectional data. In: Singh, A.C., Whitridge, P. (Eds), Proceedings of Statistics Canada Symposium on Analysis of Data in Time, Ottawa: Statistics Canada, pp. 67–74.

CSO, (2012). Incomes and living conditions of the population in Poland (report from the EU-SILC survey in 2011). Central Statistical Office of Poland, Social Surveys and Living Conditions Department, Statistical Publishing Establishment, Warszawa, Poland. in Polish.

CSO, (2013). Life quality. Social capital, poverty and social exclusion in Poland. Central Statistical Office of Poland, Social Surveys and Living Conditions Department, Statistical Publishing Establishment, Warszawa, Poland. in Polish.

DATTA, G. S., RAO, J. N. K., SMITH, D. D., (2005). On measuring the variability of small area estimators under a basic area level model. Biometrika, 92 (1), pp. 183–196.

ELBERS, C., LANJOUW, J. O., LANJOUW, P., (2003). Micro-level Estimation of Poverty and Inequality. Econometrica, 71 (1), pp. 355–364.

ESTEBAN, M.D., MORALES, D., PÉREZ, A., SANTAMARÍA, L., (2011). Two area-level time models for estimating small area poverty indicators. Journal of the Indian Society of Agricultural Statistics, 66, pp. 75–89.

FAY, R. E. HERRIOT, R. A., (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, pp. 269–277.

GIUSTI, C., MARCHETTI, S., PRATESI, M., SALVATI, N., (2012). Semiparametric Fay-Herriot model using penalized splines. Journal of the Indian Society of Statistics, 66 (1), pp. 1–14.

GREENE, W. H., (2003). Econometric Analysis. Pearson Education (Singapore) Pte. Ltd., Indian Branch., Delhi, India, 5th, edition.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M.J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model, Computational Statistics and Data Analysis, vol. 52, issue 12, pp. 5242–5252.

GUADARRAMA M., MOLINA I., RAO, J.N.K., (2016). A Comparison of Small Area Estimation Methods for Poverty Mapping, Statistics in Transition new series, Vol. 17, pp. 41–66.

HAUGHTON, J. KHANDKER, S. R., (2009). Handbook on Poverty and Inequality. The World Bank, Washington D. C., U.S.A.

KAMARAJ, K., KATHIRAVAN, C., JAYAKUMAR, A., (2014). Entrepreneurship – A Possible Solution To The Surplus Population, Journal of Business Management & Social Sciences Research (JBM&SSR), 3 (1), pp. 27–32.

MARHUENDA, Y., MOLINA, I., MORALES, D., (2013). Small area estimation with spatio-temporal Fay-Herriot models. Computational Statistics & Data Analysis, 58, pp. 308–325.

MŁODAK, A., (2013). Coherence and comparability as criteria of quality assessment in business statistics. Statistics in Transition-new series, 14, pp. 287–318.

MOLINA, I., RAO, J.N.K., (2010). Small area estimation of poverty indicators. The Canadian Journal of Statistics, 38, pp. 369–385.

MOLINA, I. NANDRAM, B., RAO, J.N.K., (2014). Small area estimation of general parameters with application to poverty indicators: A Hierarchical Bayes approach. The Annals of Applied Statistics, 8(2), pp. 852–885.

MORALES, D., PAGLIARELLA, M. C., SALVATORE, R., (2015). Small area estimation of poverty indicators under partitioned area-level time models. SORT, 39 (1), pp. 19–34.

PETRUCCI, A., SALVATI, N., (2006). Small area estimation for spatial correlation in watershed erosion assessment. Journal of agricultural, biological, and environmental statistics, 11, pp. 169–182.

PFEFFERMANN, D., (2013). New Important Developments in Small Area Estimation. Statistical Science, 28, pp. 40–68.

PRATESI M. (editor), (2016). Analysis of Poverty Data by Small Area Estimation, Wiley Series in Survey Methodology, Wiley.

PRATESI M., SALVATI, N., (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. Statistical Methods & Applications, 17, pp. 113–141.

QUINTAES, V., HANSEN, N., SILVA, D., PESSOA, D., SILVA, P., (2011). A Fay-Herriot Model for Estimating the Proportion of Households in Poverty in Brazilian Municipalities. Proceedings from the 58th World Statistical Congress, Dublin (Session CPS016), International Statistical Institute, pp. 4218–4223.

QUINTANO, C., CASTELLANO R., PUNZO, G., (2007). Estimating Poverty in the Italian Provinces using Small Area Estimation Models. Metodološki zvezki, 1 (4), pp. 37–70.

RAO, J.N.K., (2003). Small Area Estimation. John Wiley & Sons, Inc., Hoboken, New Jersey, U.S.A.

RAO, J.N.K., YU, M., (1994). Small-Area Estimation by Combining Time-Series and Cross-Sectional Data. The Canadian Journal of Statistics, 22, pp. 511–528.

SALVATI, N., GIUSTI, C., PRATESI, M., (2014). The Use of Spatial Information for the Estimation of Poverty Indicators at the Small Area Level [in:] Betti, G., Lemmi, H, (eds.) Poverty and Social Exclusion, New Methods of Analysis, Routledge, London, United Kingdom.

SINGH, B.B., SHUKLA G.K., KUNDU, D., (2005). Spatio-Temporal Models in Small Area Estimation. Survey Methodology, 26, pp. 173–181.

WAWROWSKI, Ł., (2014). Wykorzystanie metod statystyki małych obszarów do tworzenia map ubóstwa w Polsce, Wiadomości Statystyczne, 9, pp. 46–56.

YBARRA, L. M. R., LOHR, S. L., (2008). Small area estimation when auxiliary information is measured with error. Biometrika 95: pp. 919–931.

Appendix – results

Table 2. Estimated scope of poverty (in %) together with values of the standard error (in percentage points).

Subregion (NUTS 3)	Direct estimate	Standard error	EBLUP estimate	Standard error	Precision gain
Jeleniogórski	15.7	3.4	17.1	1.7	2.00
Legnicko–Głogowski	14.4	3.8	14.5	1.6	2.34
Wałbrzyski	15.3	2.9	20.5	1.9	1.55
Wrocławski	11.3	3.2	12.6	1.6	1.96
City of Wrocław	6.2	1.9	7.5	1.4	1.30
Bydgosko–Toruński	11.5	2.9	12.1	1.5	1.98
Grudziądzki	26.1	4.4	22.9	1.9	2.30
Włocławski	18.3	3.9	22.6	1.9	2.08
Bialski	35.2	6.0	29.4	2.2	2.76
Chełmsko–Zamojski	34.7	3.9	30.2	2.1	1.86
Lubelski	24.0	3.9	18.5	2.1	1.86
Puławski	35.4	4.9	29.5	2.1	2.33
Gorzowski	31.0	6.1	16.4	2.2	2.79
Zielonogórski	21.7	4.2	17.7	1.8	2.27
Łódzki	14.1	3.5	15.1	1.8	1.92
City of Łódź	13.9	2.9	14.2	1.8	1.59
Piotrkowski	23.6	3.7	21.6	1.8	2.02
Sieradzki	21.5	4.3	24.4	1.8	2.38
Skierniewicki	21.5	4.5	23.4	1.8	2.48
Krakowski	17.7	3.9	17.4	2.0	1.99
City of Kraków	8.4	2.4	8.7	1.5	1.57
Nowosądecki	28.8	4.7	23.2	2.3	2.05
Oświęcimski	12.0	3.2	14.3	1.6	2.03
Tarnowski	40.9	5.5	24.6	2.6	2.12
Ciechanowsko–Płocki	18.2	3.6	21.3	1.8	2.06
Ostrołęcko–Siedlecki	21.1	3.4	25.7	1.8	1.87
Radomski	23.5	3.8	24.5	2.0	1.93
City of Warszawa	6.2	1,3	6.3	1.1	1.16
Warszawski wschodni	12.8	2.8	14.4	1.7	1.68
Warszawski zachodni	10.8	2.1	10.3	1.4	1.47
Nyski	12.2	3.4	16.5	1.8	1.86
Opolski	14.2	3.7	11.5	1.8	2.01
Krośnieński	25.9	5.1	24.1	1.9	2.73
Przemyski	28.6	6.1	26.1	2.1	2.96

continued on the next page

Subregion (NUTS 3)	Direct estimate	Standard error	EBLUP estimate	Standard error	Precision gain
Rzeszowski	14.7	3.2	18.0	1.7	1.85
Tarnobrzeski	19.7	4.3	20.9	1.8	2.37
Białostocki	12.0	4.2	13.4	1.7	2.50
Łomżyński	21.4	5.2	24.6	2.1	2.51
Suwalski	18.5	7.5	22.2	1.9	3.87
Gdański	11.0	2.9	11.9	1.7	1.75
Słupski	29.7	4.8	20.8	2.1	2.27
Starogardzki	17.3	4.2	22.0	1.8	2.30
Tri–City*	13.3	3.6	7.4	1.9	1.85
Bielski	10.5	2.2	11.1	1.5	1.51
Bytomski	24.1	5.3	13.9	1.9	2.71
Częstochowski	15.2	3,2	14.6	1.6	2.03
Gliwicki	13.4	3.8	14.1	1.7	2.20
Katowicki	13.6	2.6	14.6	1.6	1.66
Rybnicki	10.1	2.0	10.4	1.6	1.25
Sosnowiecki	9.5	2.1	10.2	1.5	1.43
Tyski	10.3	2.9	9.9	1.6	1.85
Kielecki	22.2	3.5	21.3	1.8	1.98
Sandomiersko–Jędrzejowski	34.0	5.9	29.8	2.0	2.91
Elbląski	17.6	4.0	20.7	1.8	2.21
Ełcki	17.5	6.3	20.8	1.8	3.39
Olsztyński	14.8	3.4	17.2	1.8	1.89
Kaliski	17.5	3.5	16.7	1.7	2.11
Koniński	21.3	3.8	19.4	1.7	2.20
Leszczyński	18.0	5.1	17.0	2.2	2.31
Pilski	21.6	5.7	19.8	1.9	2.97
Poznański	13.4	3.8	11.0	1.9	1.98
City of Poznań	7.7	3,1	8,5	2,0	1,58
Koszaliński	21.9	4.6	16.6	2.0	2.32
Stargardzki	17.3	8.3	18.7	2.1	4.03
City of Szczecin	11.6	4.1	9.6	1.7	2.38
Szczeciński	16.5	6.6	12.1	1.8	3.73
Mean	18.4	4.0	17.6	1.8	2.17
Standard deviation	7.6	1.3	6.0	0.2	0.57
Minimum	6.2	1.3	6.3	1.1	1.16
Lower quartile	12.9	3.2	12.8	1.7	1.86
Median	17.4	3.8	17.2	1.8	2.04
Upper quartile	21.9	4.7	21.9	2.0	2.36
Maximum	40.9	8.3	30.2	2.6	4.03

* Tri–City is a metropolitan area in Poland consisting of three cities: Gdańsk, Gdynia and Sopot.

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 637–650, DOI 10.21307/stattrans-2017-004

A NEW ESTIMATOR OF MEAN USING DOUBLE SAMPLING

Kalyan Rao Vadlamudi[1], **Stephen A. Sedory, Sarjinder Singh**

ABSTRACT

In this paper, we consider the problem of estimation of population mean of a study variable by making use of first-phase sample mean and first-phase sample median of the auxiliary variable at the estimation stage. The proposed new estimator of the population mean is compared to the sample mean estimator, ratio estimator and the difference type estimator for the fixed cost of the survey by using the concept of two-phase sampling. The magnitude of the relative efficiency of the proposed new estimator has been investigated through simulation study.

Key words: Two-phase sampling, relative efficiency, analytical and empirical comparison.

1.Introduction

Consider a population Ω consisting of N units. Let (y_i, x_i), $i = 1,2,..., N$ be the values of the study variable Y and auxiliary variable X for the ith unit in the population.

Let

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1.1}$$

and

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.2}$$

be the population means of the study and auxiliary variables respectively. Survey statisticians are often interested in estimating the population mean $\bar{Y}$ of the study variable. It is also well known that if the population mean $\bar{X}$ of an auxiliary variable is known then it can be used to improve estimation strategies in survey sampling. Examples of such estimators are the ratio estimator due to Cochran

[1] Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX 78363, USA.

(1940) and the linear regression estimator due to Hansen, Hurwitz and Madow (1953).

If such auxiliary information of the population mean $\bar{X}$ is not known or complete auxiliary information is not known, then it can be relatively cheaper to obtain information on the auxiliary variable by taking a large preliminary sample for estimating population mean of the auxiliary variable to be used at the estimation stage. In other words, in the case of single auxiliary variable X, if the population mean $\bar{X}$ of the auxiliary variable is unknown then we consider taking a preliminary large sample of m units by using simple random and without replacement sampling (SRSWOR) from the population of N units. In the sample s_1 of m units, we observe only the auxiliary variable x_i , $i = 1,2,...,m$. From the given first-phase sample s_1 of m units, we select another subsample s_2 of n units by using SRSWOR sample. In the sample s_2, we measure the ordered pairs (y_i, x_i), $i = 1,2,...,n$ of the study and auxiliary variables. Then an unbiased estimator of the population mean $\bar{X}$ based on the first-phase sample information as the sample mean is given by:

$$\bar{x}_m = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1.3}$$

The unbiased estimators of the population means $\bar{Y}$ and $\bar{X}$ of the study and auxiliary variables based on the second-phase sample information are, respectively, given by:

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.4}$$

and

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.5}$$

Neyman (1938) invented this sampling technique called double sampling or two-phase sampling, and later work related to this scheme is extensively reviewed in Singh (2003). It leads to ratio and regression type estimators of the population mean $\bar{Y}$ in two-phase sampling as:

$$\bar{y}_{\text{rat(d)}} = \bar{y}_n\left(\frac{\bar{x}_m}{\bar{x}_n}\right) \tag{1.6}$$

and

$$\bar{y}_{\text{reg(d)}} = \bar{y}_n + \beta\left(\bar{x}_m - \bar{x}_n\right) \tag{1.7}$$

The variances of the sample mean, ratio and regression type estimators are, respectively, given by;

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

$$V(\bar{y}_{\mathrm{rat(d)}}) = \left(\frac{1}{m} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)\left[S_y^2 + R^2 S_x^2 - 2RS_{xy}\right] \tag{1.8}$$

and

$$V(\bar{y}_{\mathrm{reg(d)}}) = \left(\frac{1}{m} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right) S_y^2\left[1 - \rho_{xy}^2\right] \tag{1.9}$$

where

$$R = \frac{\bar{Y}}{\bar{X}};\quad \rho_{xy} = \frac{S_{xy}}{S_x S_y};\quad S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X});$$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2, \text{ and}$$

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2.$$

To our knowledge, the pioneer contributors, to the problem of estimating median, are Kuk and Mak (1989) by proposing very clear estimators of median in the presence of auxiliary information. Singh, Joarder and Tracy (2001) extended their idea to the situation of median estimation in two-phase sampling. The importance of double sampling and improvements on the estimation of population mean can also be seen in several publications by Vishwakarma and Kumar (2015), Vishwakarma and Gangele (2014), Vishwakarma and Singh (2011), Amin *et al*. (2016), and Sanaullah *et al*. (2014). However, none of these papers deal with the situation of making use of estimator of median of the auxiliary variable at the estimation stage of population mean of the study variable in two-phase sampling. This motivated the authors to think on these lines if some improvements can be seen by making the use of first-phase median of the auxiliary variable.

In the next section, we introduce a new estimator of the population mean in two-phase sampling which makes use of first-phase sample mean and sample median of the auxiliary variable.

2. Estimator

Let $\hat{M}_x^*$ be the median of the auxiliary variable X based on the first phase sample s_1 of m units. Let $\hat{M}_x$ be the median of the auxiliary variable X based on the second phase sample s_2 of n units.

Suppose $x_{(1)}, x_{(2)},..., x_{(m)}$ are the x values of first-phase sample units in the ascending order. Further, let t_1 be an integer such that $X_{(t_1)} \le M_x \le X_{(t_1+1)}$ and let $p_1 = t_1/n$ be the proportion of X values in the first phase sample that are less than or equal to the median value M_x, an unknown population parameter. If $\hat{p}_1$ is a predictor of p_1, the first-phase sample median $\hat{M}_x^*$ can be written in terms of quintiles as $\hat{Q}_x(\hat{p}_1)$, where $\hat{p}_1 = 0.5$.

Suppose $x_{(1)}, x_{(2)},..., x_{(n)}$ are the x values of second-phase sample units in the ascending order. Further, let t_2 be an integer such that $X_{(t_2)} \le M_x \le X_{(t_2+1)}$ and let $p_2 = t_2/n$ be the proportion of X values in the second phase sample that are less than or equal to the median value M_x, an unknown population parameter. If $\hat{p}_2$ is a predictor of p_2, the first-phase sample median $\hat{M}_x$ can be written in terms of quintiles as $\hat{Q}_x(\hat{p}_2)$, where $\hat{p}_2 = 0.5$.

Now we define a new estimator of the population mean $\bar{Y}$ in two-phase sampling as:

$$\bar{y}_{kal} = \bar{y}_n + \beta_1^*(\bar{x}_m - \bar{x}_n) + \beta_2^*(\hat{M}_x^* - \hat{M}_x) \tag{2.1}$$

where β_1^* and β_2^* are unknown partial regression coefficients to be determined such that the variance of the estimator is minimum.

It may be worth pointing out that the proposed estimator $\bar{y}_{kal}$ is an extension of the recent estimator of due to Lamichhane, Singh and Diawara (2015) from single-phase sampling to two-phase sampling.

To study the asymptotic properties of the proposed estimator, $\bar{y}_{kal}$, let us define the following error terms:

$$\varepsilon_0 = \frac{\bar{y}_n}{\bar{Y}} - 1;\ \ \varepsilon_1 = \frac{\bar{x}_n}{\bar{X}} - 1;\ \ \varepsilon_2 = \frac{\hat{M}_x}{M_x} - 1;\ \ \varepsilon_3 = \frac{\bar{x}_m}{\bar{X}} - 1 \text{ and } \varepsilon_4 = \frac{\hat{M}_x^*}{M_x} - 1$$

such that

$$E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0;\ E(\varepsilon_3) \approx E(\varepsilon_4) \approx 0$$

$$E(\varepsilon_0^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_y^2;\ \ E(\varepsilon_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_x^2;$$

$$E(\varepsilon_2^2) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{4\{f_x(M_x)\}^2 M_x^2}$$

$$E(\varepsilon_3^2)=\left(\frac{1}{m}-\frac{1}{N}\right)C_x^2;\;\; E(\varepsilon_4^2)=\left(\frac{1}{m}-\frac{1}{N}\right)\frac{1}{4\{f_x(M_x)\}^2M_x^2};$$

$$E(\varepsilon_0\varepsilon_1)=\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{xy}C_xC_y;\;\; E(\varepsilon_1\varepsilon_3)=\left(\frac{1}{m}-\frac{1}{N}\right)C_x^2;$$

$$E(\varepsilon_0\varepsilon_2)=-\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{YMx}\{f_x(M_x)\}^{-1}}{\bar{Y}\,M_x};\; E(\varepsilon_0\varepsilon_3)=\left(\frac{1}{m}-\frac{1}{N}\right)\rho_{xy}C_xC_y;$$

$$E(\varepsilon_0\varepsilon_4)=-\left(\frac{1}{m}-\frac{1}{N}\right)\frac{S_{YM_x}\{f_x(M_x)\}^{-1}}{\bar{Y}\,M_x};$$

$$E(\varepsilon_1\varepsilon_2)=-\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{XMx}\{f_x(M_x)\}^{-1}}{\bar{X}M_x}$$

$$E(\varepsilon_1\varepsilon_4)=-\left(\frac{1}{m}-\frac{1}{N}\right)\frac{S_{XM_x}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x};$$

$$E(\varepsilon_2\varepsilon_3)=-\left(\frac{1}{m}-\frac{1}{N}\right)\frac{S_{XM_x}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x}$$

$$E(\varepsilon_2\varepsilon_4)=\left(\frac{1}{m}-\frac{1}{N}\right)\frac{1}{4\{f_x(M_x)\}^2M_x^2};\;\text{ and}$$

$$E(\varepsilon_3\varepsilon_4)=-\left(\frac{1}{m}-\frac{1}{N}\right)\frac{S_{XMx}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x}$$

where

$$S_{XM_x}=\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)\left(I_{x_i}-0.5\right),$$

$$f=n/N,$$

$$S_{YM_x}=\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)\left(I_{x_i}-0.5\right)$$

and

$$I_{x_i}=\begin{cases}1, & \text{if} \quad X_i\le M_x\\ 0, & \text{otherwise}\end{cases}$$

Note that we used the following main result from Kuk and Mak (1989) in deriving the variance and co-variance expressions for the sample means and

sample median, that is, if F_x be the cumulative distribution function of X, then the sample median can be approximated as:

$$\hat{M}_x = M_x + (0.5 - p_x) f_x^{-1}(M_x) + ...$$

where p_x be the proportion of I_{x_i} values taking a value of 1.

The new estimator $\bar{y}_{kal}$ in terms of ε_i, $i = 0,1,2,3,4$ can be written as

$$\begin{aligned}\bar{y}_{kal} &= \bar{y}_n + \beta_1^*(\bar{x}_m - \bar{x}_n) + \beta_2^*(\hat{M}_x^{|} - \hat{M}_x) \\ &= \bar{Y}(1+\varepsilon_0) + \beta_1^*[\bar{X}(1+\varepsilon_3) - \bar{X}(1+\varepsilon_1)] + \beta_2^*[M_x(1+\varepsilon_4) - M_x(1+\varepsilon_2)] \\ &= \bar{Y} + \bar{Y}\varepsilon_0 + \beta_1^*\bar{X}(\varepsilon_3 - \varepsilon_1) + \beta_2^* M_x(\varepsilon_4 - \varepsilon_2) \qquad (2.2)\end{aligned}$$

Now we have the following theorems:

Theorem 2.1. The new proposed estimator $\bar{y}_{kal}$ is an unbiased estimator of the population mean $\bar{Y}$.

Proof. Taking expected value on both sides of (2.2), we get

$$\begin{aligned}E(\bar{y}_{kal}) &= E\left[\bar{Y} + \bar{Y}\ \varepsilon_0 + \beta_1^*\bar{X}(\varepsilon_3 - \varepsilon_1) + \beta_2^* M_x(\varepsilon_4 - \varepsilon_2)\right] \\ &= \bar{Y} + \bar{Y}\ E(\varepsilon_0) + \beta_1^*\bar{X}(E(\varepsilon_3) - E(\varepsilon_1)) + \beta_2^* M_x(E(\varepsilon_4) - E(\varepsilon_2)) \\ &= \bar{Y} + 0 \\ &= \bar{Y} \qquad (2.3)\end{aligned}$$

Thus the new estimator $\bar{y}_{kal}$ is an unbiased estimator of $\bar{Y}$ and it proves the theorem.

Theorem 2.2. The minimum variance, to the first order of approximation, of the new proposed estimator $\bar{y}_{kal}$ is given by

$$Min.V(\bar{y}_{kal}) = \left(\frac{1}{m} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy}S_{XM_x})^2}{S_y^2 S_x^2\left(0.25S_x^2 - S_{XM_x}^2\right)}\right) \qquad (2.4)$$

Proof. By the definition of variance, the variance of the unbiased estimator $\bar{y}_{kal}$ is given by

$$V(\bar{y}_{kal}) = E\left[\bar{y}_{kal} - \bar{Y}\right]^2$$

$$= E\left[\bar{Y}\varepsilon_0 + \beta_1^* \bar{X}(\varepsilon_3 - \varepsilon_1) + \beta_2^* M_x(\varepsilon_4 - \varepsilon_2)\right]^2$$

$$= E\Bigg[\bar{Y}^2\varepsilon_0^2 + \beta_1^{*2}\bar{X}^2(\varepsilon_3 - \varepsilon_1)^2 + \beta_2^{*2}M_x^2(\varepsilon_4 - \varepsilon_2)^2$$

$$-2\beta_1^*\bar{X}\bar{Y}\varepsilon_0(\varepsilon_3 - \varepsilon_1) - 2\beta_2^* M_x\bar{Y}\varepsilon_0(\varepsilon_4 - \varepsilon_2) + 2\beta_1^*\beta_2^*\bar{X}M_x(\varepsilon_3 - \varepsilon_1)(\varepsilon_4 - \varepsilon_2)\Bigg] \tag{2.5}$$

Further note that

$$E(\varepsilon_3 - \varepsilon_1)^2 = \left(\frac{1}{n} - \frac{1}{m}\right)C_x^2 \tag{2.6}$$

$$E(\varepsilon_4 - \varepsilon_2)^2 = \left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{4\{f_x(M_x)\}^2 M_x^2} \tag{2.7}$$

$$E(\varepsilon_0\varepsilon_3 - \varepsilon_0\varepsilon_1) = -\left(\frac{1}{n} - \frac{1}{m}\right)\rho_{xy}C_xC_y \tag{2.8}$$

$$E(\varepsilon_0\varepsilon_4 - \varepsilon_0\varepsilon_2) = \left(\frac{1}{n} - \frac{1}{m}\right)\frac{S_{YM_X}\{f_x(M_x)\}^{-1}}{\bar{Y}\,M_x} \tag{2.9}$$

and

$$E(\varepsilon_3\varepsilon_4 - \varepsilon_1\varepsilon_4 - \varepsilon_3\varepsilon_2 + \varepsilon_1\varepsilon_2) = -\left(\frac{1}{n} - \frac{1}{m}\right)\frac{S_{XM_X}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x} \tag{2.10}$$

On substituting (2.6) – (2.10) into (2.5), we have

$$V(\bar{y}_{kal}) = \bar{Y}^2\varepsilon_0^2 + \beta_1^{*2}\bar{X}^2E(\varepsilon_3 - \varepsilon_1)^2 + \beta_2^{*2}M_x^2E(\varepsilon_4 - \varepsilon_2)^2$$

$$-2\beta_1^*\bar{X}\bar{Y}E(\varepsilon_0\varepsilon_3 - \varepsilon_0\varepsilon_1) - 2\beta_2^*M_x\bar{Y}E(\varepsilon_0\varepsilon_4 - \varepsilon_0\varepsilon_2) + 2\beta_1^*\beta_2^*\bar{X}M_xE(\varepsilon_3 - \varepsilon_1)(\varepsilon_4 - \varepsilon_2)$$

$$= \bar{Y}^2\left(\frac{1}{n} - \frac{1}{N}\right)C_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)\beta_1^{*2}\bar{X}^2C_x^2 + \beta_2^{*2}M_x^2\left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{4\{f_x(M_x)\}^2M_x^2}$$

$$+2\beta_1^*\bar{X}\,\bar{Y}\left(\left(\frac{1}{n} - \frac{1}{m}\right)\rho_{xy}C_xC_y\right) - 2\beta_2^*\bar{Y}M_x\left(\frac{1}{n} - \frac{1}{m}\right)\frac{S_{YM_X}\{f_x(M_x)\}^{-1}}{\bar{Y}\,M_x}$$

$$-2\beta_1^*\beta_2^*\bar{X}M_x\left(\frac{1}{n} - \frac{1}{m}\right)\frac{S_{XM_X}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x} \tag{2.11}$$

On differentiating $V(\bar{y}_{kal})$ with respect to β_1^* and β_2^* and each equating to zero, we have,

$$\frac{\partial V(\bar{y}_{kal})}{\partial \beta_1^*} = 0 \text{, and } \frac{\partial V(\bar{y}_{kal})}{\partial \beta_2^*} = 0$$

On solving the system of linear equations, the optimum values of β_1^* and β_2^* are given by

$$\beta_1^* = \frac{\{f_x(M_x)\}^{-2} S_{YM_x} S_{XM_x} - S_{xy} S_y^2}{\{f_x(M_x)\}^{-2}\left(\frac{1}{4}S_x^2 - S_{XM_x}^2\right)} \tag{2.12}$$

and

$$\beta_2^* = \frac{S_x^2 S_{YM_x} - S_{xy} S_{XM_x}}{\{f_x(M_x)\}^{-1}\left(\frac{1}{4}S_x^2 - S_{XM_x}^2\right)} \tag{2.13}$$

On substituting the optimum values of β_1^* and β_2^*, the minimum variance is given by

$$Min.V(\bar{y}_{kal}) = \left(\frac{1}{m} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)\left(S_y^2 - \frac{S_{xy}^2}{S_x^2} - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_x^2\left(0.25S_x^2 - S_{XM_x}^2\right)}\right)$$

$$= \left(\frac{1}{m} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25S_x^2 - S_{XM_x}^2\right)}\right) \tag{2.14}$$

which proves the theorem.

Remark: It may be worth pointing out that the replacement of β_1^* and β_2^* with their consistent estimates lead to a new estimator of the population mean in two two-phase sampling which has same mean square error up to the first order of approximation as the minimum variance in (2.14). Such changes do not affect the results up to the first order of approximation. ((6.1) and (6.2) in Singh, Singh, and Upadhayaya (2007)).

3. Comparison of different estimators

Note that

$$0.25S_x^2 - S_{XM_x}^2 > 0$$

$$\frac{S_{XM_x}^2}{S_x^2} < \frac{1}{4} \tag{3.1}$$

Thus the proposed new estimator $\bar{y}_{kal}$ is always more efficient than the sample mean, ratio and the regression type estimator in two-phase sampling. It may be worth pointing out that the final minimum variance of the proposed estimator is free from the value of $f_x(M_x)$, hence from computational point.

In the next section, we focus on the cost analysis in two-phase sampling because it is considered as one among the list of cost effective sampling schemes in survey sampling.

4. Cost Analysis

In this section we consider comparison of different estimators with cost aspects. Let C_0 be the overhead cost, C_1 be the cost of information from one unit in the first phase sample; and C_2 be the cost of information from one unit in the second phase sample. Note that the value of C_1 is always smaller than that of C_2 . Thus the total cost function is given by

$$C = C_0 + mC_1 + nC_2 \tag{4.1}$$

Now we have the following results

For the fixed cost C of the survey, the minimum variance of the proposed estimator $\bar{y}_{kal}$ is given by

$$Min.V(\bar{y}_{kal})_{fixed\_cost} =$$

$$= \frac{\left(\sqrt{C_1 S_y^2\left(\rho_{xy}^2 + \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)} + \sqrt{C_2 S_y^2\left[1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right]}\right)^2}{C - C_0} - \frac{S_y^2}{N} \tag{4.2}$$

The optimum values of m and n are given by

$$m = m_{kal} = \frac{(C - C_0)\sqrt{S_y^2\left(\rho_{xy}^2 + \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)}}{\sqrt{C_1}\left[\sqrt{C_1 S_y^2\left(\rho_{xy}^2 + \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)} + \sqrt{C_2 S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)}\right]} \tag{4.3}$$

and

$$n = n_{kal} = \frac{(C - C_0)\sqrt{S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2 \left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)}}{\sqrt{C_1}\left[\sqrt{C_1 S_y^2\left(\rho_{xy}^2 + \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2 \left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)} + \sqrt{C_2 S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2 \left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)}\right]} \tag{4.4}$$

which proves the theorem.

In case of single-phase sampling, the total cost function is given by

$$C = C_0 + nC_2 \tag{4.5}$$

From (4.11), we have the optimum sample size as:

$$n = \frac{C - C_0}{C_2} = n_u \text{ (say)} \tag{4.6}$$

The variance of the sample mean estimator is given by

$$V(\bar{y}_{\mathrm{n}})_{fixed\_\mathrm{cost}} = \left(\frac{C_2}{C - C_0} - \frac{1}{N}\right) S_y^2 \tag{4.7}$$

The optimum values of m and n for the ratio estimator are given by

$$m = \frac{(C - C_0)\sqrt{2RS_{xy} - R^2 S_x^2}}{\sqrt{C_1}\left[\sqrt{C_1(2RS_{xy} - R^2 S_x^2)} + \sqrt{C_2(S_y^2 + R^2 S_x^2 - 2RS_{xy})}\right]} = m_{rat} \tag{4.8}$$

and

$$n = \frac{(C - C_0)\sqrt{(S_y^2 + R^2 S_x^2 - 2RS_{xy})}}{\sqrt{C_1}\left[\sqrt{C_1(2RS_{xy} - R^2 S_x^2)} + \sqrt{C_2(S_y^2 + R^2 S_x^2 - 2RS_{xy})}\right]} = n_{rat} \tag{4.9}$$

The minimum variance of the ratio estimator for the fixed cost is given by:

$$Min.V(\bar{y}_{rat})_{fixt\_\mathrm{cost}} = \frac{\left[\sqrt{C_1(2RS_{xy} - R^2 S_x^2)} + \sqrt{C_2(S_y^2 + R^2 S_x^2 - 2RS_{xy}}\right]^2}{C - C_0} - \frac{S_y^2}{N} \tag{4.10}$$

The optimum values of m and n for the regression type estimator $\bar{y}_{reg(d)}$ are given by

$$m = \frac{(C - C_0)\sqrt{S_y^2 \rho_{xy}^2}}{\sqrt{C_1}\left[\sqrt{C_1 S_y^2 \rho_{xy}^2} + \sqrt{C_2 S_y^2 \left(1 - \rho_{xy}^2\right)}\right]} = m_{reg} \tag{4.11}$$

and

$$n = \frac{(C - C_0)\sqrt{S_y^2 \left(1 - \rho_{xy}^2\right)}}{\sqrt{C_1}\left[\sqrt{C_1 S_y^2 \rho_{xy}^2} + \sqrt{C_2 S_y^2 \left(1 - \rho_{xy}^2\right)}\right]} = n_{reg} \tag{4.12}$$

Note that for the fixed cost of the survey, the variances of the regression type estimator $\bar{y}_{reg(d)}$ is given by

$$Min.V(\bar{y}_{\text{reg}})_{fixed\_\text{cost}} = \frac{\left(\sqrt{C_1 S_y^2 \rho_{xy}^2} + \sqrt{C_2 S_y^2 (1 - \rho_{xy}^2)}\right)^2}{C - C_0} - \frac{S_y^2}{N} \tag{4.13}$$

The percent relative efficiency of the new proposed estimator $\bar{y}_{kal}$ with respect to the sample mean estimator $(\bar{y}_n)$, ratio estimator $(\bar{y}_{\text{rat(d)}})$, and regression type estimators $(\bar{y}_{\text{reg(d)}})$:

$$RE(0) = RE(\text{sample mean}) = \frac{Min.V(\bar{y}_n)_{fixed\_\text{cost}}}{Min.V(\bar{y}_{kal})_{fixed\_\text{cost}}} \times 100\% \tag{4.14}$$

$$RE(1) = RE(\text{ratio}) = \frac{Min.V(\bar{y}_{rat})_{fixed\_\text{cost}}}{Min.V(\bar{y}_{kal})_{fixed\_\text{cost}}} \times 100\% \tag{4.15}$$

$$RE(2) = RE(\text{reg}) = \frac{Min.V(\bar{y}_{reg})_{fixed\_\text{cost}}}{Min.V(\bar{y}_{kal})_{fixed\_\text{cost}}} \times 100\% \tag{4.16}$$

In order to see the magnitude of the relative efficiency of the proposed estimator $\bar{y}_{kal}$ over the mean, ratio and the regression type estimator, we did simulation study.

5. Simulation Study

From (3.1), it is clear that the maximum value of $S_{XM_x} < \frac{S_x}{2}$. We consider many populations of size $N = 50{,}000$, $\bar{X} = 50$, $\bar{Y} = 20$, $S_y^2 = 5$, $S_x^2 = 10$ and

different values of the correlation coefficient ρ_{xy}. Note that it is very likely that the possible value of $0 < S_{XM_X} < 0.5$ and $0 < S_{YM_X} < 0.5$. Consider a situation of having total cost $C = \$5000$, overhead cost $C_0 = \$500$, cost of collecting information from one unit in the first-phase sample $C_1 = \$4$, and the cost of collecting information from one unit in the sample $C_2 = \$10$. We wrote R-codes to compute the percent relative efficiency values and the optimum sample sizes for the four estimators. There are many situations where the proposed estimator performs better than the existing estimators, and in the simulation study we stored only those combinations of ρ_{xy}, S_{XM_X} and S_{YM_X} where the value of $RE(2)$ is greater than 105%. In other words, the proposed estimator is at least 105% more efficient than the linear regression type estimator in double sampling. The results so obtained are presented in Table 5.1

Table 5.1. Percent relative efficiency of the proposed new estimator over the three estimators and optimum sample sizes

ρ_{xy}	S_{XM_X}	S_{YM_X}	n_u	m_{rat}	n_{rat}	m_{reg}	n_{reg}	m_{kal}	n_{kal}	$RE(0)$	$RE(1)$	$RE(2)$
0.80	0.05	0.40	450	483	406	515	386	540	370	107.7	108.3	105.4
0.80	0.05	0.45	450	483	406	515	386	550	364	110.1	110.6	107.8
0.80	0.05	0.50	450	483	406	515	386	565	354	113.4	114.0	111.0
0.80	0.10	0.45	450	483	406	515	386	544	368	108.7	109.2	106.4
0.80	0.10	0.50	450	483	406	515	386	556	360	111.4	112.0	109.1
0.80	0.15	0.45	450	483	406	515	386	539	371	107.5	108.1	105.3
0.80	0.15	0.50	450	483	406	515	386	549	364	109.8	110.4	107.5
0.80	0.20	0.50	450	483	406	515	386	543	368	108.5	109.0	106.2
0.80	0.25	0.50	450	483	406	515	386	538	371	107.4	107.9	105.1
0.85	0.05	0.35	450	516	385	568	352	600	332	115.4	112.8	106.8
0.85	0.05	0.40	450	516	385	568	352	616	322	118.9	116.3	110.1
0.85	0.05	0.45	450	516	385	568	352	637	309	123.9	121.2	114.8
0.85	0.05	0.50	450	516	385	568	352	670	288	131.7	128.8	122.0
0.85	0.10	0.35	450	516	385	568	352	594	336	113.8	111.3	105.4
0.85	0.10	0.40	450	516	385	568	352	606	328	116.6	114.1	108.0
0.85	0.10	0.45	450	516	385	568	352	623	317	120.7	118.1	111.8
0.85	0.10	0.50	450	516	385	568	352	649	301	126.7	123.9	117.3
0.85	0.15	0.40	450	516	385	568	352	598	333	114.8	112.3	106.4
0.85	0.15	0.45	450	516	385	568	352	613	324	118.2	115.6	109.4
0.85	0.15	0.50	450	516	385	568	352	633	311	122.9	120.2	113.8
0.85	0.20	0.45	450	516	385	568	352	604	330	116.1	113.6	107.5
0.85	0.20	0.50	450	516	385	568	352	620	319	120.0	117.3	111.1
0.85	0.25	0.45	450	516	385	568	352	596	334	114.4	111.9	106.0
0.85	0.25	0.50	450	516	385	568	352	610	326	117.6	115.0	108.9
0.85	0.30	0.50	450	516	385	568	352	602	331	115.7	113.2	107.1

Discussion: From the simulation study, it is observed that a high value of correlation coefficient ρ_{xy} and a high value of S_{YM_X} is required for the proposed new estimator to show percent relative efficiency of at least 105%. For example, if $\rho_{xy} = 0.80$, $S_{XM_X} = 0.05$, $S_{YM_X} = 0.40$, with a total cost of \$5000, if instead of we use only single phase sample mean estimator with optimum sample size $n_u = 450$, we should use the proposed estimator with optimum sample sizes $m_{kal} = 540$ and $n_{kal} = 370$, then the percent relative efficiency value is $RE(0) = 107.7\%$. Instead of using the ratio estimator with optimum sample sizes $m_{rat} = 483$ and second-phase sample size $n_{rat} = 406$, if one uses the proposed estimator with optimum sample sizes $m_{kal} = 540$ and $n_{kal} = 370$ then the percent relative efficiency value is $RE(1) = 108.3\%$. In the same way, if we use the proposed estimator with optimum sample sizes $m_{kal} = 540$ and $n_{kal} = 370$, then the relative efficiency of the proposed estimator over the regression method of estimation with optimum sample sizes $m_{reg} = 515$ and $n_{reg} = 386$ is $RE(2) = 105.4\%$.

Note that $n_{kal}(= 370) < n_{reg}(= 386) < n_{rat}(= 406) < n_u(= 450)$. The optimum second-phase sample size remains lowest in case of the proposed estimator, thus the proposed estimator reduces efforts for collecting data on the second-phase of the sample for the fixed cost of the survey and provides efficient results. The rest of the results in Table 5.1 can also be interpreted in the same way. We conclude that there exist several situations where the proposed new estimator can be used more efficiently for a fixed cost of the survey.

Acknowledgements

The authors are grateful to the Admin: Patryk Barszcz, Editorial Office, Statistics in Transition new series, and a referee for bringing the original manuscript in the presnt form.

REFERENCES

AMIN, M. N. U., SHAHBAZ, M. Q., KADILAR, C. (2016). Ratio estimators for population mean using robust regression in double sampling, Gael University Journal of Science, 29 (4), pp. 793–798.

COCHRAN, W. G., (1940). Some properties of estimators based on sampling scheme with varying probabilities, Austral. J. Statist., 17, pp. 22–28.

HANSEN, M. H., HURWITZ, W. N., MADOW, W. G., (1953). Sample survey methods and theory, New York, John Wiley and Sons, pp. 456–464.

KUK, A. Y. C., MAK, T. K. (1989). Median estimation in the presence of auxiliary information, J. R. Statist. Soc., B, 51, pp. 261–269.

LAMICHHANE, R., SINGH, S., DIAWARA, N., (2015). Improved Estimation of Population Mean Using Known Median of Auxiliary Variable, Communications in Statistics-Simulation and Computation, DOI: 10.1080/03610918.2015.1062102.

NEYMAN, J., (1938). Contribution to the theory of sampling human populations. J. Amer. Statist. Assoc., 33, pp. 101–116.

SANAULLAH, A., ALI, H. A., AMIN, M. N. U., HANIF, M., (2014). Generalized exponential chain ratio estimator under stratified two-phase random sampling. Applied Mathematics and Computation, 226, pp. 541–547.

SINGH, S., (2003). Advanced sampling theory with applications: How Michael selected Amy. Kluwer: The Academic Publisher, The Netherlands.

SINGH, S., JOARDER A. H., TRACY D. S., (2001). Median estimation using double sampling. Australian & New Zealand J. Statist., pp. 43, 33–46.

SINGH, S., SINGH, H. P., UPADHAYAYA, L. N., (2007). Chain ratio and regression type estimators for median estimation in survey sampling. Statistical Papers, 48 (1), pp. 23–46.

VISHWAKARMA, G. K., SINGH, H. P., (2011). Separate ratio-product estimator for estimating population mean using auxiliary information. Journal of Statistical Theory and Applications, 10 (4), pp. 653–664.

VISHWAKARMA, G. K., GANGELE, R. K., (2014). A class of chain ratio-type exponential estimators in double sampling using two auxiliary variates. Applied Mathematics and Computation, 227, pp. 171–175.

VISHWAKARMA, G. K., KUMAR, M., (2015). An efficient class of estimators for the mean of a finite population in two-phase sampling using multi-auxiliary variates, Commun. Math. Stat., 3, pp. 477–489.

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 651–668, DOI 10.21307/stattrans-2017-005

RELATIONS FOR MOMENTS OF PROGRESSIVELY TYPE-II RIGHT CENSORED ORDER STATISTICS FROM ERLANG-TRUNCATED EXPONENTIAL DISTRIBUTION

Mansoor Rashid Malik[1], **Devendra Kumar**[2]

ABSTRACT

In this paper, we establish some new recurrence relations for the single and product moments of progressively Type-II right censored order statistics from the Erlang-truncated exponential distribution. These relations generalize those established by Aggarwala and Balakrishnan (1996) for standard exponential distribution. These recurrence relations enable computation of mean, variances and covariances of all progressively Type-II right censored order statistics for all sample sizes in a simple and efficient manner. Further an algorithm is discussed which enable us to compute all the means, variances and covariances of Erlang-truncated exponential progressive Type-II right censored order statistics for all sample sizes n and all censoring schemes $(R_1, R_2,..., R_m)$, $m < n$. By using these relations, we tabulate the means and variances of progressively Type-II right censored order statistics of the Erlang-truncated exponential distribution.

Key words: Censoring, progressive Type-II right censored order statistics, single moments, product moments, recurrence relations, Erlang-truncated exponential distribution.

1. Introduction

Practitioners and statisticians are often faced with incomplete or censored data. In life testing, censored samples are present whenever the experimenter does not observe the failure times of all units placed on the life test. This may happen intentionally or unintentionally and may be caused, e.g. by time constraints on the test duration like in Type-I censoring, by requirements on the minimum number of observed failures, or by the structure of a technical system. Naturally, the probabilistic structure of the resulting incomplete data depends heavily on the censoring mechanism and so suitable inferential procedures become necessary. Progressive censoring can be described as a censoring method where units under test are removed from the life

[1]Department of Statistics, Amity University, Noida, India. E-mail: mansoormalik884@gmail.com
[2]Corresponding Author: Department of Statistics, Central University of Haryana. E-mail: devendrastats@gmail.com

test at some prefixed or random inspection times. It allows for both failure and time censoring. Many modifications of the standard model have been developed, but the basic idea can be easily described by progressive Type-II censoring, which can also be considered as the most popular model. Under this scheme of censoring, from a total of n units placed simultaneously on a life test, only m are completely observed until failure. Then, given a censoring plan $(R_1, R_2, \cdots, R_m)$:
i) At the time $x_{1,m:n}$ of the first failure, R_1 of the $n-1$ surviving units are randomly withdrawn (or censored) from the life-testing experiment.
ii) At the time $x_{2,m:n}$ of the next failure, R_2 of the $n-2-R_1$ surviving units are censored, and so on.
iii) Finally, at the time $x_{m,m:n}$ of the mth failure, all the remaining $R_m = n-m-R_1-R_2-\cdots-R_{m-1}$ surviving units are censored.

Note that censoring takes place here progressively in m stages. Clearly, this scheme includes the complete sample situation and the conventional Type-II right censoring scenario as special cases. The ordered failure times $X_{1,m:n}^{(R_1,R_2,\cdots,R_m)} \leq X_{2,m:n}^{(R_1,R_2,\cdots,R_m)} \leq \cdots \leq X_{m,m:n}^{(R_1,R_2,\cdots,R_m)}$ arising from such a progressively Type-II right censored sample are called progressively Type-II censored order statistics. These are natural generalizations of the usual order statistics that were studied quite extensively during the past century. For more details one can refer to Balakrishnan and Cramer (2014). The following notations are used throughout this paper. In progressive censoring, the following notations are used:

(i) $n, m, R_1, R_2 .... R_m$ are all integers.

(ii) m is the sample size (which may be random in some models).

(iii) n is the total number of units in the experiment.

(iv) R_j is the number of (effectively employed) removals at the $j-th$ censoring time.

(v) $(R_1, R_2, \cdots, R_m)$ denotes the censoring scheme.

Recurrence relations for single and product moments for any continuous distribution can be used to compute all means, variances and covariance of such a distribution. Several authors obtained the recurrence relation for progressively type-II right censored order statistics for different distributions such as Cohin (1963), Mann [(1971), Thomas and Wilson (1972), Arnold (1992), Balakrishnan et al. (2012), Nikulin and Haghighi (2006), Nadarajah and Haghighi (2011), Joshi (1978), Balakrishnan and Malik (1986), Arnold et al. (1992), Viveros and Balakrishnan (1994), Balakrishnan and Sandhu (1995), Aggarwala and Balakrishnan (1996), Balakrishnan and Aggarwala (1998), Balakrishnan and Sultan (1998), Abd El-Baset and Mohammed (2003), David and Nagaraja (2003), Fernandez (2004), Balakrishnan et al. (2004), Sultan et al. (2006), Mahmoud et al. (2006), Balakrishnan (2007), Bal-

akrishnan et al. (2011), Balakrishnan and Saleh (2013). The moments of order statistics have been recursively derived, see Salah et al. (2008), Kumar and Sanku (2017) and Kumar et al. (2017), for a complete sample.

If the failure times are based on an absolutely continuous distribution function $F(x)$ with probability density function $f(x)$, the joint probability density function of the progressively censored failure times $X_{1:m:n}, X_{2:m:n}, \cdots, X_{m:m:n}$, is given by [see Balakrishnan and Aggarwala (2000).

$$f_{X_{1:m:n},X_{2:m:n},\ldots,X_{m:m:n}}(x_1,x_2,\cdots,x_m)$$
$$= A(n,m-1)\prod_{i=0}^{m} f(x_i)[1-F(x_i)]^{R_i}, -\infty < x_1 < x_2 < \cdots < x_m < \infty, \qquad (1)$$

where $f(x)$ and $F(x)$ are respectively the pdf and the cdf of the random sample and

$$A(n,m-1) = n(n-R_1-1)\cdots(n-R_1-R_2-\cdots-R_{m-1}-m+1). \qquad (2)$$

The exponential distribution is the simplest distribution in terms of expression and analytical tractability. It is also widely used in reliability engineering. There is no doubt that the wide applicability of the exponential distribution even in inappropriate scenarios is motivated by its simplicity. However, the exponential distribution has a major problem of constant failure/hazard rate property, which makes it inappropriate for modelling data-sets from various complex life phenomena that may exhibit increasing, decreasing or bathtub hazard rate characteristics. El-Alosey (2007) proposed a two parameter Erlang-truncated exponential distribution. The probability density function (pdf) is given by

$$f(x;\beta,\lambda) = \beta(1-e^{-\lambda})e^{-\beta x(1-e^{-\lambda})}, \quad x \geq 0, \beta, \lambda > 0, \qquad (3)$$

and the corresponding cumulative density function (cdf) is

$$F(x;\beta,\lambda) = 1-e^{-\beta x(1-e^{-\lambda})}, \qquad x \geq 0, \beta, \lambda > 0. \qquad (4)$$

Now in the view of (3) and (4), we have

$$f(x) = \beta(1-e^{-\lambda})[1-F(x)]. \qquad (5)$$

where β is the shape parameter and λ is the scale parameter. It is important to note that the Erlang-truncated exponential distribution has a constant hazard rate function. The standard exponential distribution is the special case of Erlang-truncated

distribution for $\beta(1-e^{-\lambda})=1$. We shall use the (5) in the following sections to derive some recurrence relations for single and product moments of poressively Type-II right censored order statistics arising from Erlang-truncated exponential distributions.

The inability of the Erlang-truncated exponential distribution to adequately model a variety of complex real data-sets, particularly lifetime ones, has stirred huge concern amongst distribution users and researchers alike and has summoned enormous research attention over the last two decades. Khan et al. (2010) obtained the recurrence relations for single and product moments of generalized order statistics of this distribution. Kulshrestha et al. (2013) obtained the marginal and joint moment generating functions of generalized order statistics and Kumar (2014) obtained the explicit expression for generalized order statistics.

Let $X_1, X_2, \cdots, X_n$ be a random sample from the Erlang-truncated exponential distribution with pdf and cdf given in (3) and (4) respectively. The corresponding progressive Type-II right censored order statistics with censoring scheme $(R_1, R_2, \cdots R_m)$, $m \leq n$ will be

$$X_{1:m:n}^{(R_1,R_2,\cdots,R_m)}, X_{2:m:n}^{(R_1,R_2,\cdots,R_m)}, \cdots, X_{m:m:n}^{(R_1,R_2,\cdots,R_m)}.$$

The single moments of the progressive Type-II right censored order statistics from the Erlang-truncated exponential distribution can be written as follows:

$$\begin{aligned}
\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}} &= E\left[x_{i:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}}\right] \\
&= A(n,m-1)\int\int\cdots\int_{0<x_1<x_2<\cdots<x_m<\infty} x_i^k f(x_1) \\
&\times [1-F(x_1)]^{R_1} f(x_2)[1-F(x_2)]^{R_2} f(x_3)[1-F(x_3)]^{R_3}\cdots f(x_m) \\
&\times [1-F(x_m)]^{R_m} dx_1 dx_2 dx_3 \cdots dx_m, \qquad (6)
\end{aligned}$$

where $f(.)$ and $F(.)$ are given respectively in (3), (4), and $A(n,m-1)$ as defined in (2). When $k=1$, the superscript in the notation of the mean of the progressive Type-II right censored order statistics may be omitted without any confusion.

The outline of this note is as follows. Recurrence relations for single moments of progressive Type-II right censored order statistics from Erlang-truncated exponential distribution are given in section 2. Section 3 describes the recurrence relations for product moments of progressive Type-II right censored order statistics from Erlang-truncated exponential distribution. The recurrence algorithm is carried out in section 4 for Erlang-truncated exponential distribution. Further computations of means and variances from Erlang-truncated exponential progressive Type-II right

censored order statistics for some sample sizes n and some censoring schemes (R_1, R_2,..., R_m), $m < n$ are tabulated in section 5.

2. Single moments of progressivley Type-II censored order statistics

In this section, we establish several new recurrence relations satisfied by the single moments of progressive Type-II right censored order statistics from the Erlang-truncated exponential distribution. These recurrence relations may be used to compute the means, variances and covariances of Erlang-truncated exponential progressive Type-II right censored order statistics for all sample sizes n and all censoring schemes $(R_1, R_2, \cdots, R_m),\ m \leq n$.

Theorem 2.1: For $2 \leq m \leq n$ and $k \geq 0$,

$$\begin{aligned} \mu_{1:m:n}^{(R_1,R_2,\cdots,R_m)^{(k+1)}} &= \frac{k+1}{(1+R_1)\beta(1-e^{-\lambda})}\mu_{1:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}} \\ &- \frac{(n-R_1-1)}{(1+R_1)}\mu_{1:m-1:n}^{(R_1+1+R_2,\cdots,R_m)^{(k+1)}}. \end{aligned} \tag{7}$$

Proof: From equations (5) and (6), we have

$$\begin{aligned} \mu_{1:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}} &= A(n,m-1)\int\int\cdots\int_{0<x_1<x_2<\cdots<x_m<\infty} \\ &\times L(x_2)f(x_2)[1-F(x_2)]^{R_2}f(x_3)[1-F(x_3)]^{R_3}\cdots f(x_m) \\ &\times [1-F(x_m)]^{R_m}dx_2dx_3\cdots dx_m, \end{aligned} \tag{8}$$

where

$$L(x_2) = \int_0^{x_2} x_1^k f(x_1)[1-F(x_1)]^{R_1}dx_1. \tag{9}$$

Using (5) in (9), we get

$$\begin{aligned} L(x_2) &= \int_0^{x_2} x_1^k \left\{\beta(1-e^{-\lambda})[1-F(x_1)]\right\}[1-F(x_1)]^{R_1}dx_1 \\ &= \beta(1-e^{-\lambda})\int_0^{x_2} x_1^k[1-F(x_1)]^{R_1+1}dx_1. \end{aligned} \tag{10}$$

Integrating (10) by parts, we get after simplification

$$\begin{aligned} &= \frac{\beta(1-e^{-\lambda})}{k+1}\Big[[1-F(x_2)]^{R_1+1}x_2^{k+1} + (R_1+1)\int_0^{x_2} x_1^{k+1} \\ &\times [1-F(x_1)]^{R_1}f(x_1)dx_1\Big]. \end{aligned} \tag{11}$$

Substituting the value of $L(x_2)$ from (11) in (8) and using (6), we simply have

$$\begin{aligned}\mu_{1:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}} &= \frac{\beta(1-e^{-\lambda})}{k+1}\Big[(n-R_1-1)\mu_{1:m-1:n}^{(R_1+1+R_2,\cdots,R_m)^{(k+1)}} \\ &+ (1+R_1)\mu_{1:m:n}^{(R_1,R_2,\cdots,R_m)^{(k+1)}}\Big],\end{aligned}$$

rearranging the above equation gives the result in (7).

Theorem 2.2: For $m=1, n=1,2,\cdots$ *and* $k\geq 0$,

$$\mu_{1:1:n}^{(n-1)^{(k+1)}} = \frac{k+1}{n\beta(1-e^{-\lambda})}\mu_{1:1:n}^{(n-1)^{(k)}}. \tag{12}$$

Proof: Theorem 2.2 may be proved by following exactly the same steps as those used in proving Theorem 2.1, which is presented above.

Remark 1. We may use the fact that the first progressive Type-II right censored order statistics is the same as the first usual order statistic from a sample of size n, regardless of the censoring scheme employed.

Theorem 2.3: For $2\leq i\leq m-1$, $m\leq n$ and $k\geq 0$,

$$\begin{aligned}\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)^{(k+1)}} &= \frac{1}{1+R_i}\Big[\frac{k+1}{\beta(1-e^{-\lambda})}\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}} \\ &- (n-R_1-R_2-\cdots-R_i-i) \\ &\times \mu_{i:m-1:n}^{(R_1,R_2,\cdots,R_{i-1},R_i+R_{i+1}+1,R_{i+2},\cdots,R_m)^{(k+1)}} \\ &+ (n-R_1-R_2-\cdots-R_{i-1}-i+1) \\ &\times \mu_{i-1:m-1:n}^{(R_1,R_2,\cdots,R_{i-2},R_{i-1}+R_i+1,R_{i+1},\cdots,R_m)^{(k+1)}}\Big].\end{aligned} \tag{13}$$

Proof: Theorem 2.3 may be proved by following exactly the same steps as those used in proving Theorem 2.1, which is presented above.

Theorem 2.4: For $2\leq m\leq n$, and $k\geq 0$,

$$\begin{aligned}\mu_{m:m:n}^{(R_1,R_2,\cdots,R_m)^{(k+1)}} &= \frac{k+1}{(1+R_m)\beta(1-e^{-\lambda})}\mu_{m:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}} \\ &+ \mu_{m-1:m-1:n}^{(R_1,R_2,\cdots,R_{m-2},R_{m-1}+R_m+1)^{(k+1)}}.\end{aligned} \tag{14}$$

Proof: Theorem 2.4 may be proved by following exactly the same steps as those used in proving Theorem 2.1, which is presented above.

Remark 2. Using these recurrence relations, we can obtain all the single moments of all progressive Type-II right censored order statistics for all sample sizes and

censoring schemes $(R_1, R_2, \cdots, R_m)$ in a sample recursive manner. The recursive algorithm will be described in detail in section 4.

Corollary 2.1: For $\beta(1-e^{-\lambda})=1$ in (7), we get the recurrence relation for single moments of progressively Type-II consored order statistics from the standard exponential distribution.

$$\begin{aligned}\mu_{1:m:n}^{(R_1,R_2,\cdots,R_m)^{(k+1)}} &= \frac{1}{(1+R_1)}\Big[(k+1)\mu_{1:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}} \\ &- (n-R_1-1)\mu_{1:m-1:n}^{(R_1+1+R_2,\cdots,R_m)^{(k+1)}}\Big], \end{aligned} \tag{15}$$

as obtained by Aggarwala and Balakrishnan [2].

Corollary 2.2: For $\beta(1-e^{-\lambda})=1$ in (12), we get

$$\mu_{1:1:n}^{(n-1)^{(k+1)}} = \frac{k+1}{(n)}\mu_{1:1:n}^{(n-1)^{(k)}}, \tag{16}$$

as obtained by Aggarwala and Balakrishnan [2].

Corollary 2.3: For $\beta(1-e^{-\lambda})=1$ in (13), we get

$$\begin{aligned}\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)^{(k+1)}} &= \frac{1}{1+R_i}\Big[(k+1)\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}} \\ &- (n-R_1-R_2-\cdots-R_i-i) \\ &\times \mu_{i:m-1:n}^{(R_1,R_2,\cdots,R_{i-1},R_i+R_{i+1}+1,R_{i+2},\cdots,R_m)^{(k+1)}} \\ &+ (n-R_1-R_2-\cdots-R_{i-1}-i+1) \\ &\times \mu_{i-1:m-1:n}^{(R_1,R_2,\cdots,R_{i-2},R_{i-1}+R_i+1,R_{i+1},\cdots,R_m)^{(k+1)}}\Big], \end{aligned} \tag{17}$$

as obtained by Aggarwala and Balakrishnan [2].

Corollary 2.4: For $\beta(1-e^{-\lambda})=1$ in (14), we get

$$\begin{aligned}\mu_{m:m:n}^{(R_1,R_2,\cdots,R_m)^{(k+1)}} &= \frac{k+1}{1+R_m}\mu_{m:m:n}^{(R_1,R_2,\cdots,R_m)^{(k)}} \\ &+ \mu_{m-1:m-1:n}^{(R_1,R_2,\cdots,R_{m-2},R_{m-1}+R_m+1)^{(k+1)}}, \end{aligned} \tag{18}$$

as obtained by Aggarwala and Balakrishnan [2].

Deductions: For the special case $R_1 = R_2 = \cdots = R_m = 0$ so that $m = n$ in which the progressive censored order statistics become the usual order statistics $X_{1:n}, X_{2:n}, \cdots, X_{n:n}$, then

(i) From Eq. (7): For $k \geq 0$,we get

$$\begin{aligned}\mu_{1:n}^{(k+1)} &= \frac{k+1}{\beta(1-e^{-\lambda})}\mu_{1:n}^{k} \\ &- (n-1)\mu_{1:n-1:n}^{(1,0,0,\cdots,0)^{(k+1)}} \end{aligned} \tag{19}$$

(ii) From Eq. (13): For $k \geq 0$,we get

$$\begin{aligned}\mu_{i:n}^{(k+1)} &= \frac{k+1}{\beta(1-e^{-\lambda})}\mu_{i:n}^{(k)} \\ &- (n-i)\mu_{i:n}^{(k+1)} + (n-i+1)\mu_{i-1:n}^{(k+1)} \end{aligned} \tag{20}$$

3. Product moments of progressivley Type-II censored order statistics

In this section, we establish some recurrence relations for product moments of the progressive Type-II right censored order statistics from the Erlang-truncated exponential distribution. The $(r,s)^{th}$ product moment of the progressive type-II right censored order statistics can be written as

$$\begin{aligned}\mu_{r,s:m:n}^{(R_1,R_2,\cdots,R_m)} &= E\left[x_{r:m:n}^{(R_1,R_2,\cdots,R_m)} x_{s:m:n}^{(R_1,R_2,\cdots,R_m)}\right] \\ &= A(n,m-1)\int\int\ldots\int_{0<x_1<x_2<\cdots<x_m<\infty} x_r \\ &\times x_s f(x_1)[1-F(x_1)]^{R_1} f(x_2) \\ &\times [1-F(x_2)]^{R_2}\cdots f(x_m)[1-F(x_m)]^{R_m} dx_1 dx_2 dx_3\cdots dx_m, \end{aligned} \tag{21}$$

where $f(.)$ and $F(.)$ are given respectively in (3) and (4) and $A(n,m-1)$ is defined in (2).

Theorem 3.1: For $1 \leq i < j \leq m-1$ and $m \leq n$,

$$\begin{aligned}\mu_{i,j:m:n}^{(R_1,R_2,\cdots,R_m)} &= \frac{1}{R_j+1}\Big[\frac{1}{\beta(1-e^{-\lambda})}\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)} \\ &- (n-R_1-1-\cdots-R_j-j)\mu_{i,j:m-1:n}^{(R_1,R_2,\cdots,R_{j-1},R_j+R_{j+1}+1,\cdots R_m)} \\ &+ (n-R_1-1-\cdots-R_{j-1}-j+1) \\ &\times \mu_{i,j-1:m-1:n}^{(R_1,R_2,\cdots,R_{j-1}+R_j+1,\cdots,R_m)}\Big]. \end{aligned} \tag{22}$$

Proof: Using (5) and (6), we have

$$\begin{aligned}\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)} &= A(n,m-1)\int\int\cdots\int_{0<x_1<\cdots<x_{j-1}<x_{j+1}<\cdots<x_m<\infty} \\ &\times \left\{\int_{x_{j-1}}^{x_{j+1}} \beta(1-e^{-\lambda})[1-F(x_j)]^{R_j+1}dx_j\right\} \\ &\times x_i f(x_1)[1-F(x_1)]^{R_1}\cdots f(x_{j-1}) \\ &\times [1-F(x_{j-1})]^{R_{j-1}} f(x_{j+1})[1-F(x_{j+1})]^{R_{j+1}}\cdots f(x_m) \\ &\times [1-F(x_m)]^{R_m} dx_1 dx_2\cdots dx_{j-1}dx_{j+1}\cdots dx_m. \qquad (23)\end{aligned}$$

Integrating the innermost integral by parts, we obtain

$$\begin{aligned}\beta(1-e^{-\lambda})\int_{x_{j-1}}^{x_{j+1}}[1-F(x_j)]^{R_j+1}dx_j &= \beta(1-e^{-\lambda})\Big[x_{j+1}[1-F(x_{j+1})]^{1+R_j} \\ &- x_{j-1}[1-F(x_{j-1})]^{1+R_j} + (1+R_j) \\ &\times \int_{x_{j-1}}^{x_{j+1}}[1-F(x_j)]^{R_j} f(x_j)x_j dx_j\Big],\end{aligned}$$

which, when substituted into equation (23) and using (21), we have

$$\begin{aligned}\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)} &= \beta(1-e^{-\lambda})\Big[(n-R_1-1-\cdots-R_j-j) \\ &\times \mu_{i,j:m-1:n}^{(R_1,R_2,\cdots,R_{j-1},R_j+R_{j+1}+1,\cdots R_m)} - (n-R_1-1-\cdots-R_{j-1}-j+1) \\ &\times \mu_{i,j-1:m-1:n}^{(R_1,R_2,\cdots,R_{j-1}+R_j+1,\cdots,R_m)} + (R_j+1)\mu_{i,j:m:n}^{(R_1,R_2,\cdots,R_m)}\Big].\end{aligned}$$

Upon rearrangement of this equation, we obtain the relation in (22).

Theorem 3.2: For $1 \le i \le m-1$ and $m \le n$,

$$\begin{aligned}\mu_{i,m:m:n}^{(R_1,R_2,\cdots,R_m)} &= \frac{1}{R_m+1}\Big[\frac{1}{\beta(1-e^{-\lambda})}\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)} \\ &+ (n-R_1-1-\cdots-R_{m-1}-m+1) \\ &\times \mu_{i,m-1:m-1:n}^{(R_1,R_2,\cdots,R_{m-1}+R_m+1,\cdots,R_m)}\Big]. \qquad (24)\end{aligned}$$

Proof: The theorem 3.2 may be proved by following exactly the same steps as those used earlier in proving Theorem 3.1.

Remark 3. Using these recurrence relations, we can obtain all the product moments of progressive Type-II right censored order statistics for all sample sizes and censoring schemes $(R_1,R_1,\cdots,R_m)$. The detailed recursive algorithm will be described in section 4.

Corollary 3.1: For $\beta(1-e^{-\lambda})=1$ in (22), we get the recurrence relation for product moments of progressively Type-II consored order statistics from the standard exponential distribution.

$$\begin{aligned}\mu_{i,j:m:n}^{(R_1,R_2,\cdots,R_m)} &= \frac{1}{R_j+1}\Big[\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)} - (n-R_1-1-\cdots-R_j-j)\\ &\times \mu_{i,j:m-1:n}^{(R_1,R_2,\cdots,R_{j-1},R_j+R_{j+1}+1,\cdots R_m)}\\ &+ (n-R_1-1-\cdots-R_{j-1}-j+1)\\ &\times \mu_{i,j-1:m-1:n}^{(R_1,R_2,\cdots,R_{j-1}+R_j+1,\cdots,R_m)}\Big],\end{aligned}$$

as obtained by Aggarwala and Balakrishnan [2].

Corollary 3.2: For $\beta(1-e^{-\lambda})=1$ in (24), we get

$$\begin{aligned}\mu_{i,m:m:n}^{(R_1,R_2,\cdots,R_m)} &= \frac{1}{R_m+1}\Big[\mu_{i:m:n}^{(R_1,R_2,\cdots,R_m)}\\ &+ (n-R_1-1-\cdots-R_{m-1}-m+1)\\ &\times \mu_{i,m-1:m-1:n}^{(R_1,R_2,\cdots,R_{m-1}+R_m+1,\cdots,R_m)}\Big],\end{aligned}$$

as obtained by Aggarwala and Balakrishnan [2].

4. Illustration of the recursive computational algorithm

In this section, we describe the recursive computational algorithm that will produce all the means, variances and covariances of all progressively Type-II right censored order statistics for all sample sizes n and all choices of m and $(R_1,R_2,\cdots R_m)$ from the Erlang-truncated exponential distribution.

4.1. Single moments

All the frist and second order moments with $m=1$ for all sample sizes n can be obtained by setting $k=0$ in equation (12) and then again setting $k=1$ in the same equation. Next using equation (7), we can determine all the moments of the form $\mu_{1,2:n}^{(R_1,R_2)}, n=2,3\cdots$,which can in turn be used again with (7), to determine all moments of the form $\mu_{1,2:n}^{(R_1,R_2)^2}, n=2,3\cdots$. (14) can then be used to obtain $\mu_{2,2:n}^{(R_1,R_2)}$ for all R_1,R_2 and $n\geq 2$ and these values can be used to obtain all moments of the form $\mu_{1,2:n}^{(R_1,R_2)^2}$ by using equation (14) again. (7) can now be used again to obtain $\mu_{1,3:n}^{(R_1,R_2,R_3)}$, $\mu_{1,3:n}^{(R_1,R_2,R_3)^2}$ for all n, R_1,R_2 and R_3 and equation (13)can be used next to

obtain all moments of the form $\mu_{2,3:n}^{(R_1,R_2,R_3)}$, $\mu_{2,3:n}^{(R_1,R_2,R_3)^2}$. Finally, equation (14) can be used to obtain all moments of the form $\mu_{3,3:n}^{(R_1,R_2,R_3)}$, $\mu_{3,3:n}^{(R_1,R_2,R_3)^2}$.
This process can be continued until all desired first and second order moments and hence all variances are obtained.

4.2. Product moments

From (24), all moments of the form form $\mu_{m-1,m:m:n}^{(R_1,R_2,\cdots,R_m)}$, $m = 2,3,\cdots,n$, can be determined, since only single moments, which have already been obtained, are needed to calculate them. Then, using (22), all moments of the form $\mu_{i-1,i:m:n}^{(R_1,R_2,\cdots,R_m)}$, $2 \leq i < m$, can be obtained. From this point, using (24), we can obtain all moments of the form $\mu_{m-2,m:m:n}^{(R_1,R_2,\cdots,R_m)}$, $m = 3,4,\cdots,n$, and subsequently, using (22), all the moments of the form $\mu_{i-2,i:m:n}^{(R_1,R_2,\cdots,R_m)}$, $3 \leq i < m$. Continuing this way, all the desired product moments and hence all covariances can be obtained.

5. Numerical results

The recurrence relations obtained in the preceding sections allow us to evaluate the means,,variances and covairances of Erlang-truncated exponential progressive Type-II right censored order statistics for all sample sizes n and all censoring schemes $(R_1, R_2,..., R_m)$, $m < n$. These quantities can be used for various inferential purposes; for example, they are useful in determining BLUEs of location/scale parameters and BLUPs of censored failure times. In this section we compute the means and variances of Erlang-truncated exponential progressive Type-II right censored order statistics for sample sizes up to 20 and for different choices of m and progressive schemes $(R_1, R_2,..., R_m)$, $m < n$.
In Table 1-2, we have computed the values of means of the progressive Type-II right censored order statistics for $\lambda = 2,3$, $\beta = 3,5$ and different values of m and n. One can see that the means are increasing with respect to m and n but decreasing with respect to β and λ. In Table 3-4, we have computed the variances of the progressive Type-II right censored order statistics for $\lambda = 2,3$, $\beta = 3,5$ and different values of m and n. We can see that variances are decreasing with respect to m, n and β, λ.
Tables for the skewness, kurtosis, product moments and covariances of the progressive Type-II right censored order statistics are not presented here but are available from the author on request. All computations here were performed using Mathematica. Mathematica like other algebraic manipulation packages allows for arbitrary precision, so the accuracy of the given values is not an issue.

Table 1. Means of progressively Type-II right censored order statistics for $\lambda = 2$ $\beta = 3$.

$m\downarrow$	$n\downarrow$	Scheme	Mean				
2	5	(0,3)	0.077101	0.173477			
2	5	(3,0)	0.077101	0.462607			
2	8	(6,0)	0.048188	0.433694			
2	8	(0,6)	0.048188	0.103260			
2	10	(8,0)	0.038550	0.424056			
2	10	(0,8)	0.038550	0.081384			
2	12	(10,0)	0.032125	0.417631			
2	12	(0,10)	0.032125	0.067171			
2	15	(13,0)	0.025700	0.411206			
2	15	(0,13)	0.025700	0.053236			
2	18	(16,0)	0.021416	0.406922			
2	18	(0,16)	0.021416	0.044093			
2	20	(18,0)	0.019275	0.404781			
2	20	(0,18)	0.019275	0.039565			
3	5	(2,0,0)	0.077101	0.269854	0.655359		
3	5	(0,0,2)	0.077101	0.173477	0.301979		
3	8	(5,0,0)	0.048188	0.240941	0.626447		
3	8	(0,0,5)	0.048188	0.103260	0.167511		
3	10	(7,0,0)	0.038550	0.231303	0.616809		
3	10	(0,0,7)	0.038550	0.081384	0.129572		
3	12	(9,0,0)	0.032125	0.224878	0.610384		
3	12	(0,0,9)	0.032125	0.067171	0.105722		
3	15	(12,0,0)	0.025700	0.218453	0.603959		
3	15	(0,0,12)	0.025700	0.053236	0.082890		
3	18	(15,0,0)	0.021416	0.214169	0.599675		
3	18	(0,0,15)	0.021416	0.044093	0.068187		
3	20	(17,0,0)	0.019275	0.212028	0.597534		
3	20	(0,0,17)	0.019275	0.039565	0.060982		
4	5	(1,0,0,0)	0.077101	0.205603	0.398356	0.783861	
4	5	(0,0,0,1)	0.077101	0.173477	0.301979	0.494732	
4	8	(4,0,0,0)	0.048188	0.176690	0.369443	0.754949	
4	8	(0,0,0,4)	0.048188	0.103260	0.167511	0.244612	
4	10	(6,0,0,0)	0.038550	0.167052	0.359805	0.745311	
4	10	(0,0,0,6)	0.038550	0.081384	0.129572	0.184645	
4	12	(8,0,0,0)	0.032125	0.160627	0.353380	0.738886	
4	12	(0,0,0,8)	0.032125	0.067171	0.105722	0.148556	
4	15	(11,0,0,0)	0.025700	0.154202	0.346955	0.732461	
4	15	(0,0,0,11)	0.025700	0.053236	0.082890	0.115016	
4	18	(14,0,0,0)	0.021416	0.149918	0.342671	0.728177	
4	18	(0,0,0,14)	0.021416	0.044093	0.068187	0.093888	
4	20	(16,0,0,0)	0.019275	0.147777	0.340530	0.726036	
4	20	(0,0,0,16)	0.019275	0.039565	0.060982	0.083658	
5	5	(0,0,0,0,0)	0.077101	0.173477	0.301979	0.494732	0.880238
5	8	(3,0,0,0,0)	0.048188	0.144564	0.273066	0.465819	0.851325
5	8	(0,0,0,0,3)	0.048188	0.103260	0.167511	0.244612	0.340989
5	10	(5,0,0,0,0)	0.038550	0.134927	0.263429	0.456181	0.841687
5	10	(0,0,0,0,5)	0.038550	0.081384	0.129572	0.184645	0.248896
5	12	(7,0,0,0,0)	0.032125	0.128501	0.257003	0.449756	0.835262
5	12	(0,0,0,0,7)	0.032125	0.067171	0.105722	0.148556	0.196744
5	15	(10,0,0,0,0)	0.025700	0.122076	0.250578	0.443331	0.828837
5	15	(0,0,0,0,10)	0.025700	0.053236	0.082890	0.115016	0.150062
5	18	(13,0,0,0,0)	0.021416	0.117793	0.246295	0.439048	0.824554
5	18	(0,0,0,0,13)	0.021416	0.044093	0.068187	0.093888	0.121424
5	20	(15,0,0,0,0)	0.019275	0.115651	0.244153	0.436906	0.822412
5	20	(0,0,0,0,15)	0.019275	0.039565	0.060982	0.083658	0.107753

Table 2. Means of progressively Type-II right censored order statistics for $\lambda = 3$ $\beta = 5$.

$m\downarrow$	$n\downarrow$	Scheme	Mean				
5	5(0,3)	(0,3)	0.042095	0.094715			
5	5(3,0)	(3,0)	0.042095	0.252574			
8	8(6,0)	(6,0)	0.026309	0.236789			
8	8(0,6)	(0,6)	0.026309	0.056378			
10	10(8,0)	(8,0)	0.021047	0.231527			
10	10(0,8)	(0,8)	0.021047	0.044434			
12	12(10,0)	(10,0)	0.017539	0.228019			
12	12(0,10)	(0,10)	0.017539	0.036674			
15	15(13,0)	(13,0)	0.014031	0.224511			
15	15(0,13)	(0,13)	0.014031	0.029066			
18	18(16,0)	(16,0)	0.011693	0.222172			
18	18(0,16)	(0,16)	0.011693	0.024074			
20	20(18,0)	(18,0)	0.010523	0.221003			
20	20(0,18)	(0,18)	0.010523	0.021601			
5	5(2,0,0)	(2,0,0)	0.042095	0.147335	0.357814		
5	5(0,0,2)	(0,0,2)	0.042095	0.094715	0.164875		
8	8(5,0,0)	(5,0,0)	0.026309	0.131549	0.342028		
8	8(0,0,5)	(0,0,5)	0.026309	0.056378	0.091458		
10	10(7,0,0)	(7,0,0)	0.021047	0.126287	0.336766		
10	10(0,0,7)	(0,0,7)	0.021047	0.044434	0.070744		
12	12(9,0,0)	(9,0,0)	0.017539	0.122779	0.333258		
12	12(0,0,9)	(0,0,9)	0.017539	0.036674	0.057722		
15	15(12,0,0)	(12,0,0)	0.014031	0.119271	0.329750		
15	15(0,0,12)	(0,0,12)	0.014031	0.029066	0.045256		
18	18(15,0,0)	(15,0,0)	0.011693	0.116932	0.327411		
18	18(0,0,15)	(0,0,15)	0.011693	0.024074	0.037229		
20	20(17,0,0)	(17,0,0)	0.010523	0.115763	0.326242		
20	20(0,0,17)	(0,0,17)	0.010523	0.021601	0.033295		
5	5(1,0,0,0)	(1,0,0,0)	0.042095	0.112255	0.217495	0.427974	
5	5(0,0,0,1)	(0,0,0,1)	0.042095	0.094715	0.164875	0.270114	
8	8(4,0,0,0)	(4,0,0,0)	0.026309	0.096469	0.201709	0.412188	
8	8(0,0,0,4)	(0,0,0,4)	0.026309	0.056378	0.091458	0.133554	
10	10(6,0,0,0)	(6,0,0,0)	0.021047	0.091207	0.196447	0.406926	
10	10(0,0,0,6)	(0,0,0,6)	0.021047	0.044434	0.070744	0.100812	
12	12(8,0,0,0)	(8,0,0,0)	0.017539	0.087699	0.192939	0.403418	
12	12(0,0,0,8)	(0,0,0,8)	0.017539	0.036674	0.057722	0.081108	
15	15(11,0,0,0)	(11,0,0,0)	0.014031	0.084191	0.189431	0.399910	
15	15(0,0,0,11)	(0,0,0,11)	0.014031	0.029066	0.045256	0.062796	
18	18(14,0,0,0)	(14,0,0,0)	0.011693	0.081852	0.187092	0.397571	
18	18(0,0,0,14)	(0,0,0,14)	0.011693	0.024074	0.037229	0.051261	
20	20(16,0,0,0)	(16,0,0,0)	0.010523	0.080683	0.185923	0.396402	
20	20(0,0,0,16)	(0,0,0,16)	0.010523	0.021601	0.033295	0.045676	
5	5(0,0,0,0,0)	(0,0,0,0,0)	0.042095	0.094715	0.164875	0.270114	0.480594
8	8(3,0,0,0,0)	(3,0,0,0,0)	0.026309	0.078929	0.149089	0.254328	0.464808
8	8(0,0,0,0,3)	(0,0,0,0,3)	0.026309	0.056378	0.091458	0.133554	0.186173
10	10(5,0,0,0,0)	(5,0,0,0,0)	0.021047	0.073667	0.143827	0.249066	0.459546
10	10(0,0,0,0,5)	(0,0,0,0,5)	0.021047	0.044434	0.070744	0.100812	0.135892
12	12(7,0,0,0,0)	(7,0,0,0,0)	0.017539	0.070159	0.140319	0.245558	0.456038
12	12(0,0,0,0,7)	(0,0,0,0,7)	0.017539	0.036674	0.057722	0.081108	0.107418
15	15(10,0,0,0,0)	(10,0,0,0,0)	0.014031	0.066651	0.136811	0.242051	0.452530
15	15(0,0,0,0,10)	(0,0,0,0,10)	0.014031	0.029066	0.045256	0.062796	0.081931
18	18(13,0,0,0,0)	(13,0,0,0,0)	0.011693	0.064313	0.134472	0.239712	0.450191
18	18(0,0,0,0,13)	(0,0,0,0,13)	0.011693	0.024074	0.037229	0.051261	0.066295
20	20(15,0,0,0,0)	(15,0,0,0,0)	0.010523	0.063143	0.133303	0.238543	0.449022
20	20(0,0,0,0,15)	(0,0,0,0,15)	0.010523	0.021601	0.033295	0.045676	0.058831

Table 3. Variances of progressively Type-II right censored order statistics for $\lambda = 2$ $\beta = 3$.

$m\downarrow$	$n\downarrow$	Scheme	Variance				
2	5	(0,3)	0.005944	0.015233			
2	5	(3,0)	0.005944	0.154559			
2	8	(6,0)	0.002322	0.150936			
2	8	(0,6)	0.002322	0.005355			
2	10	(8,0)	0.001486	0.150100			
2	10	(0,8)	0.001486	0.003320			
2	12	(10,0)	0.001032	0.149646			
2	12	(0,10)	0.001032	0.002260			
2	15	(13,0)	0.000660	0.149275			
2	15	(0,13)	0.000660	0.001418			
2	18	(16,0)	0.000458	0.149073			
2	18	(0,16)	0.000458	0.000972			
2	20	(18,0)	0.000371	0.148986			
2	20	(0,18)	0.000371	0.000783			
3	5	(2,0,0)	0.005944	0.043098	0.191713		
3	5	(0,0,2)	0.005944	0.015233	0.031745		
3	8	(5,0,0)	0.002322	0.039475	0.188090		
3	8	(0,0,5)	0.002322	0.005355	0.009483		
3	10	(7,0,0)	0.001486	0.038639	0.187254		
3	10	(0,0,7)	0.001486	0.003320	0.005643		
3	12	(9,0,0)	0.001032	0.038185	0.186800		
3	12	(0,0,9)	0.001032	0.002260	0.003746		
3	15	(12,0,0)	0.000660	0.037814	0.186428		
3	15	(0,0,12)	0.000660	0.001418	0.002298		
3	18	(15,0,0)	0.000458	0.037612	0.186227		
3	18	(0,0,15)	0.000458	0.000972	0.001553		
3	20	(17,0,0)	0.000371	0.037525	0.186140		
3	20	(0,0,17)	0.000371	0.000783	0.001241		
4	5	(1,0,0,0)	0.005944	0.022457	0.059611	0.208225	
4	5	(0,0,0,1)	0.005944	0.015233	0.031745	0.068899	
4	8	(4,0,0,0)	0.002322	0.018834	0.055988	0.204603	
4	8	(0,0,0,4)	0.002322	0.005355	0.009483	0.015427	
4	10	(6,0,0,0)	0.001486	0.017998	0.055152	0.203767	
4	10	(0,0,0,6)	0.001486	0.003320	0.005643	0.008675	
4	12	(8,0,0,0)	0.001032	0.017544	0.054698	0.203313	
4	12	(0,0,0,8)	0.001032	0.002260	0.003746	0.005581	
4	15	(11,0,0,0)	0.000660	0.017173	0.054326	0.202941	
4	15	(0,0,0,11)	0.000660	0.001418	0.002298	0.003330	
4	18	(14,0,0,0)	0.000458	0.016971	0.054125	0.202739	
4	18	(0,0,0,14)	0.000458	0.000972	0.001553	0.002213	
4	20	(16,0,0,0)	0.000371	0.016884	0.054037	0.202652	
4	20	(0,0,0,16)	0.000371	0.000783	0.001241	0.001756	
5	5	(0,0,0,0,0)	0.005944	0.015233	0.031745	0.068899	0.217514
5	8	(3,0,0,0,0)	0.002322	0.011610	0.028123	0.065276	0.213891
5	8	(0,0,0,0,3)	0.002322	0.005355	0.009483	0.015427	0.024716
5	10	(5,0,0,0,0)	0.001486	0.010774	0.027287	0.064441	0.213055
5	10	(0,0,0,0,5)	0.001486	0.003320	0.005643	0.008675	0.012804
5	12	(7,0,0,0,0)	0.001032	0.010320	0.026833	0.063986	0.212601
5	12	(0,0,0,0,7)	0.001032	0.002260	0.003746	0.005581	0.007903
5	15	(10,0,0,0,0)	0.000660	0.009948	0.026461	0.063615	0.212230
5	15	(0,0,0,0,10)	0.000660	0.001418	0.002298	0.003330	0.004558
5	18	(13,0,0,0,0)	0.000458	0.009747	0.026259	0.063413	0.212028
5	18	(0,0,0,0,13)	0.000458	0.000972	0.001553	0.002213	0.002972
5	20	(15,0,0,0,0)	0.000371	0.009659	0.026172	0.063326	0.211941
5	20	(0,0,0,0,15)	0.000371	0.000783	0.001241	0.001756	0.002336

Table 4. Variances of progressively Type-II right censored order statistics for $\lambda = 3$ $\beta = 5$.

$m\downarrow$	$n\downarrow$	Scheme	Variance				
5	5(0,3)	(0,3)	0.001772	0.004540			
5	5(3,0)	(3,0)	0.001772	0.046073			
8	8(6,0)	(6,0)	0.000692	0.044993			
8	8(0,6)	(0,6)	0.000692	0.001596			
10	10(8,0)	(8,0)	0.000443	0.044744			
10	10(0,8)	(0,8)	0.000443	0.000989			
12	12(10,0)	(10,0)	0.000307	0.044609			
12	12(0,10)	(0,10)	0.000307	0.000673			
15	15(13,0)	(13,0)	0.000196	0.044498			
15	15(0,13)	(0,13)	0.000196	0.000422			
18	18(16,0)	(16,0)	0.000136	0.044438			
18	18(0,16)	(0,16)	0.000136	0.000290			
20	20(18,0)	(18,0)	0.000110	0.044412			
20	20(0,18)	(0,18)	0.000110	0.000233			
5	5(2,0,0)	(2,0,0)	0.001772	0.012847	0.057148		
5	5(0,0,2)	(0,0,2)	0.001772	0.004540	0.009463		
8	8(5,0,0)	(5,0,0)	0.000692	0.011767	0.056069		
8	8(0,0,5)	(0,0,5)	0.000692	0.001596	0.002826		
10	10(7,0,0)	(7,0,0)	0.000443	0.011518	0.055819		
10	10(0,0,7)	(0,0,7)	0.000443	0.000989	0.001682		
12	12(9,0,0)	(9,0,0)	0.000307	0.011383	0.055684		
12	12(0,0,9)	(0,0,9)	0.000307	0.000673	0.001116		
15	15(12,0,0)	(12,0,0)	0.000196	0.011272	0.055573		
15	15(0,0,12)	(0,0,12)	0.000196	0.000422	0.000685		
18	18(15,0,0)	(15,0,0)	0.000136	0.011212	0.055513		
18	18(0,0,15)	(0,0,15)	0.000136	0.000290	0.000463		
20	20(17,0,0)	(17,0,0)	0.000110	0.011186	0.055487		
20	20(0,0,17)	(0,0,17)	0.000110	0.000233	0.000370		
5	5(1,0,0,0)	(1,0,0,0)	0.001772	0.006694	0.017769	0.062071	
5	5(0,0,0,1)	(0,0,0,1)	0.001772	0.004540	0.009463	0.020538	
8	8(4,0,0,0)	(4,0,0,0)	0.000692	0.005614	0.016689	0.060991	
8	8(0,0,0,4)	(0,0,0,4)	0.000692	0.001596	0.002826	0.004598	
10	10(6,0,0,0)	(6,0,0,0)	0.000443	0.005365	0.016440	0.060742	
10	10(0,0,0,6)	(0,0,0,6)	0.000443	0.000989	0.001682	0.002586	
12	12(8,0,0,0)	(8,0,0,0)	0.000307	0.005230	0.016305	0.060606	
12	12(0,0,0,8)	(0,0,0,8)	0.000307	0.000673	0.001116	0.001663	
15	15(11,0,0,0)	(11,0,0,0)	0.000196	0.005119	0.016194	0.060496	
15	15(0,0,0,11)	(0,0,0,11)	0.000196	0.000422	0.000685	0.000992	
18	18(14,0,0,0)	(14,0,0,0)	0.000136	0.005059	0.016134	0.060435	
18	18(0,0,0,14)	(0,0,0,14)	0.000136	0.000290	0.000463	0.000659	
20	20(16,0,0,0)	(16,0,0,0)	0.000110	0.005033	0.016108	0.060409	
20	20(0,0,0,16)	(0,0,0,16)	0.000110	0.000233	0.000370	0.000523	
5	5(0,0,0,0,0)	(0,0,0,0,0)	0.001772	0.004540	0.009463	0.020538	0.064840
8	8(3,0,0,0,0)	(3,0,0,0,0)	0.000692	0.003461	0.008383	0.019458	0.063760
8	8(0,0,0,0,3)	(0,0,0,0,3)	0.000692	0.001596	0.002826	0.004598	0.007367
10	10(5,0,0,0,0)	(5,0,0,0,0)	0.000443	0.003211	0.008134	0.019209	0.063511
10	10(0,0,0,0,5)	(0,0,0,0,5)	0.000443	0.000989	0.001682	0.002586	0.003816
12	12(7,0,0,0,0)	(7,0,0,0,0)	0.000307	0.003076	0.007998	0.019074	0.063375
12	12(0,0,0,0,7)	(0,0,0,0,7)	0.000307	0.000673	0.001116	0.001663	0.002355
15	15(10,0,0,0,0)	(10,0,0,0,0)	0.000196	0.002965	0.007888	0.018963	0.063264
15	15(0,0,0,0,10)	(0,0,0,0,10)	0.000196	0.000422	0.000685	0.000992	0.001358
18	18(13,0,0,0,0)	(13,0,0,0,0)	0.000136	0.002905	0.007827	0.018903	0.063204
18	18(0,0,0,0,13)	(0,0,0,0,13)	0.000136	0.000290	0.000463	0.000659	0.000886
20	20(15,0,0,0,0)	(15,0,0,0,0)	0.000110	0.002879	0.007801	0.018877	0.063178
20	20(0,0,0,0,15)	(0,0,0,0,15)	0.000110	0.000233	0.000370	0.000523	0.000696

6. Concluding remarks

In this paper, we have established several recurrence relations for the single and product moments of progressively Type-II right censored order statistics from the Erlan-truncated exponential distribution. These relations produce in a simple systmatic manner all the means, variances and covariances of progressively Type-II right censored order statistics for all sample sizes and all progressive censoring schemes. we have computed the means and variances of Erlang-truncated exponential progressive Type-II right censored order statistics for sample sizes up to 20 and for different choices of m and progressive schemes $(R_1, R_2,..., R_m)$, $m < n$.

Acknowledgements

The author would like to thank two anonymous referees and the editors for many valuable suggestions which have helped to improve the paper significantly.

REFERENCES

ARNOLD, B. C., BALAKRISHNAN, N., NAGARAJA, H. N., (1992). A First Course in Order Statistics. John Wiley and Sons, New York.

AGGARWALA, R., BALAKRISHNAN, N., (1996). Recurrence relations for single and product moments of progressively Type-II censored order statistics from a exponential and truncated exponential distribution. Ann. Inst. Statist. Math., 48, pp. 757–771.

BALAKRISHNAN, N., MALIK, H. J., (1986). Order statistics from the linear-exponential distribution, part I: Increasing hazard rate case. Comm. Stat. Theory Meth., 15, pp. 179–203.

BALAKRISHNAN, N., CRAMER, H. J., (2014). The art of progressive censoring: Applications to reliability and quality. Springer Science Business Media New York.

BALAKRISHNAN, N., Sandhu, R. A., (1995). A simple simulational algorithm for generating progressive Type-II censored samples. Amer. Statist., 49, pp. 229–230.

BALAKRISHNAN, N., AGGARWALA, R. (1998). Recurrence relations for single and product moments of order statistics from a generalized logistic distribution with applications to inference and generalizations to double truncation. Handbook of Statistics, 17, pp. 85–126.

BALAKRISHNAN, N., SULTAN, K. S., (1998). Recurrence relations and identities for moments of order statistics. In: N. Balakrishnan and C. R. Rao, eds. Handbook of Statistics, 16, pp. 149–228, Order Statistics: Theory and Methods, North-Holland, Amsterdam, The Netherlands.

BALAKRISHNAN, N. AGGARWALA, R., (2000). Progressive Censoring: Theory, Method and Applications, Birkhauser, Bosto.

BALAKRISHNAN, N., KANNAN, N., LIN, C. T., WU, S. J. S., (2004). Inference for the extreme value distribution under progressive Type-II censoring, Journal of Statistical Computation and Simulation, 74, pp. 25–45.

BALAKRISHNAN, N., (2007). Progressive Censoring Methodology. An Appraisal Test. 16, pp. 211–296.

BALAKRISHNAN, N., AL-HUSSAINI, E. K., SALEH, H. M., (2011). Recurrence relations for moments of progressively censored order statistics from logistic distribution with applications to inference. Journal of Statistical Planning and Inference, 14, pp. 17–30.

BALAKRISHNAN, N., SALEH, H. M., (2012). Relations for moments of progressively Type-II censored order statistics from log-logistic distribution with applications to inference. Comm. Stat. Theory Meth., 41, pp. 880–906.

BALAKRISHNAN, N., SALEH, H. M., (2013). Recurrence relations for single and product moments of progressively Type-II censored order statistics from a generalized halflogistic distribution with application to inference., Journal of Statistical Computation and Simulation, 83, 1704–1721.

COHEN, A. C., (1963). Progressively censored samples in life testing, Technometrics, 5, pp. 327–329.

DAVID, H. A., NAGARAJA, H. N., (2003). Order Statistics, Third Edition. John Wiley and Sons, New York.

EL-BASET, A., AHMED, A., FAWZY, A. M., (2003). Recurrence relations for single moments of generalized order statistics from doubly truncated distribution. Journal of Statistical Planning and Inference, 117, pp. 241–249.

El-Alosey, A. R., (2007). Random sum of new type of mixture of distribution. International Journal of Statistics and Systems, 2, pp. 49–57.

FERNANDEZ, A. J., (2004). On estimating exponential parameters with general type II progressive censoring. Journal of Statistical Planning and Inference, 121, pp. 135–147.

JOSHI, P. C., (1978). Recurrence relations between moments of order statistics from exponential and truncated exponential distributions. Sankhya Ser. B, 39, pp. 362–371.

KHAN, R. U., KUMAR, D., HASEEB, A., (2010). Moments of generalized order statistics from Erlang-truncated exponential distribution and its characterization. International Journal of Statistics and Systems, 5, pp. 455–464.

KUMAR, D., (20014). Relations of generalized order statistics from Erlang-Truncated exponential distribution. Pacific Journal of Applied Statistics, 6, pp. 55–77.

KUMAR, D., DEY, S., NADARAJAH, S., (2017). Extended exponential distribution based on order statistics. Communications in Statistics-Theory and Methods, 46, pp. 9166–9184.

KUMAR, D., DEY, S., (2017). Power generalized Weibull distribution based on order statistics. Journal of Statistical Research, 51, pp. 61–78.

KULSHRESTHA, A., KHAN, R. U., KUMAR, D., (2013). On moment generating functions of generalized order statistics from Erlang-truncated exponential distribution. Open Journal of Statistics, 2, pp. 557–564.

MAHMOUD, R. M., SULTAN, K. S., SALEH, H. M., (2006). Progressively censored data from the linear exponential distribution, moments and estimation. METRON, LXIV(2), pp. 99–215.

MANN, N. R., (1971). Best linear invariant estimation for Weibull parameters under progressive censoring, Technometrics, 13, pp. 521–534.

NADARAJAH, S., HAGHIGHI, F., (2011). An extension of the exponential distribution. Statistics, 45, pp. 543–558.

NIKULIN, M., HAGHIGHI, F., (2006). A chi-squared test for the generalized power Weibull family for the head-and-neck cancer censored data. Journal of Mathematics Sciences, 133, pp. 1333–1341.

SALAH, M. M., RAQAB, M. Z., AHSANULLAH, M., (2008). Marshall-Olkin Exponential Distribution: Moments of Order Statistics. Journal of Applied Statistical Science, 17, pp. 81–92.

SULTAN, K. S., MAHMOUD, M. R., SALEH, H. M., (2006). Moments of estimation from progressively censored data of the half logistic distribution. International Journal of Reliability and Applications, 7, pp. 187–201.

THOMAS, D. R., WILSON, W. M., (1972). Linear order statistic estimation for the two-parameter Weibull and extreme-value distributions from Type-II progressively censored samples. Technometrics, 14, pp. 679–691.

VIVEROS, R., BALAKRISHNAN, N., (1994). Interval estimation of life characteristics from progressively censored data. Technometrics, 36, pp. 84–91.

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 669–686, DOI 10.21307/stattrans-2017-006

A GENERALIZED RANDOMIZED RESPONSE MODEL

Housila P. Singh, Swarangi M. Gorey[1]

ABSTRACT

In this paper we have suggested a generalized version of the Gjestvang and Singh (2006) model and have studied its properties. We have shown that the randomized response models due to Warner (1965), Mangat and Singh (1990), Mangat (1994) and Gjestvang and Singh (2006) are members of the proposed RR model. The conditions are obtained in which the suggested RR model is more efficient than the Warner (1965) model, Mangat and Singh (1990) model and Mangat (1994) model and Gjestvang and Singh (2006) model. A numerical illustration is given in support of the present study.

Key words: sensitive variable, population proportion, Gjestvang and Singh's model, variance, efficiency.
AMS Subject Classification: 62D05.

1. Introduction

The collection of data through personal interviews surveys on sensitive issues such as induced abortions, alcohol and drug abuse (Weissman et al., 1986, Fisher et al., 1992) as well as on attitudes (Antonak and Livnech, 1995), on sexual behaviour (Williams and Suen, 1994, Jarman, 1997) and family income is a serious issue. Warner (1965) introduced an ingenious technique known as the randomized response technique for estimating the proportion π of people bearing a sensitive attribute, say A, in a given community from which a sample is collected. For estimating π, a simple random sample of n respondents is selected with replacement from the population. For collecting information on the sensitive characteristic, Warner (1965) made use of randomization device. The randomization device consists of a deck of cards with each card having one of the following two statements:

(i) I belong to sensitive group A;

(ii) I do not belong to sensitive group A,

[1] School of Studies in Statistics, Vikram University, Ujjain-456010, M.P., India.

represented with probabilities p_0 and $(1-p_0)$ respectively in the deck of cards. Each respondent in the sample is asked to select a card at random from the well-shuffled deck. Without showing the card to the interviewer, the interviewee answers the question, "Is the statement true for you?" the number of respondents n_1 that answer "yes" is binomially distributed with parameters $p_0\pi+(1-p_0)(1-\pi)$. The maximum likelihood estimator π exists for $p_0 \neq \frac{1}{2}$ and is given by

$$\hat{\pi}_w = \frac{(n_1/n)-(1-p_0)}{(2p_0-1)} \tag{1.1}$$

which is unbiased and has the variance

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p_0(1-p_0)}{n(2p_0-1)^2}. \tag{1.2}$$

Mangat and Singh (1990) envisaged a two-stage randomized response model. In the first stage, each respondent was requested to use a randomization device, R_1, such as a deck of cards with each card containing one of the following two statements: (i) "I belong to sensitive group A", (ii) "Go the randomization device R_2". The statements occur with probabilities T_0 and $(1-T_0)$, respectively, in the first device R_1. In the second stage, if directed by the outcome of R_1, the respondent is requested to use the randomization device R_2, which is the same as the Warner (1965) device. Under the two-stage randomized response model, an unbiased estimator of the population proportion π, due to Mangat and Singh (1990) is given by

$$\hat{\pi}_{ms} = \frac{(n_1/n)-(1-T_0)(1-p_0)}{(2p_0-1)+2T_0(1-p_0)} \tag{1.3}$$

with the variance

$$V(\hat{\pi}_{ms}) = \frac{\pi(1-\pi)}{n} + \frac{(1-p_0)(1-T_0)[1-(1-p_0)(1-T_0)]}{n[2p_0-1+2T_0(1-p_0)]^2} \tag{1.4}$$

Mangat (1994) investigated another randomized response model where each respondent selected in the sample was requested to report "yes" if he/she belonged to the sensitive group A; otherwise, he/she was instructed to use the

Warner (1965) device. Under this model, Mangat (1994) obtained an unbiased estimator of the population proportion π given by

$$\hat{\pi}_m = \frac{(n_1/n)-(1-p_0)}{p_0} \tag{1.5}$$

with the variance

$$V(\hat{\pi}_m) = \frac{\theta_m(1-\theta_m)}{np_0^2}, \tag{1.6}$$

where $\theta_m = \pi + (1-\pi)(1-p_0)$.

It is to be mentioned that the Mangat (1994) RR model is more efficient than both the Warner (1965) and Mangat and Singh (1990) models.

A rich growth of literature on randomized response procedure has been accumulated in Chaudhuri and Mukherjee (1987, 1988). Further, a detailed review on randomized response sampling can be found in Singh (2003). Some related work on the randomized response sampling can be also be found in Odumade and Singh (2008, 2009a, 2009b, 2010) Bouza et al. (2010) and Chaudhuri et al. (2016).

It is noted that the Mangat (1994) model has been improved by Gjestvang and Singh (2006). In this paper we have made an effort to suggest a generalized randomized response model which includes Warner (1965), Mangat and Singh (1990), Mangat (1994), Gjestvang and Singh (2006) randomized response model. It has been shown that the proposed model is superior to the models suggested by Warner (1965), Mangat and Singh (1990), Mangat (1994) and Gjestvang and Singh (2006) under some realistic conditions. Numerical illustration is given in support of the present study.

2. Suggested Randomized Response Model

In this section we propose a generalized randomized response model. For estimating π, the proportion of respondents in the population belonging to the sensitive group A, a simple random sample of n respondents is selected with replacement from the population. If the person who is selected in the sample belongs to the sensitive group A, then he or she is requested to use the randomization device R_1 that is described below. Similar to Gjestvang and Singh (2006), let α_1 and β_1 be any two positive real numbers such that $p = \alpha_1/(\alpha_1+\beta_1)$ is the probability in the randomization device R_1 directing the selected respondent to report a scrambled response (or indirect response) as

$(1+w_1\beta_1 S_1)$, and $(1-p)=\beta_1/(\alpha_1+\beta_1)$ is the probability in the randomization device R_1 directing the selected respondent to report a scrambled response as $(1-w_1\alpha_1 S_1)$, where w_1 is a known real number and S_1 is any non-directional scrambling variable, i.e. S_1 can take positive, zero and negative values. If the person who is selected in the sample does not belong to the sensitive group A, then he or she is requested to use the randomization device R_2 that is described below. Let α_2 and β_2 be any two positive real numbers (similar to Gjastvang and Singh (2006)) such that $T=\alpha_2/(\alpha_2+\beta_2)$ is the probability in the randomization device R_2 directing the selected respondent to report a scrambled response $w_2\beta_2 S_2$, and let $(1-T)=\beta_2/(\alpha_2+\beta_2)$ be the probability in the randomization device R_2 directing the selected respondent to report scrambled response as $-w_2\alpha_2 S_2$, where w_2 is a known real number and S_2 is any non-directional scrambling variables. The main difference from the existing randomization response models is that here the distribution of the scrambling variables S_1 and S_2 may or may not be known. Gjestvang and Singh (2006) have noted that the negative response will not disclose the privacy of any respondent belonging to non-sensitive or sensitive group because they come from both groups. Here we also note that if the mean θ_i and variance γ_i^2 of the ith scrambling variable S_i (i=1,2) are known before start of the survey, then in such a situation, the value of w_i may be the function of the known quantities (θ_i,γ_i^2), i=1,2.

Theorem 2.1 An unbiased estimator of the population proportion π is given by

$$\hat{\pi}_{HS}=\frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.1}$$

Proof The observed response in the proposed method has the distribution

$$y_i=\begin{cases} 1+w_1\beta_1 S_1 & \text{with probability } p\pi, \\ 1-w_1\alpha_1 S_1 & \text{with probability } (1-p)\pi, \\ w_2\beta_2 S_2 & \text{with probability } T(1-\pi), \\ -w_2\alpha_2 S_2 & \text{with probability } (1-T)(1-\pi). \end{cases} \tag{2.2}$$

Let E_1 and E_2 denote the expected values over all possible samples and over the randomization device. Then we have

$$E(\hat{\pi}_{HS})=E_1E_2(\hat{\pi}_{HS})$$

$$= \frac{1}{n} E_1 \sum_{i=1}^{n} E_2(y_i). \tag{2.3}$$

where

$$y_i = \pi p(1 + w_1\beta_1 S_1) + (1-p)\pi(1 - w_1\alpha_1 S_1) + T(1-\pi)\beta_2 S_2 w_2 - w_2\alpha_2(1-T)(1-\pi)S_2$$

Let $E_2(S_1) = \theta_1$ and $E_2(S_2) = \theta_2$. Then we have

$$E_2(y_i) = \pi p(1 + w_1\beta_1\theta_1) + (1-p)\pi(1 - w_1\alpha_1\theta_1) + T(1-\pi)\beta_2\theta_2 w_2 - w_2\alpha_2(1-T)(1-\pi)\theta_2,$$

$$= \pi\{p(1 + w_1\beta_1\theta_1) + (1-p)(1 - w_1\alpha_1\theta_1)\} + (1-\pi)w_2\theta_2(T\beta_2 - (1-T)\alpha_2),$$

$$= \pi\left\{1 + w_1\theta_1\left[\frac{\alpha_1\beta_1}{(\alpha_1+\beta_1)} - \frac{\alpha_1\beta_1}{(\alpha_1+\beta_1)}\right]\right\} + (1-\pi)w_2\theta_2\left\{\frac{\alpha_2\beta_2}{(\alpha_2+\beta_2)} - \frac{\alpha_2\beta_2}{(\alpha_2+\beta_2)}\right\},$$

$$= \pi + (1-\pi),$$

$$= \pi. \tag{2.4}$$

Putting (2.4) in (2.3) we get

$$E(\hat{\pi}_{HS}) = \frac{1}{n} E_1 \sum_{i=1}^{n} \pi$$
$$= \pi$$

which proves the theorem.

Theorem 2.2 The variance of the estimator $\hat{\pi}_{HS}$ is given by

$$V(\hat{\pi}_{HS}) \frac{\pi(1-\pi)}{n} + \frac{1}{n}\left\{\pi w_1^2\alpha_1\beta_1(\gamma_1^2 + \theta_1^2) + (1-\pi)w_2^2\alpha_2\beta_2(\gamma_2^2 + \theta_2^2)\right\} \tag{2.5}$$

Proof The responses are independent, thus the variance of the estimator $\hat{\pi}_{HS}$ is given by

$$V(\hat{\pi}_{HS}) = V\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} V(y_i). \tag{2.6}$$

Let V_1 and V_2 denote the variance over all possible samples and the variance over the randomization device respectively. Then we have

$$V(y_i) = E_1 V_2(y_i) + V_1 E_2(y_i)$$
$$= V_1 V_2(y_i) + V_1(\pi)$$

$$= E_1 V_2(y_i). \tag{2.7}$$

Let the variance of the scrambling variables be $V(S_1) = \gamma_1^2$ and $V(S_2) = \gamma_2^2$. Then

$$\begin{aligned} V_2(y_i) &= E_2(y_i^2) - \{E_2(y_i)\}^2 \\ &= \pi\left\{pE_2(1 + w_1\beta_1 S_1)^2 + (1-p)E_2(1 - w_1\alpha_1 S_1)^2\right\} \\ &\quad + (1-\pi)\left\{TE_2(w_2\beta_2 S_2)^2 + (1-T)E_2(-w_2\alpha_2 S_2)^2\right\} - \pi^2, \end{aligned}$$

$$\begin{aligned} &= \pi\left\{ p\left[1 + w_1^2\beta_1^2(\gamma_1^2 + \theta_1^2) + 2w_1\beta_1\theta_1\right]\right. \\ &\quad \left. + (1-p)\left[1 + w_1^2\alpha_1^2(\gamma_1^2 + \theta_1^2) - 2w_1\alpha_1\theta_1\right]\right\} \\ &\quad + (1-\pi)\left\{ T\, w_2^2\beta_2^2(\gamma_2^2 + \theta_2^2) + (1-T)w_2^2\alpha_2^2(\gamma_2^2 + \theta_2^2)\right\} - \pi^2, \end{aligned}$$

$$\begin{aligned} &= \pi(1-\pi) + \pi\left[\frac{w_1^2\alpha_1\beta_1^2(\gamma_1^2 + \theta_1^2)}{(\alpha_1 + \beta_1)} + \frac{2w_1\alpha_1\beta_1\theta_1}{(\alpha_1 + \beta_1)} + \frac{w_1^2\alpha_1^2\beta_1(\gamma_1^2 + \theta_1^2)}{(\alpha_1 + \beta_1)} - \frac{2w_1\alpha_1\beta_1\theta_1}{(\alpha_1 + \beta_1)}\right] \\ &\quad + (1-\pi)\left\{\frac{w_2^2\alpha_2\beta_2^2(\gamma_2^2 + \theta_2^2)}{(\alpha_2 + \beta_2)} + \frac{\beta_2 w_2^2\alpha_2^2(\gamma_2^2 + \theta_2^2)}{(\alpha_2 + \beta_2)}\right\} - \pi^2, \end{aligned}$$

$$= \pi(1-\pi) + \frac{\pi w_1^2\alpha_1\beta_1(\gamma_1^2 + \theta_1^2)(\alpha_1 + \beta_1)}{(\alpha_1 + \beta_1)} + \frac{(1-\pi)w_2^2(\gamma_2^2 + \theta_2^2)\alpha_2\beta_2(\alpha_2 + \beta_2)}{(\alpha_2 + \beta_2)}$$

$$= \pi(1-\pi) + \pi w_1^2\alpha_1\beta_1(\gamma_1^2 + \theta_1^2) + (1-\pi)w_2^2(\gamma_2^2 + \theta_2^2)\alpha_2\beta_2. \tag{2.8}$$

Thus from (2.6), (2.7) and (2.8) we have

$$V(\hat{\pi}_{HS}) = \frac{\pi(1-\pi)}{n} + \frac{1}{n}\left\{\pi w_1^2\alpha_1\beta_1(\gamma_1^2 + \theta_1^2) + (1-\pi)w_2^2(\gamma_2^2 + \theta_2^2)\alpha_2\beta_2\right\} \tag{2.9}$$

which proves the theorem.

Corollary 2.1 Assuming that

$$\frac{\alpha_2\beta_2}{\alpha_1\beta_1} = \frac{\gamma_1^2 + \theta_1^2}{\gamma_2^2 + \theta_2^2},$$

$(\gamma_2^2 + \theta_2^2) = 1$ [similar to Gjestvang and Singh (2006,p.525)] and

$w_1 = w_2 = w$ (say), the variance of the estimator $\hat{\pi}_{HS}$ in (2.5) reduces to

$$V(\hat{\pi}_{HS}) = \frac{\pi(1-\pi)}{n} + \frac{w^2\alpha_2\beta_2}{n}. \tag{2.10}$$

Proof is simple so omitted.

The variance in (2.10) of the proposed estimator $\hat{\pi}_{HS}$ can be estimated as

$$\hat{V}(\hat{\pi}_{HS}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \hat{\pi}_{HS})^2. \tag{2.11}$$

It should be remembered here that (w, α_2, β_2) are known quantities in the variance expression (2.10). As mentioned in Gjestvang and Singh (2006), we also show that models due to Warner (1965), Mangat and Singh (1990), Mangat (1994) and Gjestvang and Singh (2006) are special cases of the suggested RR procedure (model). If we set

(i) $\{p(1+w_1\beta_1\theta_1)+(1-p)(1-w_1\alpha_1\theta_1)\} = p_0$ and

$w_2\{T\beta_2\theta_2 - (1-T)\alpha_2\theta_2\} = (1-p_0)$,

(ii) $\{p(1+w_1\beta_1\theta_1)+(1-p)(1-w_1\alpha_1\theta_1)\} = (1-p_0)(1-T_0)$ and

$w_2\{T\beta_2\theta_2 - (1-T)\alpha_2\theta_2\} = \{1-(1-p_0)(1-T_0)\}$,

(iii) $\{p(1+w_1\beta_1\theta_1)+(1-p)(1-w_1\alpha_1\theta_1)\} = 1$ and

$w_2\{T\beta_2\theta_2 - (1-T)\alpha_2\theta_2\} = (1-p_0)$

(iv) $(w_1, w_2) = (1,1)$

the proposed RR model respectively reduces to the Warner (1965), Mangat and Singh (1990), Mangat (1994) and Gjestvang and Singh (2006) models.

3. Efficiency Comparison

In the proposed procedure, if we set $w_1 = w_2 = 1$, then the procedure investigated by Gjestvang and Singh (2006, sec.2, p.524) becomes special case (or member of the present proposed procedure).

In the Gjestvang and Singh (2006) model, the observed response has the distribution

$$y_i^* = \begin{cases} 1+\beta_1 S_1 & \text{with probability } p\pi, \\ 1-\alpha_1 S_1 & \text{with probability } (1-p)\pi, \\ \beta_2 S_2 & \text{with probability } T(1-\pi), \\ -\alpha_2 S_2 & \text{with probability } (1-T)(1-\pi). \end{cases} \tag{3.1}$$

This can be also obtained just by putting $w_1 = w_2 = 1$ in (2.2).

An unbiased estimator of π due to Gjestvang and Singh (2006) is given by

$$\hat{\pi}_{GS} = \frac{1}{n}\sum_{i=1}^{n} y_i^* . \tag{3.2}$$

The variance of $\hat{\pi}_{GS}$ is given by

$$V(\hat{\pi}_{GS}) = \frac{\pi(1-\pi)}{n} + \frac{1}{n}\left\{\pi\left(\gamma_1^2 + \theta_1^2\right)\alpha_1\beta_1 + (1-\pi)\left(\gamma_2^2 + \theta_2^2\right)\alpha_2\beta_2\right\}. \tag{3.3}$$

Assuming that

$$\frac{\alpha_2\beta_2}{\alpha_1\beta_1} = \frac{\gamma_1^2 + \theta_1^2}{\gamma_2^2 + \theta_2^2}$$

and $\gamma_2^2 + \theta_2^2 = 1$, then variance of $\hat{\pi}_{GS}$ in (3.3) reduces to

$$V(\hat{\pi}_{GS}) = \frac{\pi(1-\pi)}{n} + \frac{\alpha_2\beta_2}{n}. \tag{3.4}$$

From (2.5) and (3.3) we have

$$V(\hat{\pi}_{GS}) - V(\hat{\pi}_{HS}) = \frac{1}{n}\left[\pi\left(\gamma_1^2 + \theta_1^2\right)\alpha_1\beta_1\left(1 - w_1^2\right) + (1-\pi)\left(\gamma_2^2 + \theta_2^2\right)\alpha_2\beta_2\left(1 - w_2^2\right)\right] \tag{3.5}$$

which is positive if

$$\left(1 - w_i^2\right) > 0 \text{, i=1,2;}$$

i.e. if

$$-1 < w_i < 1 \text{, i=1,2}$$

$$\text{i.e. if } |w_i| < 1 \text{, i=1,2}$$

Thus, we established the following theorem.

Theorem 3.1 The proposed estimator $\hat{\pi}_{HS}$ (i.e. proposed procedure) is always better than Gjeatvang and Singh's (2006) estimator $\hat{\pi}_{GS}$ (i.e. Gjestvang and Singh's (2006) procedure) if

$$|w_i| < 1 \text{, i=1,2.} \tag{3.6}$$

Further, from (2.10) and (3.4) we have

$$V(\hat{\pi}_{GS}) - V(\hat{\pi}_{HS}) = \frac{\alpha_2\beta_2}{n}\left(1 - w^2\right) \tag{3.7}$$

which is always positive if

$$1-w^2>0$$
$$\text{i.e. if } \ |w|<1. \tag{3.8}$$

Thus, we established the following corollary.

Corollary 3.1 Under the assumption

$$\frac{\alpha_2\beta_2}{\alpha_1\beta_1}=\frac{\gamma_1^2+\theta_1^2}{\gamma_2^2+\theta_2^2}$$

and

$$w_1=w_2=w \text{ (a real number, say).}$$

The proposed estimator $\hat{\pi}_{HS}$ is more efficient than Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$ if $|w|<1$.

Assume that the values of $(\alpha_i,\beta_i,\theta_i,\gamma_i^2, i=1,2)$ are predetermined before conducting the survey and are assumed to be known. Note that θ_1 and θ_2 are non-directional. From (1.2) and (2.10) we have that $V(\hat{\pi}_{HS})<V(\hat{\pi}_w)$ if

$$w\alpha_2\beta_2<\frac{p_0(1-p_0)}{(2p_0-1)^2} \tag{3.9}$$

which is free from the parameter π under investigation and depends on the parameters of the randomization devices. We also note that the condition (3.9) is also very flexible.

From (1.4) and (2.10) we have

$$V(\hat{\pi}_{ms})-V(\hat{\pi}_{HS})=\frac{1}{n}\left[\frac{(1-p_0)(1-T_0)[1-(1-p_0)(1-T_0)]}{\{2p_0-1+2T_0(1-p_0)\}^2}-w\alpha_2\beta_2\right]$$

which is positive if

$$w\alpha_2\beta_2<\frac{(1-p_0)(1-T_0)[1-(1-p_0)(1-T_0)]}{[2p_0-1+2T_0(1-p_0)]^2}. \tag{3.10}$$

This condition is also free from the parameter π under investigation and depends on the parameters of the randomization devices.

Further, from (1.6) and (2.10) we have that $V(\hat{\pi}_m)<V(\hat{\pi}_{HS})$ if

$$w\alpha_2\beta_2<\frac{(1-p_0)(1-\pi)}{p_0} \tag{3.11}$$

Thus, the proposed RR model is more efficient than Warner's (1965) model, Mangat and Singh's (1990) model and Mangat's (1994) model as long as the conditions (3.9), (3.10) and (3.11) are respectively satisfied.

4. Some Members of the Proposed Procedure

I. Assume that the values of $\alpha_1, \beta_1, \alpha_2, \beta_2, \theta_1, \gamma_1^2, \theta_2$ and γ_2^2 are predetermined before conducting the survey and are assumed to be known. Note that θ_1 and θ_2 are non-directional [see Gjestvang and Singh (2006), sec.3, p.525)]. In our model, if we take $w_1 = \left(\frac{2\gamma_1\theta_1}{\gamma_1^2+\theta_1^2}\right)^{1/2}$ and $w_2 = \left(\frac{2\gamma_2\theta_2}{\gamma_2^2+\theta_2^2}\right)^{1/2}$ in (2.2), then the observed response has the distribution:

$$y_{i(1)} = \begin{cases} 1+\left(\frac{2\gamma_1\theta_1}{\gamma_1^2+\theta_1^2}\right)^{1/2}\beta_1 S_1 & \text{with probability } p\pi, \\ 1-\left(\frac{2\gamma_1\theta_1}{\gamma_1^2+\theta_1^2}\right)^{1/2}\alpha_1 S_1 & \text{with probability } (1-p)\pi, \\ \left(\frac{2\gamma_2\theta_2}{\gamma_2^2+\theta_2^2}\right)^{1/2}\beta_2 S_2 & \text{with probability } T(1-\pi), \\ -\left(\frac{2\gamma_2\theta_2}{\gamma_2^2+\theta_2^2}\right)^{1/2}\alpha_2 S_2 & \text{with probability } (1-T)(1-\pi). \end{cases} \tag{4.1}$$

Thus, an unbiased estimator of the population proportion π is given by

$$\hat{\pi}_{HS(1)} = \frac{1}{n}\sum_{i=1}^{n} y_{i(1)} . \tag{4.2}$$

Putting $w_1 = \left(\frac{2\gamma_1\theta_1}{\gamma_1^2+\theta_1^2}\right)^{1/2}$ and $w_2 = \left(\frac{2\gamma_2\theta_2}{\gamma_2^2+\theta_2^2}\right)^{1/2}$ in (2.5) we get the variance of $\hat{\pi}_{HS(1)}$ as

$$V\left(\hat{\pi}_{HS(1)}\right) = \frac{\pi(1-\pi)}{n} + \frac{1}{n}\left[2\gamma_1\theta_1\alpha_1\beta_1\pi + 2\gamma_2\theta_2\alpha_2\beta_2(1-\pi)\right]. \tag{4.3}$$

From (3.3) and (4.3) we have

$$V(\hat{\pi}_{GS}) - V(\hat{\pi}_{HS(1)}) = \frac{1}{n}\left[\pi\alpha_1\beta_1(\gamma_1 - \theta_1)^2 + (1-\pi)\alpha_2\beta_2(\gamma_2 - \theta_2)^2\right] \quad (4.4)$$

which is always positive provided $\gamma_1 \neq \theta_2$ and $\gamma_2 \neq \theta_2$. Thus, the proposed RR model (4.1) is always better than the RR model (3.1) due to Gjestvang and Singh (2006). In the situation where $\gamma_i = \theta_i$,(i=1,2), both the models are equally efficient.

II. If $w_1 = \dfrac{\theta_1}{\sqrt{\theta_1^2 + \gamma_1^2}}$ and $w_2 = \dfrac{\theta_2}{\sqrt{\theta_2^2 + \gamma_2^2}}$ in (2.2), then the observed response has the distribution:

$$y_{i(2)} = \begin{cases} 1 + \dfrac{\theta_1}{\sqrt{\theta_1^2 + \gamma_1^2}}\beta_1 S_1 & \text{with probability } p\pi, \\ 1 - \dfrac{\theta_1}{\sqrt{\theta_1^2 + \gamma_1^2}}\alpha_1 S_1 & \text{with probability } (1-p)\pi, \\ \dfrac{\theta_2}{\sqrt{\theta_2^2 + \gamma_2^2}}\beta_2 S_2 & \text{with probability } T(1-\pi), \\ -\dfrac{\theta_2}{\sqrt{\theta_2^2 + \gamma_2^2}}\alpha_2 S_2 & \text{with probability } (1-T)(1-\pi). \end{cases} \quad (4.5)$$

Thus, an estimator of the population proportion π is given by

$$\hat{\pi}_{HS(2)} = \frac{1}{n}\sum_{i=1}^{n} y_{i(2)}. \quad (4.6)$$

Inserting $w_1 = \dfrac{\theta_1}{\sqrt{\theta_1^2 + \gamma_1^2}}$ and $w_2 = \dfrac{\theta_2}{\sqrt{\theta_2^2 + \gamma_2^2}}$ in (2.5) we get the variance of (4.6) as

$$V(\hat{\pi}_{HS(2)}) = \frac{\pi(1-\pi)}{n} + \frac{1}{n}\left\{\pi\alpha_1\beta_1\theta_1^2 + (1-\pi)\alpha_2\beta_2\theta_2^2\right\} \quad (4.7)$$

From (3.3) and (4.7) we have

$$V(\hat{\pi}_{GS}) - V(\hat{\pi}_{HS(2)}) = \frac{1}{n}\left[\pi\alpha_1\beta_1\gamma_1^2 + (1-\pi)\alpha_2\beta_2\gamma_2^2\right] \quad (4.8)$$

which is always positive. Thus, the RR model proposed in (4.5) is superior to Gjestvang and Singh's (2006) RR model (3.1).

Assuming that

$$\frac{\alpha_2\beta_2}{\alpha_1\beta_1}=\frac{\theta_1^2}{\theta_2^2}, \tag{4.9}$$

the variance of $\hat{\pi}_{HS(2)}$ reduces to

$$V\left(\hat{\pi}_{HS(2)}\right)=\frac{\pi(1-\pi)}{n}+\frac{\alpha_2\beta_2\theta_2^2}{n} \tag{4.10}$$

From (3.4) and (4.10) we have

$$V\left(\hat{\pi}_{GS}\right)-V\left(\hat{\pi}_{HS(2)}\right)=\frac{\alpha_2\beta_2}{n}\left(1-\theta_2^2\right)$$
$$>0 \text{ if } \theta_2^2<1. \tag{4.11}$$

Thus, the proposed estimator $\hat{\pi}_{HS(2)}$ is more efficient than the Gjestvang and Singh (2006) estimator $\hat{\pi}_{GS}$ as long as the condition $\theta_2^2<1$ satisfied.

III. If we set $w_1=\frac{\gamma_1}{\sqrt{\theta_1^2+\gamma_1^2}}$ and $w_2=\frac{\gamma_2}{\sqrt{\theta_2^2+\gamma_2^2}}$ in (2.3), then the observed response has the distribution:

$$y_{i(3)}\begin{cases}1+\dfrac{\gamma_1}{\sqrt{\left(\theta_1^2+\gamma_1^2\right)}}\beta_1 S_1 & \text{with probability } p\pi,\\ 1-\dfrac{\gamma_1}{\sqrt{\left(\theta_1^2+\gamma_1^2\right)}}\alpha_1 S_1 & \text{with probability } (1-p)\pi,\\ \dfrac{\gamma_2}{\sqrt{\theta_2^2+\gamma_2^2}}\beta_2 S_2 & \text{with probability } T(1-\pi),\\ \dfrac{-\gamma_2}{\sqrt{\theta_2^2+\gamma_2^2}}\alpha_2 S_2 & \text{with probability } (1-T)(1-\pi).\end{cases} \tag{4.12}$$

Thus, an unbiased estimator of the population proportion π is defined by

$$\hat{\pi}_{HS(3)}=\frac{1}{n}\sum_{i=1}^{n}y_{i(3)}. \tag{4.13}$$

Putting $w_1 = \dfrac{\gamma_1}{\sqrt{\left(\theta_1^2 + \gamma_1^2\right)}}$ and $w_2 = \dfrac{\gamma_2}{\sqrt{\left(\theta_2^2 + \gamma_2^2\right)}}$ in (2.5) we get the variance of the estimator $\hat{\pi}_{HS(3)}$ as

$$V\left(\hat{\pi}_{HS(3)}\right) = \frac{\pi(1-\pi)}{n} + \frac{1}{n}\left\{\pi\alpha_1\beta_1\gamma_1^2 + (1-\pi)\alpha_2\beta_2\gamma_2^2\right\}. \tag{4.14}$$

From (3.3) and (4.14) we have

$$V\left(\hat{\pi}_{GS}\right) - V\left(\hat{\pi}_{HS(3)}\right) = \frac{1}{n}\left\{\pi\alpha_1\beta_1\theta_1^2 + (1-\pi)\alpha_2\beta_2\theta_2^2\right\} \tag{4.15}$$

which is always positive. Thus, it follows from (4.15) that the proposed estimator $\hat{\pi}_{HS(3)}$ is more efficient than Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$, i.e. the RR model suggested in (4.9) is superior to the RR model in (3.1) due to Gjestvang and Singh (2006).

Assuming that

$$\frac{\alpha_2\beta_2}{\alpha_1\beta_1} = \frac{\gamma_1^2}{\gamma_2^2}, \tag{4.16}$$

the variance of the estimator $\hat{\pi}_{HS(3)}$ in (4.14) is reduced to

$$V\left(\hat{\pi}_{HS(3)}\right) = \frac{\pi(1-\pi)}{n} + \frac{\alpha_2\beta_2\gamma_2^2}{n}. \tag{4.17}$$

It can be seen from (3.4) and (4.17) that

$$V\left(\hat{\pi}_{GS}\right) - V\left(\hat{\pi}_{HS(3)}\right) = \frac{\alpha_2\beta_2}{n}\left(1-\gamma_2^2\right)$$

which is positive if

$$\gamma_2^2 < 1 \tag{4.18}$$

Thus, the proposed estimator $\hat{\pi}_{HS(3)}$ is more efficient than Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$ as long as the condition (4.18) is satisfied.

Remark 4.1 For $w_i = \dfrac{\theta_i^2}{\left(\theta_i^2 + \gamma_i^2\right)}$, (i=1,2) $w_i = \dfrac{\gamma_i^2}{\left(\theta_i^2 + \gamma_i^2\right)}$, (i=1,2) and $w_i = \dfrac{\left|\left(\theta_i^2 - \gamma_1^2\right)\right|}{\left(\theta_i^2 + \gamma_i^2\right)}$ one can get the randomized response models always better than Gjestvan and Singh's (2006) randomized response models.

Many more suitable choices of w_1 and w_2 can be considered (which may be either the function of $(\theta_i, \gamma_i, i=1,2)$ or not) for which we can obtain the model superior to the Gjestvang and Singh (2006).

5. Relative Efficiency

It is assumed that the values of $\alpha_1, \beta_1, \alpha_2, \beta_2, \theta_1, \gamma_1^2, \theta_2$ and γ_2^2 are known before the start of the survey. It is to be noted that the Mangat (1994) model remains more efficient than the Mangat and Singh (1990) model. Also, Gjestvang and Singh (2006) have proved that the estimator $\hat{\pi}_{GS}$ proposed by them can always be made more efficient than the Warner (1965), Mangat and Singh (1990) and Mangat (1994) estimators for various choices of known parameters of the model. Thus, it is acceptable to compare the proposed model only with Gjestvang and Singh (2006).

To see the magnitude of the gain efficiency of the suggested randomized response model, we compute the percent relative efficiency (PRE) of the proposed estimator $\hat{\pi}_{HS}$ with respect to Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$ as follows.

$$PRE(\hat{\pi}_{HS}, \hat{\pi}_{GS}) = \frac{V(\hat{\pi}_{GS})}{V(\hat{\pi}_{HS})} \times 100 \tag{5.1}$$

or equivalently (by using (2.9) and (3.3) in (5.1))

$$PRE(\hat{\pi}_{HS,}\hat{\pi}_{GS}) = \frac{\left[\pi(1-\pi) + \left\{\pi\alpha_1\beta_1\left(\gamma_1^2 + \theta_1^2\right) + (1-\pi)\alpha_2\beta_2\left(\gamma_2^2 + \theta_2^2\right)\right\}\right]}{\left[\pi(1-\pi) + \left\{\pi w_1^2\alpha_1\beta_1\left(\gamma_1^2 + \theta_1^2\right) + (1-\pi)w_2^2\alpha_2\beta_2\left(\gamma_2^2 + \theta_2^2\right)\right\}\right]} \times 100 \tag{5.2}$$

Further, for the simplicity we have assumed $\left(\gamma_1^2 + \theta_1^2\right) = \left(\gamma_2^2 + \theta_2^2\right) = 1$ [similar to Gjestvang and Singh (2006), p.526] and $w_1 = w_2 = w$ (a real constant) under these assumptions, the $PRE(\hat{\pi}_{HS}, \hat{\pi}_{GS})$ in (5.2) reduces to :

$$PRE(\hat{\pi}_{HS}, \hat{\pi}_{GS}) = \frac{\left[\pi(1-\pi) + \left\{\pi\alpha_1\beta_1 + (1-\pi)\alpha_2\beta_2\right\}\right]}{\left[\pi(1-\pi) + w^2\left\{\pi\alpha_1\beta_1 + (1-\pi)\alpha_2\beta_2\right\}\right]} \times 100 \tag{5.3}$$

We have computed the $PRE(\hat{\pi}_{HS}, \hat{\pi}_{GS})$ by using (5.3) for $\pi = 0.05, 0.1(0.1)0.9$, and for three sets of α_i's, β_i's (i=1,2) values as (i) $\alpha_1 = 0.6, \beta_1 = 0.4, \alpha_2 = 0.3, \beta_2 = 0.7$ (ii) $\alpha_1 = 0.8, \beta_1 = 0.2, \alpha_2 = 0.4,$

$\beta_2 = 0.6$ (iii) $\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 0.5$ $|w| = 0.05, 0.1(0.1)0.9$. Findings are compiled in Table 5.1.

Table 5.1. The percent relative efficiency of the proposed model with respect to Gjestvang and Singh's (2006) model

$\alpha_1 = 0.6, \beta_1 = 0.4, \alpha_2 = 0.3, \beta_2 = 0.7$

π \ $\lvert w \rvert$	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.05	104.85	104.81	104.65	104.40	104.05	103.60	103.06	102.42	101.70	100.89
0.1	104.81	108.51	108.23	107.77	107.13	106.32	105.34	104.21	102.94	101.53
0.3	104.65	119.93	119.21	118.03	116.41	114.40	112.03	109.36	106.43	103.29
0.5	104.40	133.35	132.02	129.85	126.94	123.38	119.29	114.79	110.00	105.04
0.7	104.05	160.55	157.66	153.06	147.06	140.00	132.24	124.11	115.89	107.80
0.9	293.17	288.97	273.32	250.69	224.65	198.18	173.24	150.80	131.20	114.35

$\alpha_1 = 0.8, \beta_1 = 0.2, \alpha_2 = 0.4, \beta_2 = 0.6$

π \ $\lvert w \rvert$	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.05	102.90	102.87	102.78	102.64	102.43	102.16	101.84	101.46	101.03	100.54
0.1	105.22	105.17	105.01	104.74	104.36	103.87	103.28	102.60	101.82	100.95
0.3	112.66	112.56	112.13	111.42	110.45	109.23	107.77	106.10	104.23	102.19
0.5	121.56	121.36	120.58	119.30	117.55	115.38	112.84	109.97	106.84	103.50
0.7	139.58	139.16	137.53	134.89	131.37	127.10	122.24	116.96	111.40	105.71
0.9	225.60	223.49	215.43	203.21	188.27	172.00	155.57	139.79	125.15	111.86

$\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 0.5$

π \ $\lvert w \rvert$	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.05	104.37	104.34	104.20	103.98	103.66	103.25	102.76	102.19	101.54	100.80
0.1	107.92	107.85	107.59	107.17	106.58	105.84	104.94	103.90	102.72	101.42
0.3	119.42	119.25	118.56	117.42	115.87	113.93	111.65	109.07	106.24	103.20
0.5	133.22	132.89	131.58	129.45	126.58	123.08	119.05	114.61	109.89	104.99
0.7	161.16	160.42	157.53	152.95	146.96	139.92	132.18	124.07	115.87	107.79
0.9	294.21	289.98	274.19	251.39	225.17	198.54	173.47	150.94	131.27	114.38

It is observed from Table 5.1 that the values of $\mathrm{PRE}\left(\hat{\pi}_{\mathrm{HS},}\hat{\pi}_{\mathrm{GS}}\right)$ are larger than 100 for the given parametric values. It follows that the suggested estimator $\hat{\pi}_{\mathrm{HS}}$ can always be made more efficient than Gjestvang and Singh's (2006) estimator $\hat{\pi}_{\mathrm{GS}}$ and hence more efficient than the Warner (1965), Mangat and

Singh (1990) and Mangat (1990) estimators. For larger values (or even moderately large values) of $|w|$ and π, the considerable gain in efficiency is observed by using the proposed estimator $\hat{\pi}_{HS}$ over Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$. Thus, we see that the proposed procedure is an improvement over Gjestvang and Singh's (2006) procedure.

We have further computed the percent relative efficiencies (PRE's) of the proposed estimators $\hat{\pi}_{HS(1)}, \hat{\pi}_{HS(2)}$ and $\hat{\pi}_{HS(3)}$ with respect to Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$ by using the formulae:

$$PRE\left(\hat{\pi}_{HS(1)}, \hat{\pi}_{GS}\right) = \frac{\left[\pi(1-\pi) + \pi\left(\gamma_1^2 + \theta_1^2\right)\alpha_1\beta_1 + (1-\pi)\left(\gamma_2^2 + \theta_2^2\right)\alpha_2\beta_2\right]}{\left[\pi(1-\pi) + 2\left\{\gamma_1\theta_1\alpha_1\beta_1\pi + \gamma_2\theta_2\alpha_2\beta_2(1-\pi)\right\}\right]} \times 100 \tag{5.4}$$

$$PRE\left(\hat{\pi}_{HS(2)}, \hat{\pi}_{GS}\right) = \frac{\left[\pi(1-\pi) + \pi\left(\gamma_1^2 + \theta_1^2\right)\alpha_1\beta_1 + (1-\pi)\left(\gamma_2^2 + \theta_2^2\right)\alpha_2\beta_2\right]}{\left[\pi(1-\pi) + \pi\alpha_1\beta_1\theta_1^2 + (1-\pi)\alpha_2\beta_2\theta_2^2\right]} \times 100 \tag{5.5}$$

$$PRE\left(\hat{\pi}_{HS(2)}, \hat{\pi}_{GS}\right) = \frac{\left[\pi(1-\pi) + \pi\left(\gamma_1^2 + \theta_1^2\right)\alpha_1\beta_1 + (1-\pi)\left(\gamma_2^2 + \theta_2^2\right)\alpha_2\beta_2\right]}{\left[\pi(1-\pi) + \pi\alpha_1\beta_1\gamma_1^2 + (1-\pi)\alpha_2\beta_2\gamma_2^2\right]} \times 100 \tag{5.6}$$

for $\left(\theta_1, \gamma_1^2\right) = (0.6, 0.50)$, $\left(\theta_2, \gamma_2^2\right) = (0.8, 0.36)$, $(\alpha_1, \beta_1) = (0.6, 0.4)$, $(\alpha_2, \beta_2) = (0.05, 0.95)$ [similar to Gjestvang and Singh (2006), section 4, p.527]. Findings are given in Table 5.2.

Table 5.2. The percent relative efficiencies of $\hat{\pi}_{HS(1)}, \hat{\pi}_{HS(2)}$ and $\hat{\pi}_{HS(3)}$ with respect to Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$

π	$PRE\left(\hat{\pi}_{HS(1)}, \hat{\pi}_{GS}\right)$	$PRE\left(\hat{\pi}_{HS(2)}, \hat{\pi}_{GS}\right)$	$PRE\left(\hat{\pi}_{HS(3)}, \hat{\pi}_{GS}\right)$
0.1	101.31	121.74	130.67
0.2	100.87	118.69	121.04
0.3	100.71	118.65	118.30
0.4	100.64	119.90	117.70
0.5	100.62	122.23	118.33
0.6	100.63	125.93	120.07
0.7	100.68	131.88	123.27
0.8	100.78	142.27	128.99
0.9	100.96	164.23	140.46

It is observed from Table 5.2 that the percent relative efficiencies of the proposed estimators $\hat{\pi}_{HS(1)}$, $\hat{\pi}_{HS(2)}$ and $\hat{\pi}_{HS(3)}$ with respect to Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$ are larger than 100. It follows that the proposed estimators are more efficient than Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$. We note that there is a marginal gain in efficiency by using the proposed estimator $\hat{\pi}_{HS(1)}$ over Gjestvang and Singh's (2006) estimator $\hat{\pi}_{GS}$ while the gain in efficiency is substantial by using the suggested estimators $\hat{\pi}_{HS(2)}$ and $\hat{\pi}_{HS(3)}$. The proposed estimator $\hat{\pi}_{HS(2)}$ is more efficient than the estimator $\hat{\pi}_{HS(3)}$ as long as $\pi < 0.3$. On the other hand, if $\pi \geq 0.3$ the proposed estimator $\hat{\pi}_{HS(3)}$ is better than the estimator $\hat{\pi}_{HS(2)}$. However, the proposed estimators $\hat{\pi}_{HS(2)}$ and $\hat{\pi}_{HS(3)}$ are more efficient than the estimator $\hat{\pi}_{HS(1)}$. Thus, we conclude that the proposed estimator $\hat{\pi}_{HS(2)}$ is a suitable choice for $\pi < 0.3$, whereas for $\pi \geq 0.3$, the estimator $\hat{\pi}_{HS(3)}$ is the appropriate choice for estimating the population proportion $\hat{\pi}_{HS(3)}$.

Finally, we conclude that the suggested general procedure is justifiable in the sense of obtaining better estimators from the proposed generalized estimator $\hat{\pi}_{HS}$ for appropriate values of (w_1, w_2).

REFERENCES

ANTONAK, R. F., LIVNEH, H., (1995). Randomized response technique: A review and proposed extension to disability attitude research. Genetic, Social, and general Psychology Monographs, 121, pp. 97–145.

BOUZA, C. N., HERRERA, C., MITRA, P. G., (2010). A Review of Randomized Responses Procedures: The Qualitative Variable Case. Revista Investigación Operacional, 31 (3), 240–247.

CHAUDHURI, A., MUKERJEE, R., (1987). Randomized Response Technique: A Review. Statistica Neerlandica, 41, pp. 27–44.

CHAUDHURI, A., MUKERJEE, R., (1988). Randomized Response. Statistics: Textbooks and Monographs, Vol. 85, Marcel Dekker, Inc., New York, NY.

CHAUDHURI, A., CHRISTOFIDES, T. C., RAO, C. R., (2016). Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits. Handbook of Statistics 34.

FISHER, M., KUPFERMAN, L. B., LESSER, M., (1992). Substance use in a school-based clinic population: Use of the randomized response technique to estimate prevalence. Journal of Adolescent Health, 13, pp. 281–285.

GJESTVANG, C. R., SINGH, S., (2006). A New Randomized Response Model. Journal of the Royal Statistical Society, B. 68, pp. 523–530.

JARMAN, B. J., (1997). The Prevalence and Precedence of Socially Condoned Sexual Aggression Within a Dating Context as Measured by Direct Questioning and the Randomized Response Technique

MANGAT, N. S., SINGH, R., (1990). An Alternative Randomized Procedure. Biometrika, 77, pp. 439–442.

MANGAT, N. S., (1994). An Improved Randomized Response Strategy. Journal of the Royal Statistical Society., B, 56 (1), pp. 93–95.

ODUMADE, O., SINGH, S., (2008). Generalized Force Quantitative Randomized Response Model: A Unified Approch. Journal of the Indian Society and Agricultural Statistics, 62 (3), pp. 244–252.

ODUMADE, O., SINGH, S., (2009). Improved Bar-Lev. Bobovitch, Boukai Randomized Response Model. Communications in Statistics – Simulation and Computation, 38, pp. 473–502.

ODUMADE, O., SINGH, S., (2009). Efficient Use of Two Decks of Cards in Randomized Response Sampling. Communications in Statistics – Theory and Methods 38, pp. 439–446.

ODUMADE, O., SINGH, S., (2010). An Alternative to The Bar-Lev, Bobovitch and Boukai Randomized Response Model. Sociological Methods And Research. Doi10.1177/0049124110378094.

SINGH, S., (2003). Advanced Sampling Theory With Applications. Kluwer Academic Publishers, Dordrecht.

WARNER, S. L., (1965). Randomized Response: A Survey Technique For Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60, pp. 63–69.

WEISSMAN, A. N., STEER, R. A., LIPTON, D. S., (1986). Estimating illict drug use through telephone interviews and the randomized response technique. Drug and Alcohol Dependence, 18, 225–233.

WILLIAMS, B. L., SUEN, H., (1994). A Methodological Comparison of Survey Techniques in Obtaining Self-Reports of Condom-Related Behaviors. Psychological Reports, 7, pp. 1531–1537.

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 687–699, DOI 10.21307/stattrans-2017-007

THE INTER-COUNTRY COMPARISON OF THE COST OF CHILDREN MAINTENANCE USING HOUSING EXPENDITURE[1]

Małgorzata Kalbarczyk[2], Agata Miazga[3], Anna Nicińska[4]

ABSTRACT

It is interesting to compare maintenance costs of children between countries with similar yet different family policy regimes because this could yield valuable lessons for researchers and policy-makers and also for the sake of methodological development.

In this study, we aim to conduct a comparative analysis of the equivalence scales in Austria, Italy, Poland and France taking into account the age of children. To this end, we use data from the European Income and Living Condition (EU-SILC) to calculate equivalence scales for mono- and duo-parental households for the first and second child. The four countries share common European cultural context, yet differ with respect to social environment, in particular to family policy. We apply the Engel estimation method proposing the share of housing spending in total expenditures as a tool to obtain commodity-specific equivalence scales.

Our results are consistent with other studies showing that the cost of a first child is higher than that of a later child. The scale values are not the same across all the countries concerned, with the highest cost observed in Italy and the lowest in Poland.

Key words: equivalence scales, EU-SILC, housing expenses, Engel curves

[1] This study was financed by the National Science Centre within the project 'Preventing aging of the population using the instruments of fiscal policy and migration policy' (DEC-2012/07/B/HS4/03254). The presented analysis is based on the results of the EU-SILC survey (2010). The content of this article does not reflect the official opinion of the European Union, the European Commission, Eurostat or any European statistical institution providing the data. Responsibility for the information and views expressed in the article lies entirely with the authors. The useful comments from Agnieszka Fihel are gratefully acknowledged.

2 Faculty of Economics Sciences University of Warsaw. E-mail: mkalbarczyk@wne.uw.edu.pl.
3 Institute for Structural Research. E-mail: agata.miazga@ibs.org.pl.
4 Faculty of Economics Sciences University of Warsaw. E-mail: anicinska@wne.uw.edu.pl.

1. Introduction

The calculation of equivalence scales measuring the maintenance cost of children may be based on multiple estimation methods (Barten, 1964; Betti, Lundgren 2012; Gorman, 1976; Pashardes, 1991; Pollak, Wales, 1979; Szulc, 2009). Obviously, the choice of an appropriate method of estimation depends primarily on the research objectives and, secondly, on the availability of data. Typically, in the case of the analysis for a single country, researchers have a much wider choice of options than in the case of an analysis aimed at international comparisons, where the method selection is much more frequently restricted by the availability of data comparable between respective countries. Many methods of estimations, commonly used in international comparisons, are based on the expenditures on food as a measure of welfare. In this paper, we attempt to verify whether equivalence scales calculated on the basis of housing expenses are capable of indicating internationally comparable costs of children maintenance. In particular, we aim at distinguishing between the maintenance cost of the first and the second child and between different age categories of children.

The study was carried out for four European countries: Austria, France, Italy and Poland. The choice of these countries was driven by differences in implemented family policies and in family benefits spending, which may impact on the dominant model of care and, consequently, maintenance costs of children. As for the level of family-related public spending, in 20135 France was the most generous from all these countries, with approximately 3.65% of Gross Domestic Product (GDP) transferred to families in the form of various childcare benefits (OECD, 2016a). These transfers were lower in Austria (2.61% GDP), Italy (1.97%) and Poland (1.61% in 2012). Children's participation rate in the preschool institutions (creches and nursery schools) was highly differentiated too, with the lowest value in Poland and the highest in France (OECD, 2016b). Making a reference to Esping-Andersen's (1990) typology of welfare state regimes, we can distinguish Austria and Italy, where the public support for families is dominated by direct financial transfers making up for incomes lost by a parent taking care of children, and France that combines direct financial transfer, important fiscal deductions and an extended public infrastructure of pre-school institutions. In Poland, where the family policy is still underdeveloped as compared to the three other countries of analysis, relatively long family leaves are accompanied by financial direct transfers and limited infrastructure of pre-school institutions. Different levels of public spending and different types of family-related policy instruments in these countries may affect importantly the level of children maintenance costs.

This paper is organized as follows. In the first part, we review estimation methods used in the calculation of the equivalence scales. In the second part, we present the data used for estimations, elaborate on our method and discuss the

[5] The latest year for which the most updated data were available.

results that we obtained for four European countries. Concluding remarks include our reflexion on the use of housing expenses in international comparisons of maintenance costs of children.

2. Measures of the cost of child maintenance

There are a vast number of methods for estimating the cost of children in the economic literature. The cost of children is most frequently defined as the incremental income the parents must spend after the birth of a first or later child. Firstly, this does not account for the public cost of children incurred by the government. Secondly, alternative costs, the major one being the cost of lost income resulting from partial or full withdrawal from professional activity for the sake of childcare, are not considered. The easiest way of estimating such individual direct costs is to compare the budgets of childless persons to those who have children, i.e. by calculating equivalence scales (Panek, 2011). By using equivalence scales, one can estimate how much more a household of a certain demographic structure must spend as compared to a reference household, e.g. a childless one, in order to achieve an equal level of welfare (Szulc, 2007).

Equivalence scales are calculated according to the demographic structure and the expenses of a household, rather than its income for three main reasons. Firstly, when declaring their expenses, respondents tend to be more accurate than in assessing their incomes. Secondly, expenses are a better indicator of permanent income, that is an income earned in a lifetime perspective and, thirdly, they more accurately reflect the respondents' standards of living (Dudek, 2011). The demographic structure of a household most frequently means the composition of the household, including the number of both adults and children.

Equivalence scales may be calculated in many ways and we distinguish, most basically, two types of scales: normative and empirical. The former, also referred to as expert scales as independent experts assess the welfare needs of adults and children, include the OECD scale, Luxembourg Income Study (LIS) scale or scales devised by national offices of statistics in individual countries (Ciecieląg, 2003). Their advantage is the simplicity of calculations and ease of comparisons, the drawback being the arbitrary selection of weights (Dudek, 2009).

Empirical scales, in turn, are based on the observation of actual consumption of households (in the so-called objective approach) or their subjective declared assessment of the capability to maintain on their own (subjective approach) (Dudek, 2011). Among the objective approximate methods, the method described by Engel (1895) is the oldest and, at the same time, most popular. It involves the comparison of spending between families of different demographic structure and the same welfare level, which is measured by the share of expenditures on food in the total spending of households.

In order to calculate the Engel scale, the so-called Engel curves have to be estimated on the basis of single-equation econometric models (Panek 2011). The

dependent variable in these models is the share of expenditures on food in all expenditures, whereas the explanatory variables include income and demographic characteristics of households. In the next step, in order to calculate equivalence scales, the shares of food spend of the reference household is compared to the respective share of a household with the selected number of children. Equivalence scales are obtained by comparing total spending x of households with different demographic structure with total spending of the reference household x^0.

The main objection raised against the Engel method (Ciecieląg 2003, Dudek 2011) is that it only considers expenses on food. Another objection concerns the fallacy of the assumption as to the equality of preferences of children and adults. Despite these objections, the method is frequently used in empirical studies, mainly due to its simplicity and high availability of required data.

An alternative to the Engel method is constituted by the welfare indicator proposed by Rothbarth (1943), who measured households' welfare on the basis of the absolute level of spending on the so-called adult goods, i.e. those consumed by adults only, such as for instance alcohol and cigarettes. Most researchers claim that, contrary to the Engel method, the cost of children maintenance obtained using the Rothbarth method is underestimated (Dudek 2011). This is because no change of preferences as regards adult goods is admitted following the birth of a next family member.

Another group of methods for estimating the cost of children, which seems more accurate but also more difficult to apply, is represented by methods based on utility functions, also known as complete demand models (Muellbauer 1974). These scales are a function of utility, which is not observable in the reality. This is the strongest objection against this type of equivalence scales, known as the issue of equivalence scales identifiability (Ciecieląg 2003, Dudek 2011). In order to identify the model, which is the basis for estimating the equivalence scales, it is necessary to input additional information on households or to make additional assumptions (Blundell 1998, Lewbel and Pendakur 2006), e.g. as to the independence of the scales on the utility level according to the ESE (Equivalence Scale Exactness) or IB (Independence of Base) option (Ciecieląg 2003).

Controversies around the results obtained through the above-described objective methods led to the emergence of subjective methods that, however, are still not as commonly used as the approximate methods (Dudek 2011). Instead of real spending data, subjective methods rely on respondents' opinions about their incomes. The opinions are gathered by the means of a questionnaire in which the respondents indicate the level of income corresponding, in their opinion, to specific ranking level (Leiden method). Usually, the following ranking scale is used for the income level: very bad, bad, insufficient, barely sufficient, good and very good (van Praag, van der Sar 1988).

Each of the above methods has both advantages and disadvantages. In short, normative scales, mostly used for international comparisons, are established by expert and need not to reflect empirical results of estimations. The Engel scale does not consider the effect of scale arising when a new family member is born,

thus overestimating the maintenance costs. On the contrary, the Rothbarth method underestimates these costs as no assumption is made that the consumption of the so-called adult goods changes when a family enlarges. Deaton and Muellebauer (1986) discuss the limitations of these methods in more detail. Methods based on utility functions seem more precise, although distinctly more complicated. And subjective methods require collecting additional statistical data, which is time- and cost-consuming.

3. Empirical analysis

3.1. Data and methodology

Data used in this study were derived from the European Union Statistics on Income and Living Conditions (the EU-SILC) database for the year 2010. The EU-SILC study is carried out according to a harmonized questionnaire on a sample of around 130 thousand households in 27 countries of the European Union, as well as Island and Norway. The EU-SILC database provides comparable multidimensional microdata on incomes, poverty, social exclusion, labour, education and health, both at the household and individual level. The EU-SILC household budget survey used in this study captures the income and living conditions for majority of European countries, including social and demographic characteristics of the respondents, their income and spending.

In this analysis, we applied the Engel method for the calculation of the cost of children. However, in contrast to the original approach, we used the share of housing expenses in total spending as a tool to compute the commodity-specific equivalence scales. Our results should be interpreted very carefully because housing is rather a public household good whereas food is rather private. Methods based on food expenditure overestimate and those based on housing expenditure underestimate child costs due to economies of scale. In absolute terms, the levels of housing expenses and incomes remain varied in Austria, France, Italy and Poland (Table 1), for households with positive income. In particular, Poland registers a considerably lower level of expenses and incomes than the other three countries. In relative terms, the share of housing expenses in the average income is very similar in Austria (11.5%), France (10.4%) and Italy (9.3%), and visibly higher in Poland (14.0%). Meanwhile, the average number of children is the most elevated in Polish households (1.25), mostly because of visibly higher proportion of households with children aged 6 and over. This may be due to the facts that fertility rates were still high in Poland at the turn of the 1980s and 1990s, and that Polish adolescent leave their family houses relatively late, as compared to their counterparts in three other countries of our analysis.

Table 1. Descriptive statistics of EU-SILC data for Austria, France, Poland and Italy

Country	Austria	France	Poland	Italy
Average yearly income (Euro)	43,417	44,875	10,473	36,441
Average monthly housing expenses (Euro)	500.85	465.44	146.73	339.09
Average number of children	1.01	1.18	1.25	0.88
Share of children: under age of 3	4.99	6.25	6.63	4.56
aged 3-6	6.51	7.64	7.73	6.46
aged 6-18	21.08	23.21	26.28	20.26
aged 18-25	8.07	10.13	15.82	10.76
Number of observations	6,188	11,044	12,930	19,147

Source: Author's own analysis based on EU-SILC data.

Based on the EU-SILC data, the equivalence scales were calculated by comparing the share of housing expenses in total spending for households with different demographic structure. Several assumptions were made here. Firstly, the scales were calculated separately for single parents and households with both parents raising the children together. In the first case, a single individual without children was taken as the reference household, while in the second one it was a household constituted by a couple with no children. For all cases, other individuals cohabiting with the family in a single household, apart from the children and parents, are possible.

Secondly, two definitions of a child were considered. According to one, this means any individual up to the age of 18. According to the other, apart from individuals up to the age of 18, the term includes also those under the age of 25 who continue their studies and remain to be supported by their parents. The analysis distinguishes also various age groups of children, assuming age brackets that are at least partly aligned with the applicable education system. The group of children were broken down by age into the following brackets: age up to 3, 3-6 years, 6-18 years and 18-25 years.

Thirdly, the presented results were limited only to households with one child or two children. As the percentage of households with three children in the analysed sample was at the very low level of 3.73% (lowest in Italy – 2.35%, highest in Poland – 4.91%), the estimates of the cost of a third and later child would have been inaccurate. Accordingly, the calculated equivalence scales show the cost of a first and a second child.

3.2. Equivalence scales by child's order

Table 2 presents the equivalence scales estimated using the share of housing expenses in total spending as the welfare measure for four countries of Europe, separately for households with one parent and with two parents. The highest cost

of children raised in households with two parents is observed in Italy (Table 2). The cost of a first child reaches around 55% of the spending of a childless household of two adults, and the marginal cost of a second child in Italy corresponds to additional 17% of the reference household spending. The marginal costs of a first and second child in a two-parent household remain lower in Austria and France, but still higher than in Poland. We observe the lowest marginal costs of a first and second child were observed in the latter, both in the case of single-parent households (19 percentage points and 12 p.p., respectively) and two-parent households (approx. 29 p.p. and approx. 9 p.p., respectively).

Table 2. Marginal cost of children in Austria, France, Poland and Italy

Country	Austria		France		Poland		Italy	
Child age limit	18	25	18	25	18	25	18	25
Single parent								
1st child	0.229	0.224	0.339	0.333	0.193	0.188	0.276	0.277
2nd child	0.139	0.145	0.184	0.190	0.119	0.124	0.215	0.214
Two parents								
1st child	0.369	0.391	0.322	0.345	0.285	0.307	0.547	0.541
2rd child	0.155	0.158	0.112	0.115	0.094	0.096	0.172	0.171

Source: Authors' own analysis based on EU-SILC data.

The difference in the level of cost of child maintenance in Poland and Italy is striking as the countries are characterised by similar family policy and, at least at first glance, traditional approach to the involvement of women in the care activities. Despite numerous similarities, Poland differs from Italy in terms of professional activity of women. According to Eurostat data, in 2014 the employment rate among women in production age in Poland was 55%, and in Italy it was lower by 8 p.p., standing at 47% (Eurostat 2015). Many Polish women decide to set up a family only after they gain the eligibility to financial benefits during the leave, and return to professional activity once their children become more self-reliant. Italian women much more frequently remain permanently professionally inactive. Additionally, the maternity benefit in Poland is characterised by the highest income replacement rate (100%) while in Italy it is lower (80%) and paid only over 13 weeks (EP 2014). Salaries and benefits obtained by working mothers may explain the differences in the cost of children in Poland and Italy.

Social policy supporting single parents results in slightly lower marginal cost of a first child in all the countries. This effect is most strongly visible in Italy, especially for a first child. Consequently, the costs of a first child in single-parent households in Italy are lower than in France. However, the marginal cost of a second child again is the highest in Italy. In all countries except Austria, the

marginal cost of a second child is higher in single-parent households than in those with two parents.

In all countries, the economies of scale are visible as the marginal cost of a second child is lower than the marginal cost of a first child. The largest economies of scale in the case of single parents are observed in France, which is probably largely driven by the design of the local family policy with strong incentives for having a second child and later children. In the case of parents raising children together, the largest economies of scale are seen in Italy, which is not surprising considering the very high cost of a first child.

3.3. Equivalence scales by child's age and number of parents in household

While in France and Austria the older the child, the lower its marginal maintenance cost (Tables 3 and 4), we obtained different results for Italy and Poland. In the former the relationship was opposite, whereas in the latter it is non-linear – highest values relate to the middle child age group. In each country the same pattern was maintained, regardless of the number of parents and children, assuming that in the case of two children, both belong to the same age group.

Table 3. Marginal cost of children up to 18 years old in single-parent households by child age group

Country	Austria		France		Poland		Italy	
Child age limit	18	25	18	25	18	25	18	25
One child								
Child age								
cat1	0.278	0.256	0.459	0.437	0.187	0.170	0.209	0.213
cat2	0.244	0.224	0.419	0.399	0.236	0.214	0.271	0.276
cat3	0.166	0.192	0.140	0.163	0.156	0.181	0.348	0.341
Two children								
Child age								
cat1, cat1	0.442	0.417	0.710	0.684	0.303	0.285	0.388	0.392
cat2, cat2	0.390	0.368	0.648	0.626	0.376	0.350	0.484	0.490
cat3, cat3	0.276	0.322	0.231	0.271	0.258	0.302	0.605	0.593
cat1, cat2	0.416	0.392	0.679	0.655	0.339	0.317	0.435	0.440
cat1, cat3	0.356	0.368	0.451	0.463	0.281	0.294	0.492	0.489
cat2, cat3	0.332	0.345	0.424	0.438	0.316	0.326	0.543	0.540

Note: cat1 – children under age of 3, cat2 – children aged 3-6, cat3 – children aged 6-18.
Source: Authors' own analysis based on EU-SILC data.

In Austria, France and Poland the cost of children in households with a second little child is the lowest when the first child belongs to the oldest age group. This is probably due to the fact that the costs of children decrease with age in those

countries. The relationship is observed both in single-parent and two-parent households. In Poland, as opposed to Austria and France, the cost of children in the first years of life is lower when the first child is in the same, lowest, age bracket, rather than the middle bracket, independently of the number of parents in the household. In this respect, Poland and Italy are similar. This relationship may be explained by limited access to public care for youngest children in these countries. Economies of scale allow limiting the cost of formal and informal care in the case when both children are of similar age.

Table 4. Marginal cost of children in two-parent households by child age group

Country	Austria		France		Poland		Italy	
Child age limit	18	25	18	25	18	25	18	25
One child								
Child age								
cat1	0.451	0.463	0.356	0.368	0.281	0.294	0.492	0.489
cat2	0.424	0.438	0.332	0.345	0.316	0.326	0.543	0.540
cat3	0.231	0.271	0.276	0.322	0.258	0.302	0.605	0.593
Two children								
Child age								
cat1, cat1	0.663	0.672	0.491	0.498	0.372	0.382	0.629	0.626
cat2, cat2	0.618	0.628	0.451	0.459	0.429	0.434	0.713	0.711
cat3, cat3	0.299	0.353	0.361	0.422	0.336	0.396	0.816	0.799
cat1, cat2	0.640	0.650	0.471	0.478	0.400	0.408	0.670	0.668
cat1, cat3	0.470	0.504	0.424	0.459	0.354	0.389	0.720	0.710
cat2, cat3	0.450	0.484	0.405	0.440	0.382	0.415	0.764	0.755

Note: cat1 – children under age of 3, cat2 – children aged 3-6, cat3 – children aged 6-18.
Source: Authors' own analysis based on EU-SILC data.

Only in France the cost of children appears higher in households with a single parent than in households with two parents. It should be noted that the cost of children of single parents is calculated referring to single-person households rather than to childless couple households, as in the case of two-parent households. When comparing the two types of households with children, we are not in the position to discern the impact of child presence from the impact of different spending structures for single-person households and couple households respectively.

If, for the sake of this study, we accept an assumption that the spending structure of all households with no children is the same, regardless of the number

of adults, the following relationships may be observed. The smallest difference between households with one and two parents in the cost of children occurs in Poland, and the largest in Italy, irrespective of the child age. The largest difference in the cost of a first child between households with a single parent and two parents, respectively, is observed for the youngest children in Poland and in Italy, and for the children from the middle age bracket in Austria. As far as the cost of a large number of children is considered, there is no constant pattern reflecting the effects of the support for single parents in the countries concerned.

4. Conclusions

The study presents the calculation results of the cost of children using the share of housing expenses in total spending. Consistently with other studies (Balli, Tiezzi, 2013; Kot, 2014), the equivalence scales calculated using the Engel method indicate that the cost of a first child is higher than that of a later child, be it in Austria, France, Poland or Italy. The differences in the cost of children depending on the assumed upper child age limit are insignificant, and for two children practically unnoticeable. This means that the maintenance cost of adult children is negligible. The scale values are not the same across all the countries concerned, with the highest cost observed in Italy and the lowest in Poland.

Analyses comparing the cost of children between countries are rare. To the best of the authors' knowledge, in the literature there is no study based on an objective method to cover the four countries (Austria, France, Italy and Poland). An analysis carried out by Bishop et al. (2014) and Kalbarczyk-Stęclik et al. (2017) based on a subjective method considers a wide set of European countries. Unfortunately, the comparison of results obtained with two distinctive approaches is considerably limited. We observe that the cost of children calculated using housing expenses is higher than the one calculated with the subjective method, both in the case of the first and the second child, which is a common result of the two methods' comparison.

It should be noted that the above conclusions were drawn using a commodity-specific equivalence scale rather than overall household equivalence scale. Our results are comparable in terms of main patterns of cost distribution by child order in a family, child's age and type of a household, to the results obtained with the use of original Engel's method, which supports our approach.

However, the share of housing expenses is strongly determined by the ownership structure on the property market of a given country. In the case of countries characterised by highly diversified structure of housing property ownership and a different level of development of property rental and purchase markets the method relying on housing expenses could be more applicable to cost calculation on domestic level rather than to international comparisons.

REFERENCES

BALLI, F., TIEZZI, S., (2010). Equivalence scales, the cost of children and household consumption patterns in Italy, Review of Economics of the Household, 8 (4), pp. 527–549.

BARTEN, A. P., (1964). Family Composition, Prices and Expenditure Patterns. In: P.E. Hart, P. Mills, J.K. Whitaker, ed. 1964, Econometric Analysis for National Economic Planning. Butterworths, London, pp. 277–292.

BETTI, G., LUNDGREN, L., (2012). The impact of remittances and equivalence scales on poverty in Tajikistan, Central Asian Survey, 31(4), pp. 395–408.

BISHOP, J. A., GRODNER, A., LIU, H., AHAMDANECH-ZARCO I., (2014). Subjective Poverty Equivalence Scales for Euro Zone Countries, The Journal of Economic Inequality, 12 (2), pp. 265–278.

BLUNDELL, R., (1998). Equivalence Scales and Household Welfare: What Can Be Learned from Household Budget Data?. In: S.P. Jenkins, A. Kapteyn, B.M.S. van Praag, ed. 1998, The Distribution of Welfare and Household Production: International Perspectives, Cambridge University Press, Cambridge, New York, Melbourne, pp. 364–380.

BROWNING, M., (1992). Children and Household Economic Behaviour, Journal of Economic Literature, 30 (3), pp.1434–1475.

CIECIELĄG, J., (2003). Koszty utrzymania dzieci w Polsce, doctoral thesis, Faculty of Economic Sciences Warsaw University, Warsaw.

DEATON, A. S., MUELLBAUER, J., (1986). On measuring child costs: with applications to poor countries. The Journal of Political Economy, 94(4), pp. 720–744

DUDEK, H., (2009). Statystyczna analiza subiektywnej oceny dochodów gospodarstw domowych rolników. Rocznik Nauk Rolniczych, Seria G, T. 96, z. 4, pp. 41–49.

DUDEK, H., (2011). Skale ekwiwalentności – estymacja na podstawie kompletnych modeli popytu, Wydawnictwo SGGW, Warszawa.

ENGEL, E., (1895). Die Lebenskostenbelgischer Arbeiter-Familien Fruherund jetzt, Bulletin de Institut International de Statistique, 9 (5), pp. 897–930.

EP, (2007). The Cost of Childcare in EU Countries. Part 1 & 2, Brussel.

ESPING-ANDERSEN, G., (1990). The three worlds of welfare capitalism. Princeton University Press, Princeton.

GORMAN, W., (1976). Tricks with Utility Functions. In: M.J. Artis, A.R. Nobay, ed. 1976, Essays in Economic Analysis: Proceedings of the 1975 AUTE Conference, Sheffield, Cambridge: Cambridge University Press.

KALBARCZYK-STĘCLIK, M., MORAWSKI, L., MIŚTA, R., (2017). Subjective Equivalence Scale –Cross-Country and Time Differences, International Journal of Social Economics, in print.

KOT, S., (2014). Inter-temporal equivalence scales based on stochastic indifference criterion: the case of Poland, paper presented to the IARIW 33rd General Conference, Rotterdam, August 2014.

LEWBEL, A., PENDAKUR, K., (2006). Equivalence Scales, Entry for The New Palgrave Dictionary of Economics, 2nd edition, Macmillan Ltd., London.

MUELLBAUER, J., (1974). Household Composition, Engel Curves and Welfare Comparisons Between Households: A Duality Approach, European Economic Review, 5, pp.103–122.

OECD, (2016a). OECD Family Database. OECD, Paris.

OECD, (2016b). Enrolment in childcare and pre-school. OECD, Paris.

PANEK, T., (2011). Skale ekwiwalentności, In: T. Panek, ed. 2011. Ubóstwo, wykluczenie społeczne i nierówności. Teoria i praktyka pomiaru. Warsaw: Oficyna Wydawnicza Szkoła Główna Handlowa, pp. 43–56.

PASHARDES, P., (1991). Contemporaneous and intertemporal child costs: Equivalent expenditure vs. equivalent income scales, Journal of Public Economics 45, pp. 191–213.

POLLAK, R. A., Wales, T. J., (1979). Welfare Comparison and Equivalence Scales, American Economic Review 69, pp. 216–221.

ROTHBARTH, E., (1943). Note on a Method of Determining Equivalent Income for Families of Different Composition, Appendix 4, In: C. Madge, ed. 1943, War-Time Pattern of Saving and Spending, Cambridge: Cambridge University Press, pp. 123–130.

SZULC, A., (2007). Dochód i konsumpcja, In: T. Panek, ed. 2007, Statystyka społeczna, Warszawa: PWE, pp. 131–163.

SZULC, A., (2009). A maching estimator of household equivalence scales, Economics Letters, 103 (2), pp. 81–83.

VAN PRAAG, B. M. S., VAN DER SAR, N. L., (1988). Household Cost Functions and Equivalence Scales, Journal of Human Resources, 23(2), pp. 193–210.

APPENDIX

Table A1. Estimation results of Engel curves for Austria, France, Poland and Italy

Country	Austria		France		Poland		Italy	
Child age limit	18	25	18	25	18	25	18	25
	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.
log income	-0.019**	-0.019**	-0.022**	-0.021**	-0.020**	-0.020**	-0.029**	-0.029**
log hhsize	0.005**	0.004**	0.005**	0.004**	0.005**	0.004**	0.012**	0.013**
share of kids								
aged 0-3	0.002	0.003	0.009**	0.011**	0.000	0.001	-0.006	-0.006
aged 3-6	0.001	0.002	0.008**	0.009**	0.001	0.003	-0.003	-0.003
aged 6-18	0.002	0.003	0.005**	0.006**	0.001	0.002	-0.003	-0.003*
aged 18-25		0.006**		0.004*		0.004**		-0.002
constant	0.211**	0.210**	0.235**	0.234**	0.193**	0.193**	0.302**	0.302**
Number of observations	6,187	6,185	11,029	11,029	12,710	12,710	18,986	18,986
F test	576**	482**	801**	669**	1,013**	846**	1,434**	1,195**

Note: ** - significant at 1%, * - significant at 5%.

Source: Authors' own analysis based on EU-SILC data.

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 701–724, DOI 10.21307/stattrans-2017-008

IMPROVEMENT OF FUZZY MORTALITY MODELS BY MEANS OF ALGEBRAIC METHODS

Andrzej Szymański, Agnieszka Rossa[1]

ABSTRACT

The forecasting of mortality is of fundamental importance in many areas, such as the funding of public and private pensions, the care of the elderly, and the provision of health service. The first studies on mortality models date back to the 19th century, but it was only in the last 30 years that the methodology started to develop at a fast rate. Mortality models presented in the literature form two categories (see, e.g. Tabeau *et al*., 2001, Booth, 2006) consisting of the so-called static or stationary models and dynamic models, respectively. Models contained in the first, bigger group contains models use a real or fuzzy variable function with some estimated parameters to represent death probabilities or specific mortality rates. The dynamic models in the second group express death probabilities or mortality rates by means of the solutions of stochastic differential equations, etc.

The well-known Lee-Carter model (1992), which is widely used today, is considered to belong to the first group, similarly as its fuzzy version published by Koissi and Shapiro (2006). In the paper we propose a new class of fuzzy mortality models based on a fuzzy version of the Lee-Carter model. Theoretical backgrounds are based on the algebraic approach to fuzzy numbers (Ishikawa, 1997a, Kosiński, Prokopowicz and Ślęzak, 2003, Rossa, Socha and Szymański, 2015, Szymański and Rossa, 2014). The essential idea in our approach focuses on representing a membership function of a fuzzy number as an element of C^*-Banach algebra. If the membership function $\mu(z)$ of a fuzzy number is strictly monotonic on two disjoint intervals, then it can be decomposed into strictly decreasing and strictly increasing functions $\Phi(z)$, $\Psi(z)$, and the inverse functions $f(u)=\Phi^{-1}(u)$ and $g(u)=\Psi^{-1}(u)$, $u \in [0, 1]$ can be found.

Ishikawa (1997a) proposed foundations of the fuzzy measurement theory, which is a general measurement theory for classical and quantum systems. We have applied this approach, termed C^*-measurement, as the theoretical foundation of the mortality model. Ishikawa (1997b) introduced also the notions of objective and subjective C^*-measurement called real and imaginary C^*-measurements. In our proposal of the mortality model the function f is treated as an objective C^*-measurement and the function g as an subjective C^*-measurement, and the

[1] Institute of Statistics and Demography, University of Łódź, e-mail: agrossa@uni.lodz.pl

membership function $\mu(z)$ is represented by means of a complex-valued function $f(u) + ig(u)$, where i is the imaginary unit. We use the Hilbert space of quaternion algebra as an introduction to the mortality models.

Key words: C*-Banach algebra, non-commutative C*-algebra, quaternion algebra, fuzzy mortality model.

1. Introduction

Long-lasting observations of mortality rates or death probabilities lead to the conclusion that in developed countries they decline for most age groups, whereas the upper limit of human lifetime is moving upwards. Other life-table parameters also change in time. The mortality trends and patterns observed in developed countries in the second half of the 20th century can be summed up as follows (see also Wilmoth and Horiuchi, 1999):

- the normal lifetime drifts toward older ages,
- ages at deaths are concentrating around the normal lifetime,
- the survival curve is undergoing rectangularization (because of the aforementioned trends),
- the life expectancy is increasing,
- in the young population (especially among young males aged 20+), the number and percentage of deaths from external causes (injuries, accidents, poisoning) is rising.

These measures are therefore not constant in time. They are rather functions of time or, in broader terms, stochastic processes showing some variability. Past works on this subject have used, for instance, time-series analysis tools to examine the stochastic nature of these processes. One of the most popular is the Lee-Carter mortality model (Lee and Carter, 1992).

2. The Lee-Carter mortality model

Let $m_x(t)$ denote an age-specific (central) death rate for the subset of a population that is between exact ages x and x+1

$$m_x(t) = \frac{D_x(t)}{L_x(t)}, \qquad x=0,1,2,\ldots,X, \quad t=1,2,\ldots,T, \tag{2.1}$$

where

$D_x(t)$ – the number of deaths at age x in the year t,

$L_x(t)$ – the midyear population at the age x in the year t,

$x=0,1,\ldots,X$ – index of one-year age groups,

$t=1,2,\ldots,T$ – years of observation period.

The measure $m_x(t)$ is the ratio of deaths between ages x and $x + 1$ to the midyear population alive at age x in the given year t, also referred to as the mean population in the year t. The measure is described as the central rate because the midyear population is used in the denominator.
The Lee-Carter model can be written as

$$\ln m_x(t) = \alpha_x + \beta_x \kappa_t + \epsilon_{xt}, \qquad x=0,1,\ldots,X,\ t=1,2,\ldots,T \tag{2.2}$$

or, equivalently, as

$$m_x(t) = \exp\{\alpha_x + \beta_x \kappa_t + \epsilon_{xt}\}, \qquad x=0,1,\ldots,X,\ t=1,2,\ldots,T, \tag{2.3}$$

where $m_x(t)$, $t \in \mathbf{N}$ are age-specific mortality rates, α_x, β_x and κ_t are the model parameters, of which α_x, β_x depend on age x and κ_t on time t. The double-indexed terms $\varepsilon_{x,t}$ are error terms, which are assumed to be independent and to have the same normal distributions with an expected value of 0 and constant variance.

The parameters α_x, $x=0,1,\ldots,X$ indicate the general shape of the mortality schedule, the time-varying parameters κ_t, $t=1,2,\ldots,T$ represent the time-trend indices of the general mortality level, whereas β_x indicate the pattern of deviations from the age profile when the general level of mortality κ_t changes. In general, β_x could be negative at some ages, indicating that mortality rates at those ages tend to rise when falling at other ages. In other words, the shape of β_x profile tells which rates decline rapidly and which slowly over time in response to change of κ_t.

Because of the form of (2.2), the Lee-Carter model is called a bilinear model. The system of equations (2.2) or (2.3) cannot be explicitly solved unless additional restrictions are imposed. Let us assume, for instance, that for a set of parameters $\{\alpha_x\}$,$\{\beta_x\}$, and $\{\kappa_t\}$ the model (2.2) is valid. It is easy to see that the model holds true also for any constant c and parameters $\{\alpha_x - c\beta_x\}$, $\{\beta_x\}$, $\{\kappa_t + c\}$ or $\{\alpha_x\}$, $\{c\beta_x\}$, $\{\kappa_t/c\}$.

To make sure that an unambiguous solution is obtained, some additional restrictions must be defined. To this end, it is assumed that the sum of parameters β_x over age index x is 1 and the sum of parameters κ_t over time index t is equal to 0, i.e.

$$\sum_{x=0}^{X} \beta_x = 1, \ \sum_{t=1}^{T} \kappa_t = 0. \tag{2.4}$$

Parameters α_x and β_x do not depend on time t, which means that once they have been established they can also be used for the future period, i.e. $t > T$. The time-varying rates are κ_t. They can be further modelled using, for instance, the time series analysis methods.

Lee and Carter (1992) proposed a random walk model, but the range of proposals discussed in the literature is wider. A random walk process with a drift is given by the formula

$$\kappa_t = \delta + \kappa_{t-1} + \xi_t, \tag{2.5}$$

where δ is a constant (a drift), and ξ_t is a random term.

Parameter δ in (2.5) mostly takes negative values that point to declining mortality. Random fluctuations around this trend are represented by independent random terms ξ_t, each having a normal distribution with the expected value of 0 and finite variance.

With the values of κ_t predicted from (2.5) and the estimations of α_x and β_x the partial death rates can be forecasted, as well as other life-table mortality rates.

The method of parameter estimation proposed by Lee and Carter is based on the method of Singular Value Decomposition (SVD), which decomposes a data matrix $\mathbf{M} = [\ln m_x(t) - a_x]$ into a matrix of singular values **D** and two matrices **W** and **V** of left and right singular vectors.

Let a_x, b_x, k_t represent the estimators of parameters α_x, β_x, κ_t. Assuming that random terms ϵ_{xt} in model (2.2) have an expected value of 0, we have

$$\mathrm{E}(\epsilon_{xt}) = 0. \tag{2.6}$$

This property will be used to find a_x. To this end, we will determine the analogous first row moment from the sample, i.e. from time series $\{\ln m_x(t),\ t = 1,2,...,T\}$ for $x = 0,1,2,...,X$ we calculate the sum

$$\sum_{t=1}^{T}[\ln m_x(t) - (a_x + b_x k_t)], \tag{2.7}$$

then by comparing (2.7) with 0

$$\sum_{t=1}^{T}[\ln m_x(t) - (a_x + b_x k_t)] = 0, \tag{2.8}$$

we obtain the following equality

$$T a_x + b_x \sum_{t=1}^{T} k_t = \sum_{t=1}^{T} \ln m_x(t). \tag{2.9}$$

By allowing additionally for condition $\sum_{t=1}^{T} k_t = 0$, we arrive at

$$a_x = \frac{1}{T}\sum_{t=1}^{T} \ln m_x(t)\,. \tag{2.10}$$

To estimate β_x, κ_t the first singular value and the first vector of matrices **W** and **V** are used. For a general case, all singular values and singular vectors can be employed, which gives the following extension of the model (2.2)

$$\ln m_x(t) = \alpha_x + \sum_{i=1}^{r} \beta_x^{(i)} \kappa_t^{(i)}, \qquad x=0,1,\ldots,X,\ t=1,2,\ldots,T, \tag{2.11}$$

where r is the number of non-zero singular values.

3. The Koissi-Shapiro model

One of the most interesting generalisations of the Lee-Carter model, referring to the algebra of fuzzy numbers, was proposed by Koissi and Shapiro (2006). Their version of the Lee-Carter model (FLC model) assumes a fuzzy representation of the central death rates. It allows taking account of uncertainty involved in mortality rates and entering a random term into the fuzzy structure of the model.

Their approach builds on the assumption that the real rates of mortality are not exactly $m_x(t)$, but rather around $m_x(t)$, thus the role of the explanatory variable is played by fuzzified mortality rates.

It is well-known that death statistics are subject to reporting errors of several kinds. They may be reported for incorrect year, area, or assigned statistics that are incorrect, e.g. age. Moreover, the midyear population data that serve as the denominators of mortality rates are also the subject of errors. It is regarded as the population at July 1 and is assumed to be the point at which half of the deaths in the population during the year have occurred. Such an estimate can be actually underestimated or overestimated. For these reasons, fuzzy representation of the central death rates seems to be justified.

Koissi and Shapiro proposed fuzzy representation of the logarithms of age-specific mortality rates $\ln m_x(t)$, by converting them into symmetric, triangular fuzzy numbers (basic notions of the fuzzy numbers are given in Rossa, Socha, Szymański (2015, appendix) presented as

$$Y_{xt} = (y_{xy}, e_{xt}), \quad x = 0,1,...,X, \quad t = 1,2,...,T, \tag{3.1}$$

where $y_{xt} = \ln m_x(t)$ and e_{xt} are the spreads of the membership functions of triangular fuzzy numbers.

In fuzzification approach, a fuzzy least-squares regression based on minimum fuzziness criterion was employed, and – for simplicity – triangular symmetric fuzzy numbers were considered.

Given the log-central death rates $y_{xt} = \ln m_x(t)$ for age x in year t, the task is to find symmetric triangular fuzzy numbers $A_0 = (c_{0x}, s_{0x})$, $A_1 = (c_{1x}, s_{1x})$ and $Y_{xt} = (y_{xt}, e_{xt})$ with centers c_{0x}, c_{1x}, y_{xt} and spreads s_{0x}, s_{1x}, e_{xt} such that

$$(y_{xt}, e_{xt}) = (c_{0x}, s_{0x}) + (c_{1x}, s_{1x}) \times t . \tag{3.2}$$

To find the fuzzy numbers A_0 and A_1, the approach is as follows:

1. First, ordinary least-squares (OLS) regression is used to find the center values c_{0x} and c_{1x} such that

$$y_{xt} = c_{ox} + c_{1x}t + \varepsilon_{xt}, \quad \text{for each } x, \tag{3.3}$$

where $y_{xt} = \ln m_x(t)$ are the observed log-central death rates, t is time variable, and ε_{xt} represent random terms.

2. The spreads (s_{0x} and s_{1x}) are obtained by using the minimum fuzziness criterion. This consists in minimizing the following optimization problem, which can be solved through standard optimization software, i.e. minimize

$$T s_{0x} + s_{1x} \sum_{t=1}^{T} t \tag{3.4}$$

subject to

$$\forall_t \ \ s_{0x}, s_{1x} \geq 0$$

$$c_{0x} + c_{1x}t + (s_{0x} + s_{1x}t) \geq y_{xt}, \quad \text{and} \quad c_{0x} + c_{1x}t - (s_{0x} + s_{1x}t) \leq y_{xt} .$$

Once the log-central death rates are fuzzified, the FLC model can be defined as

$$Y_{xt} = A_x \oplus_{Tw} (B_x \otimes_{Tw} K_t), \qquad x = 0,1,...,X, \; t = 1,2,...,T, \tag{3.5}$$

where Y_{xt} are known fuzzy log-central mortality rates, A_x, B_x, K_t are unknown parameters, and $\oplus_{Tw}$, $\otimes_{Tw}$ are the addition and multiplication operators of fuzzy numbers in the norm T_w, respectively. For the definition of the norm T_w see Koissi and Shapiro (2006).

The authors assumed that the model parameters can be estimated by minimizing the criterion function based on the Diamond distance measure between fuzzy variables. The criterion can be expressed as the following sum

$$\sum_{x=0}^{X}\sum_{t=1}^{T}[3a_x^2 + 3b_x^2k_t^2 + 3y_{xt}^2 + 6a_xb_xk_t - 4y_{xt}(a_x + b_xk_t) + 2e_{xt}^2] + \tag{3.6}$$

$$+2\sum_{x=0}^{X}\sum_{t=1}^{T}\left[\left(\max\{s_{A_x}, |b_x|s_{K_t}, |k_t|s_{B_x}\}\right)^2 - 2e_{xt}\max\{s_{A_x}, |b_x|s_{K_t}, |k_t|s_{B_x}\}\right].$$

However, the FLC model poses major problems in the estimation algorithm, because expression $\max\{s_{A_x}, |b_x|s_{K_t}, |k_t|s_{B_x}\}$ in the criterion (3.6) prevents the standard use of non-linear optimization methods.

In the rest of the paper, modification to the fuzzy mortality model based on fuzzified mortality rates with exponential membership functions will be proposed. The model simplifies both operations on fuzzy numbers and the model estimation. The essential idea in this approach is representing the membership functions of fuzzy numbers as elements of C*-Banach algebra.

4. A new class of mortality models based on algebraic approach to fuzzy numbers

4.1. The theoretical background for the new mortality model

Fuzzification of data depends on the assumption about membership functions of fuzzy numbers. Koissi and Shapiro (2006) adopted triangular symmetric membership functions and used fuzzy least-squares regression. In our approach, we will assume exponential membership functions derived from relative frequencies of residuals in the least-squares regression model.

Suppose that the membership function $\mu(z)$ of a fuzzy number is strictly monotonic on two disjoint intervals. Following Nasibov and Peker (2011), we will consider an exponential membership function of the form

$$\mu(z) = \begin{cases} \exp\left\{-\left(\frac{c-z}{\tau}\right)^2\right\}, & \text{for } z \le c, \\ \exp\left\{-\left(\frac{z-c}{\nu}\right)^2\right\}, & \text{for } z > c, \end{cases} \tag{4.1}$$

where c, τ, ν are scalars.

Note that we can decompose $\mu(z)$ into two parts – strictly increasing and strictly decreasing functions $\Psi(z)$ and $\Phi(z)$ of the form

$$\Psi(z) = \exp\left\{-\left(\frac{c-z}{\tau}\right)^2\right\}, \quad \text{for } z \le c,$$
$$\Phi(z) = \exp\left\{-\left(\frac{z-c}{\nu}\right)^2\right\}, \quad \text{for } z > c. \tag{4.2}$$

Then, there exist inverse functions

$$\Psi^{-1}(u) = c + \psi(u), \quad \Phi^{-1}(u) = c + \varphi(u), \quad u \in [0,1], \tag{4.3}$$

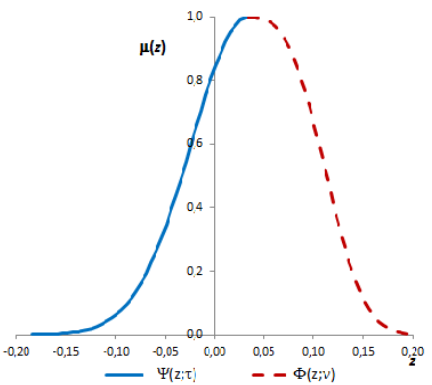
where $\psi(u)$ and $\varphi(u)$ are expressed as follows

$$\psi(u) = -\tau(-\ln u)^{\frac{1}{2}}, \quad \varphi(u) = \nu(-\ln u)^{\frac{1}{2}}, \quad u \in [0,1]. \tag{4.4}$$

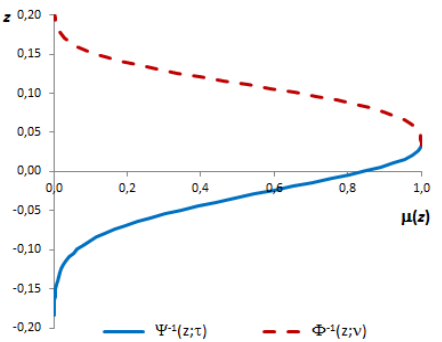
Example 1. Figure 1(a) illustrates an exponential functions (4.2) for fixed values of parameters c=0.03, $\tau = 0.08, \nu = 0.09$, Figure 1(b) presents respective inverse functions (4.3).

Figure 1. An example of an exponential membership function, c=0.03, τ =0.08, ν =0.09

(a)



(b)



Source: developed by the authors.

4.2. Transformation of membership functions into complex-valued functions

Let us consider the complex functions

$$f(u) = c + i\psi(u), \quad \text{and} \quad g(u) = c + i\varphi(u), \quad u \in [0,1], \tag{4.5}$$

where $\mathrm{i} = \sqrt{-1}$ is an imaginary unit.

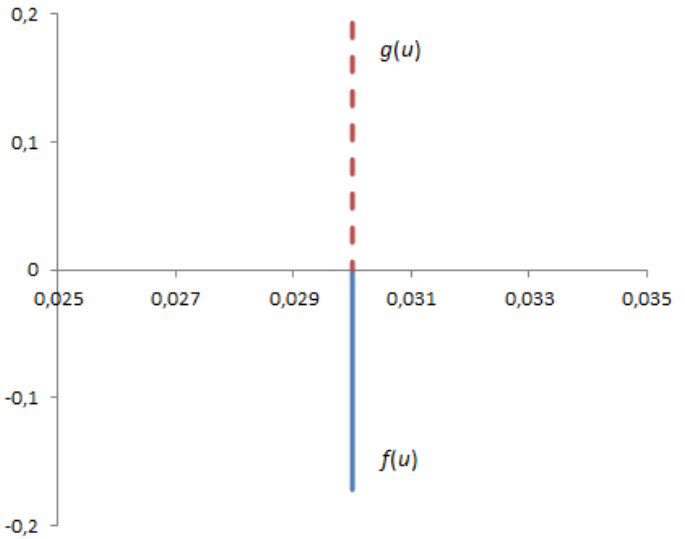
Assuming that functions $\psi(u)$ and $\varphi(u)$ are expressed as in (4.4) we get

$$f(u) = c - i\tau(-\ln u)^{\frac{1}{2}}, \quad \text{and} \quad g(u) = c + i\nu(-\ln u)^{\frac{1}{2}}, \quad u \in [0,1]. \tag{4.6}$$

The pair of two complex functions $(f(u), g(u))$ is called a quaternion.

An illustration of a quaternion $(f(u), g(u))$ on the complex plane for c=0.03, τ=0.08, ν=0.09 is presented in Figure 2.

Figure 2. A quaternion $(f(u), g(u))$, with $f(u)$ and $g(u)$ defined in (4.4) with $\tau = 0.08, \nu = 0.09$



Source: developed by the authors.

The modules of $f(u)$ and $g(u)$ are as follows

$$|f(u)|^2 = c^2 + \tau^2(-\ln u), \ u \in [0,1], \tag{4.7}$$

$$|g(u)|^2 = c^2 + \nu^2(-\ln u), \ u \in [0,1]. \tag{4.8}$$

After integrating both sides of (4.7) and (4.8) on the interval [0,1] we obtain

$$\int_0^1 |f(u)|^2 du = c^2 + \tau^2 \int_0^1 (-\ln u)\, du = c^2 + \tau^2, \tag{4.9}$$

$$\int_0^1 |g(u)|^2 du = c^2 + \nu^2 \int_0^1 (-\ln u)\, du = c^2 + \nu^2. \tag{4.10}$$

4.3. Basic properties of quaternions

It is well known that the complex numbers could be viewed as ordered pairs of real numbers. By analogy, the quaternions can be treated as ordered pairs of complex functions

$$(z, w), \quad \text{where} \quad z = a + ib, \quad w = c + id \quad \text{and} \quad i = \sqrt{-1}. \tag{4.11}$$

The algebra of quaternions is often denoted by $\boldsymbol{H}$. Quaternions were first described by Irish mathematician William Hamilton in 1843. The space $\boldsymbol{H}$ is equipped with three operations: addition, scalar multiplication and quaternion multiplication.

The sum of two elements of $\boldsymbol{H}$ is defined as the sum of their components. Therefore, we have

$$(z, w) + (u, x) = (z + u, w + x). \tag{4.12}$$

The product of an element of $\boldsymbol{H}$ by a real number $\alpha \in R$ is defined to be the same as the product by scalar of both components

$$\alpha(z, w) = (\alpha z, \alpha w). \tag{4.13}$$

To define the product of two elements in $\boldsymbol{H}$ a choice of the basis for $\boldsymbol{R^4}$ is needed. The elements of this basis are customarily denoted as $1, i, j$ and k. Each element of $\boldsymbol{H}$ can be uniquely denoted as a linear combination $a \cdot 1 + bi + cj + dk$, where a, b, c, d are real numbers.

The basis element 1 could be viewed as the identity element of $\boldsymbol{H}$. It means that multiplication by 1 does not change the value, and elements of $\boldsymbol{H}$ can be uniquely denoted as

$$(z, w) = a + bi + cj + dk, \tag{4.14}$$

where a, b, c, d are real numbers. Therefore, each element of $\boldsymbol{H}$ is determined by four numbers and hence the term "quaternion".

The possible products of basic elements i, j, k can be described as follows

$$i^2 = j^2 = k^2 = ijk = -1, \tag{4.15}$$

$$ij = k, \; ji = -k, \tag{4.16}$$

$$jk = i, \; kj = -i, \tag{4.17}$$

$$ki = j, \; ik = -j. \tag{4.18}$$

Quaternions can be represented as pairs of complex numbers as a generalization of the construction of the complex numbers being pairs of real numbers.

Let C be a two-dimensional vector space over the complex numbers. Let us choose a basis consisting of two elements 1 and j. For $z, w \in C$ of the form $z = a + bi$ and $w = c + di$, we can write

$$q = z + wj = (a + bi) + (c + di)j = a + bi + cj + dij. \tag{4.19}$$

If we denote $k = ij$ then

$$q = z + wj = a + bi + cj + dk. \tag{4.20}$$

Thus, the vector (z, w) corresponds to a quaternion $q = a + bi + cj + dk$. Then, each quaternion $q \in \boldsymbol{H}$ is uniquely represented by

$$q = z + wj. \tag{4.21}$$

Multiplication of quaternions could be defined in the form

$$(z, w)(u, x) = (zu - w\bar{x}, zx + w\bar{u}), \tag{4.22}$$

where $\bar{x}, \bar{u}$ denote conjugations of x and u.

Multiplication of quaternions is associative and distributive with respect to addition, however it is not commutative, since, for example, we have

$$(i, 0)(0,1) = (0, i), \tag{4.23}$$

but

$$(0,1)(i, 0) = (0, -i). \tag{4.24}$$

Let us denote

$$q^* = z - wj \tag{4.25}$$

as the conjugate of q.

Conjugation is an involution. It means that for $p, q \in \boldsymbol{H}$ we have

$$(q^*)^* = q, \quad (pq)^* = q^* p^*, \quad (p + q)^* = p^* + q^*. \tag{4.26}$$

The square root of the product of a quaternion with its conjugate is called a norm, and is denoted $\|q\|$. This is expressed as follows

$$\|q\| = \sqrt{qq^*} = \sqrt{q^* q} = \sqrt{a^2 + b^2 + c^2 + d^2}. \tag{4.27}$$

It is always a non-negative real number, and it is the same as the Euclidean norm on $\boldsymbol{H}$ considered as the vector space $\boldsymbol{R}^4$. Multiplying a quaternion by a real number scales its norm by the absolute value of this number

$$\|\alpha q\| = |\alpha| \|q\|. \tag{4.28}$$

This is a special case of the following property

$$\|pq\| = \|p\| \|q\| \tag{4.29}$$

for any two quaternions p and q.

The norm (4.27) allows us to define the distance $d(p,q)$ between p and q as the norm of their difference

$$d(p,q) = \|p - q\|. \tag{4.30}$$

This defines $\boldsymbol{H}$ as a metric space.

According to (4.6) we have

$$f(u) = c + i\psi(u), \quad u \in [0,1],$$

and

$$g(u) = c + i\varphi(u), \quad u \in [0,1],$$

where ψ, φ are defined in (4.4).

Hence,

$$|f(u)|^2 = c^2 + \psi^2(u), \quad \text{and} \quad |g(u)|^2 = c^2 + \varphi^2(u).$$

Let us denote

$$P(u) = \big(f(u), g(u)\big), \quad u \in [0,1]. \tag{4.31}$$

The function P is a quaternion-valued function. The norm of $P(u)$ could be expressed as follows

$$\|P(u)\|^2 = |f(u)|^2 + |g(u)|^2 = c^2 + \psi^2(u) + c^2 + \varphi^2(u), \tag{4.32}$$

and from (4.9) and (4.10) we have

$$\int_0^1 |f(u)|^2 du < \infty \quad and \quad \int_0^1 |g(u)|^2 du < \infty.$$

Integrating both sides in (4.32) we receive also

$$\int_0^1 \|P(u)\|^2 du = \int_0^1 |f(u)|^2 du + \int_0^1 |g(u)|^2 \, du < \infty. \tag{4.33}$$

Thus, the functions f and g are the elements of the Hilbert space $L_2[0,1]$, and the quaternion-valued function P is integrable with squared norm on the interval [0,1]. Let us denote the space of such functions as $L_2(\boldsymbol{H})$.

5. A mortality model based on quaternion-valued functions

5.1. Formulation of the model

We will assume that $\tilde{Y}_{x,t} = \big(f_{Y_{x,t}}, g_{Y_{x,t}}\big)$ are quaternions with complex functions $f_{Y_{x,t}}, g_{Y_{x,t}}$ of the form

$$f_{Y_{x,t}}(u) = y_{xt} - i\tau_x(-\ln u)^{\frac{1}{2}}, \qquad g_{Y_{x,t}}(u) = y_{xt} + iv_x(-\ln u)^{\frac{1}{2}}, \qquad u \in [0,1],$$

where i is an imaginary unit, $y_{xt} = \ln m_x(t)$, and τ_x, υ_x are known parameters evaluated by means of Nasibov-Peker method (see section 5.3 for more details).

Similarly, we will assume that $\tilde{A}_x = (f_{A_x}, g_{A_x})$, $\widetilde{K}_t = (f_{K_t}, g_{K_t})$ are quaternions determined by complex functions

$$f_{A_x}(u) = a_x - i(-\ln u)^{\frac{1}{2}} s_{A_x}^L, \quad g_{A_x}(u) = a_x + i(-\ln u)^{\frac{1}{2}} s_{A_x}^R, \quad u \in [0,1] \tag{5.1}$$

$$f_{K_t}(u) = k_t - i(-\ln u)^{\frac{1}{2}} s_{K_t}, \quad g_{K_t}(u) = k_t + i(-\ln u)^{\frac{1}{2}} s_{K_t}, \quad u \in [0,1]. \tag{5.2}$$

As in other models based on functional analysis, we postulate the following mortality model based on quaternion-valued functions

$$\tilde{Y}_{x,t} = \tilde{A}_x + b_x \widetilde{K}_t, \qquad x = 0,1,\dots,X,\ t = 1,2,\dots,T, \tag{5.3}$$

where $Y_{x,t}$ are fuzzified log-central mortality rates expressed in terms of quaternion-valued functions in the Hilbert space $L_2(\boldsymbol{H})$, $b_x \in \boldsymbol{R}, x = 0,1,\dots,X$, is a set of unknown scalar parameters, and quaternions $\tilde{A}_x, \widetilde{K}_t$ represent unknown parameters in $L_2(\boldsymbol{H})$ determined by the complex functions (5.1) and (5.2). The proposed model (5.3) will be termed Complex Number Mortality Model (CNMM).

Note that the quaternions $\tilde{A}_x = (f_{A_x}, g_{A_x})$, $\widetilde{K}_t = (f_{K_t}, g_{K_t})$ on the right-hand side of (5.3) reflect fuzzy numbers A_x, K_t with exponential membership functions $\mu_{A_x}(z)$ and $\mu_{K_t}(z)$ (see sections 4.1 and 4.2)

$$\mu_{A_x}(z) = \begin{cases} \exp\left\{-\left(\frac{a_x - z}{s_{A_x}^L}\right)^2\right\}, & \text{for } z \le a_x, \\ \exp\left\{-\left(\frac{z - a_x}{s_{A_x}^R}\right)^2\right\}, & \text{for } z > a_x, \end{cases} \tag{5.4}$$

$$\mu_{K_t}(z) = \begin{cases} \exp\left\{-\left(\frac{k_t - z}{s_{K_t}}\right)^2\right\}, & \text{for } z \le k_t, \\ \exp\left\{-\left(\frac{z - k_t}{s_{K_t}}\right)^2\right\}, & \text{for } z > k_t. \end{cases} \tag{5.5}$$

Using the properties (4.12) and (4.13) the complex functions defining the quaternion $\tilde{A}_x + b_x \widetilde{K}_t$ on the right-hand side of (5.3) are as follows

$$f_{A_x + b_x K_t}(u) = a_x + b_x k_t - i(-\ln u)^{\frac{1}{2}} \left(s_{A_x}^L + b_x s_{K_t}\right), \ u \in [0,1], \tag{5.6}$$

$$g_{A_x + b_x K_t}(u) = a_x + b_x k_t + i(-\ln u)^{\frac{1}{2}} \left(s_{A_x}^R + b_x s_{K_t}\right), \ u \in [0,1]. \tag{5.7}$$

It means that $\tilde{A}_x + b_x\tilde{K}_t$ reflects a fuzzy number W_{xt} with an exponential membership function

$$\mu_{W_{xt}}(z) = \begin{cases} \exp\left\{-\left(\frac{a_x+b_xk_t-z}{s^L_{A_x}+b_xs_{K_t}}\right)^2\right\}, & \text{for } z \le a_x + b_xk_t, \\ \exp\left\{-\left(\frac{z-a_x-b_xk_t}{s^R_{A_x}+b_xs_{K_t}}\right)^2\right\}, & \text{for } z > a_x + b_xk_t. \end{cases} \tag{5.8}$$

5.2. Estimation of the model parameters

In order to estimate the parameters $a_x, b_x, k_t, s^L_{A_x}, s^R_{A_x}, s_{K_t}$ we will use the notion of the norm (4.32) defined in the space of quaternion-valued functions. Thus, the following distance between left- and right-hand sides of the model (5.3) will be defined for fixed x and t

$$d_{x,t} = \int_0^1 \left\|\tilde{Y}_{x,t} - \left(\tilde{A}_x + b_x\tilde{K}_t\right)\right\|^2 du = \int_0^1 \left|f_{Y_{x,t}-(A_x+b_xK_t)}(u)\right|^2 du + \int_0^1 \left|g_{Y_{x,t}-(A_x+b_xK_t)}(u)\right|^2 du.$$

Let us find functions $f_{Y_{x,t}-(A_x+b_xK_t)}(u)$ and $g_{Y_{x,t}-(A_x+b_xK_t)}(u)$ determining the difference of quaternions $\tilde{Y}_{x,t} - \left(\tilde{A}_x + b_x\tilde{K}_t\right)$. We have

$$f_{Y_{x,t}-(A_x+b_xK_t)}(u) = y_{x,t}-(a_x + b_xk_t) - i(-\ln u)^{\frac{1}{2}}\left(\tau_x - s^L_{A_x} - b_xs_{K_t}\right), \tag{5.9}$$

$$g_{Y_{x,t}-(A_x+b_xK_t)}(u) = y_{x,t}-(a_x + b_xk_t) + i(-\ln u)^{\frac{1}{2}}\left(\nu_x - s^R_{A_x} - b_xs_{K_t}\right). \tag{5.10}$$

Hence,

$$\left|f_{Y_{x,t}-(A_x+b_xK_t)}(u)\right|^2 = \left(y_{x,t}-(a_x + b_xk_t)\right)^2 + (-\ln u)\left(\tau_x - s^L_{A_x} - b_xs_{K_t}\right)^2, \tag{5.11}$$

$$\left|g_{Y_{x,t}-(A_x+b_xK_t)}(u)\right|^2 = \left(y_{x,t}-(a_x + b_xk_t)\right)^2 + (-\ln u)\left(\nu_x - s^R_{A_x} - b_xs_{K_t}\right)^2. \tag{5.12}$$

Integrating (5.11) and (5.12) on the interval [0,1] we receive

$$d_{x,t} = \int_0^1 \left\|\tilde{Y}_{x,t} - \left(\tilde{A}_x + b_x\tilde{K}_t\right)\right\|^2 du = 2\left(y_{x,t}-(a_x + b_xk_t)\right)^2 + \left(\tau_x - s^L_{A_x} - b_xs_{K_t}\right)^2 + \left(\nu_x - s^R_{A_x} - b_xs_{K_t}\right)^2. \tag{5.13}$$

By the analogy to the Lee-Carter model and restrictions (2.4) we will assume that

$$\sum_{t=1}^{T} k_t = 0,\ \sum_{x=0}^{X} b_x = 1. \tag{5.14}$$

An additional restriction will be also imposed on the sum of s_{K_t}

$$\sum_{t=1}^{T} s_{K_t} = (X+1)\sqrt{\sum_{t=1}^{T}(\bar{y}_t - \bar{y})^2}, \tag{5.15}$$

where $\bar{y}_t = \frac{1}{X+1}\sum_{x=0}^{X} y_{xt}$ and $\bar{y} = \frac{1}{T(X+1)}\sum_{t=1}^{T}\sum_{x=0}^{X} y_{xt}$.

Thus, the criterion used to estimate model parameters takes the form

$$\begin{aligned} &F\left(a_x, b_x, k_t, s_{A_x}^L, s_{A_x}^R, s_{K_t}, \lambda_1, \lambda_2, \lambda_3\right) = \\ &\sum_{x=0}^{X}\sum_{t=1}^{T} d_{x,t} + \lambda_1\left(\sum_{x=0}^{X} b_x - 1\right) + \lambda_2 \sum_{t=1}^{T} k_t + \lambda_3 \left(\sum_{t=1}^{T} s_{K_t} - \right. \\ &\left.(X+1)\sqrt{\sum_{t=1}^{T}(\bar{y}_t - \bar{y})^2}\right), \end{aligned} \tag{5.16}$$

where $\lambda_1, \lambda_2, \lambda_3$ represent Lagrange multipliers.

To minimize (5.16) it is necessary to compute its first derivatives with respect to $a_x, b_x, k_t, s_{A_x}^L, s_{A_x}^R, s_{K_t}, \lambda_1, \lambda_2, \lambda_3$. We have

$$\begin{cases} \frac{\partial F}{\partial a_x} = -4\sum_{t=1}^{T}(y_{xt} - a_x - b_x k_t), \\ \frac{\partial F}{\partial b_x} = -2\sum_{t=1}^{T}\left[2k_t(y_{xt} - a_x - b_x k_t) + s_{K_t}\left(\tau_x + \nu_x - s_{A_x}^L - s_{A_x}^R - 2b_x s_{K_t}\right)\right] + \lambda_1 \\ \frac{\partial F}{\partial k_t} = -4\sum_{x=0}^{X} b_x(y_{xt} - a_x - b_x k_t) + \lambda_2 \\ \frac{\partial F}{\partial s_{A_x}^L} = -2\sum_{t=1}^{T}\left(\tau_x - s_{A_x}^L - b_x s_{K_t}\right) \\ \frac{\partial F}{\partial s_{A_x}^R} = -2\sum_{t=1}^{T}\left(\nu_x - s_{A_x}^R - b_x s_{K_t}\right) \\ \frac{\partial F}{\partial s_{K_t}} = -2\sum_{x=0}^{X} b_x\left(\tau_x + \nu_x - s_{A_x}^L - s_{A_x}^R - 2b_x s_{K_t}\right) + \lambda_3 \\ \frac{\partial F}{\partial \lambda_1} = \sum_{x=1}^{X} b_x - 1 \\ \frac{\partial F}{\partial \lambda_2} = \sum_{t=1}^{T} k_t \\ \frac{\partial F}{\partial \lambda_3} = \sum_{t=1}^{T} s_{K_t} - (X+1)\sqrt{\sum_{t=1}^{T}(\bar{y}_t - \bar{y})^2} \end{cases} \tag{5.17}$$

Then, setting each derivative in (5.17) equal to zero and solving for required parameters yields the set of normal equations

$$\begin{cases} a_x = \frac{1}{T}\sum_{t=1}^{T} y_{xt} = \bar{y}_x \\ b_x = \frac{\sum_{t=1}^{T}[2k_t(y_{xt}-a_x)+s_{K_t}(\tau_x+\nu_x-s_{A_x}^L-s_{A_x}^R)]-\frac{\lambda_1}{2}}{2\sum_{t=1}^{T}\left(k_t^2+s_{K_t}^2\right)} \\ k_t = \frac{\sum_{x=0}^{X} b_x(y_{xt}-a_x)-\frac{\lambda_2}{4}}{\sum_{x=0}^{X} b_x^2} \\ s_{A_x}^L = \tau_x - \frac{1}{T} b_x \sum_{t=1}^{T} s_{K_t} \\ s_{A_x}^R = \nu_x - \frac{1}{T} b_x \sum_{t=1}^{T} s_{K_t} \\ s_{K_t} = \frac{\sum_{x=0}^{X} b_x(\tau_x+\nu_x-s_{A_x}^L-s_{A_x}^R)-\frac{\lambda_3}{2}}{2\sum_{x=0}^{X} b_x^2} \\ \sum_{x=1}^{X} b_x = 1 \\ \sum_{t=1}^{T} k_t = 0 \\ \sum_{t=1}^{T} s_{K_t} - (X+1)\sqrt{\sum_{t=1}^{T}(\bar{y}_t - \bar{y})^2} = 0 \end{cases} \tag{5.18}$$

Note that the last three equations in (5.18) satisfy restrictions (5.14) and (5.15).

This set of normal equations can be solved numerically by means of an iterative procedure. After choosing a set of starting values, equations are computed sequentially using the most recent set of parameter estimates obtained from the right-hand side of each equation. In addition to numerical solution of the normal equations, there are also other minimizing algorithms, e.g. computer routines available in several mathematical packages (e.g. quasi-Newton or simplex methods).

Prediction of the log-central death rates with the CNMM can be performed in three steps. First, the random-walk model with a drift (2.5) should be used to predict time parameters k_t for future periods $t>T$. Next, functions (5.6) and (5.7) should be determined using estimated parameters $a_x, b_x, s_{A_x}^L, s_{A_x}^R, s_{K_t}$ and the sequence of predicted time indices $k_t, t > T$. Note that the functions (5.6) and (5.7) define the right-hand side of the mortality model (5.3) for $t > T$, i.e. they define quaternions $\tilde{A}_x + b_x\tilde{K}_t$ for future periods. Finally, these quaternions $\tilde{A}_x + b_x\tilde{K}_t$ can be transformed into fuzzy numbers W_{xt} using exponential membership function $\mu_{W_{xt}}(z)$ given in (5.8). They also can be further defuzzified into crisp numbers w_{xt}, if necessary, i.e. by means of the centroid defuzzification method

$$w_{xt} = \frac{\sum_{z=\epsilon}^{1} z\mu_{W_{xt}}(z)}{\sum_{z=\epsilon}^{1} \mu_{W_{xt}}(z)}, \tag{5.19}$$

where $\epsilon > 0$ denotes a small positive number.

The crisp values w_{xt} represent predicted fuzzy log-central death rates for $t>T$, whereas W_{xt} are their fuzzy counterparts.

5.3. Fuzzification of log-central death rates

Fuzzification of the log-central death rates $y_{xt} = \ln m_x(t)$ for x=0,1,...,X, t = 1,2,...,T by means of exponential membership functions (4.1) will be based on the method proposed by Nasibov and Peker (2011), which allows us to determine parameters τ_x, ν_x for a fixed x based on an empirical distribution of a sequence of data. The main results of their work are introduced in this section.

Assume that $\{r_t, t = 1,2,\dots,T\}$ is a sequence of T observations in a data set. Assume that observation are grouped into a frequency table with k mutually exclusive class intervals (Table 1).

Table 1. Frequency table

Class intervals	Midpoints z_i	Frequencies f_i	Relative frequencies p_i
r_1- r_2	z_1=(r_1+r_2)/2	f_1	p_1=f_1/T
r_2- r_3	z_2= (r_2+r_3)/2	f_2	p_2=f_2/T
…	…	…	…
r_{K-1}- r_K	$r_k = (r_{K-1}+ r_K)/2$	f_k	p_k=f_k/T

Source: developed by the authors.

Let us consider the exponential membership function (4.1). To find estimates of parameters $\tau \equiv \tau_x$, $\nu \equiv \nu_x$ the following criterion will be used

$$\sum_{i=1}^{m-1}\left(\ln(-\ln \tilde{p}_i) - 2\ln\left(\frac{c-z_i}{\tau}\right)\right)^2 + \sum_{i=m+1}^{k}\left(\ln(-\ln \tilde{p}_i) - 2\ln\left(\frac{z_i-c}{\nu}\right)\right)^2, \tag{5.19}$$

where c denotes the midpoint of m-th class interval with maximum relative frequency $p_m = \max(p_1, p_2, \dots, p_k)$, and $\tilde{p}_i, i = 1,2,\dots,k$ are normalized frequencies for separate class intervals

$$\tilde{p}_i = \frac{p_i}{p_m},\ i = 1,2,\dots,k. \tag{5.20}$$

It is worth noting that normalized frequencies (5.20) are included in the criterion (5.19) in order to find an exponential membership function of a fuzzy number similar to an empirical histogram.

The expressions (5.21) and (5.22) give the minimum of (5.19) with respect to the unknown parameters τ, ν (see Nasibov and Peker (2011) for more details). Thus, we have

$$\hat{\tau} = \exp\left(\frac{2\sum_{i=1}^{m-1}\ln(c-z_i) - \sum_{i=1}^{m-1}\ln(-\ln \tilde{p}_i)}{2(m-1)}\right), \tag{5.21}$$

$$\hat{\nu} = \exp\left(\frac{2\sum_{i=1}^{m-1}\ln(z_i-c) - \sum_{i=1}^{m-1}\ln(-\ln \tilde{p}_i)}{2(k-m)}\right). \tag{5.22}$$

Example 3. Let us consider the data aggregated in the frequency Table 2.

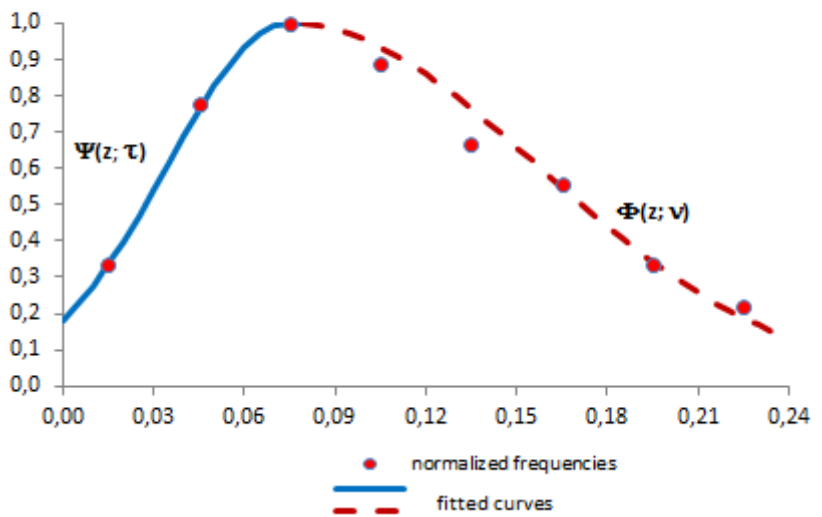
Table 2. Frequency table

Class intervals	Midpoints z_i	Frequencies f_i	Relative frequencies p_i	Normalized frequencies $\tilde{p}_i$
0.00 – 0.03	0.015	3	0.0698	0.3333
0.03 – 0.06	0.045	7	0.1628	0.7778
0.06 – 0.09	0.075	9	0.2093	1.0000
0.09 – 0.12	0.105	8	0.1860	0.8889
0.12 – 0.15	0.135	6	0.1395	0.6667
0.15 – 0.18	0.165	5	0.1163	0.5556
0.18 – 0.21	0.195	3	0.0698	0.3333
0.21 – 0.24	0.225	2	0.0465	0.2222

Source: developed by the authors.

The maximum relative frequency refers to the third class interval, thus we obtain m=3, $p_m = 0.2093$, and c=0.075. The membership function with $\hat{\nu}$, $\hat{\tau}$ derived from (5.21)–(5.22) is illustrated on Figure 3.

Figure 3. Normalized frequencies and a fitted membership function for $\hat{\tau} = 0.059$, $\hat{\nu} = 0.106$



Source: developed by the authors.

5.4. Evaluation of the proposed mortality model based on real data

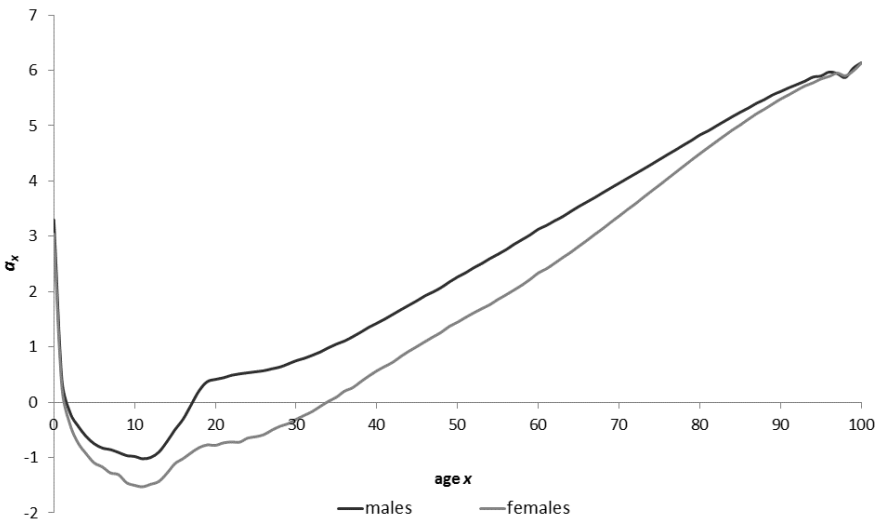
To illustrate theoretical discussions presented in the previous chapters dealing with the proposed mortality model based on quaternion-valued functions the estimates of model parameters will be calculated using real data to compare the *ex-post* forecasting errors with errors yielded by the standard Lee–Carter model (LC).

The analysis is based on the log-central death rates for males and females in Poland from the years 1958–2014. The necessary data were sourced from the Human Mortality Database (www.mortality.org) and from the GUS database (stat.gov.pl). The 2001–2014 death rates served the purpose of evaluating the models' forecasting properties and were not used in estimations.

Estimates a_x, b_x, k_t of the parameters of the quaternion mortality model (5.3) were obtained with the log-central death rates for males and females from the years 1958–2000. Parameters τ_x, ν_x were derived for each separate x using the Nasibov-Peker method, with $\{r_t, t = 1,2,\dots,T\}$ represented by standardized residuals from the ordinary least regression (3.3).

To ensure the clarity of data presentation, the parameter estimates are plotted as shown in Figures 4-6.

Figure 4. Parameters a_x, x = 0,1,...,100 estimated with the CNMM model for males and females

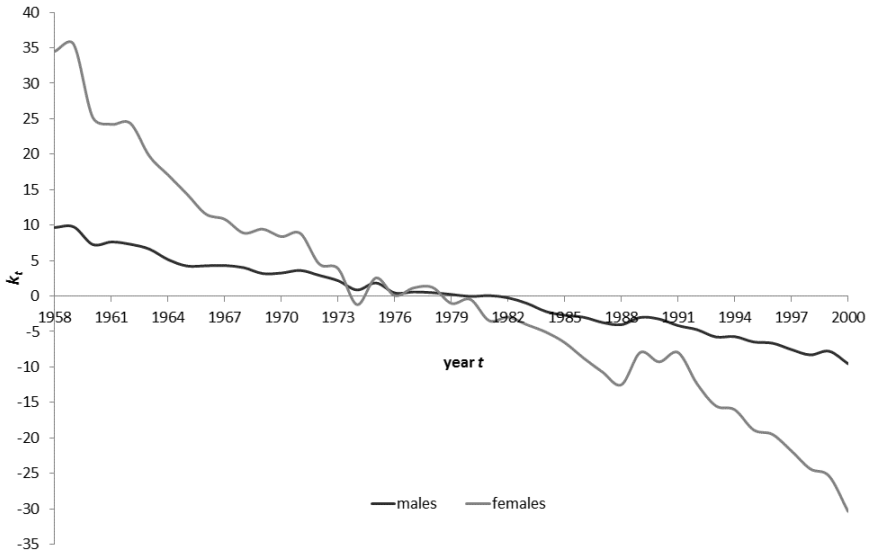


Source: developed by the authors.

Figure 5. Parameters b_x, x = 0.1,...,100 estimated with the CNMM model for males and females

Source: developed by the authors.

Figure 6. Parameters k_t, t = 1958,...,2000 estimated with model CNMM (males and females)



Source: developed by the authors.

The interpretation of the model parameters' estimates a_x, b_x, k_t is similar as in the standard Lee-Carter approach, meaning that a_x, x=0,1,...,X indicate the

general shape of the mortality schedule, the time-varying parameters k_t, $t=1,2,\ldots,T$ represent the general mortality level, and b_x, $x=0,1,\ldots,X$ indicate the pattern of deviations from the age profile when the general level of mortality k_t changes.

The conclusion that can be drawn by comparing two curves plotted in Figure 4 is that average mortality in almost all age groups was higher for men than for women. Despite this fact the shapes of mortality profiles for both sexes seem rather similar, i.e. with a high mortality among children under two years of age, relatively low mortality for children aged 8–12, rising rapidly in the older age groups.

The arrangement of curves in Figure 5 shows that in some age groups the absolute values of b_x are higher for males than for females (i.e. for young or middle ages). It means that the log-central death rates clearly are more sensitive to the temporal changes in mortality for males than those noted for females. What is more, some negative values of b_x are estimated, i.e. for males at age group (34, 67) years. They indicate that male log-central mortality rates at those ages grew in some years of the period under consideration when declining at other ages in response to change of k_t. Figure 6 also shows that the overall mortality trend was generally declining, but at a varying rate. It is also worth noting that this general mortality trend (expressed by k_t) was faster in the subpopulation of women.

The forecasting properties of LC and CNMM models were compared based on the *ex-post* errors measured for each year in the period 2001–2014, i.e. the period which was omitted from parameter estimation. The *ex-post* errors were determined using crisp forecasts of log-central death rates (5.23). Two types of prediction accuracy measures will be used, i.e. a mean squared error (*MSE*) and a mean absolute deviation (*MAD*). The results are summarized in Table 3.

Table 3. Comparison of *ex-post* errors (*MSE* and *MAD*) for LC and CNMM models

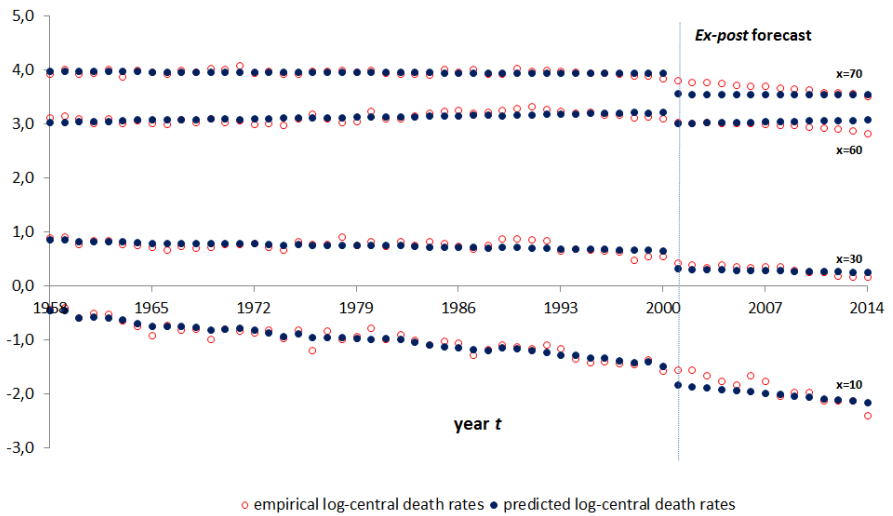
Year	Males				Females			
	MSE		*MAD*		*MSE*		*MAD*	
	LC	CNMM	LC	CNMM	LC	CNMM	LC	CNMM
2001	0.197	0.121	0.182	0.093	0.098	0.140	0.083	0.114
2002	0.204	0.119	0.185	0.091	0.122	0.120	0.107	0.096
2003	0.215	0.120	0.195	0.087	0.122	0.124	0.109	0.098
2004	0.223	0.111	0.206	0.081	0.132	0.113	0.117	0.089
2005	0.230	0.097	0.214	0.070	0.146	0.117	0.129	0.093
2006	0.232	0.110	0.214	0.081	0.152	0.105	0.130	0.083
2007	0.238	0.106	0.219	0.077	0.172	0.116	0.152	0.091
2008	0.257	0.107	0.234	0.083	0.174	0.111	0.156	0.086
2009	0.281	0.114	0.250	0.090	0.191	0.124	0.170	0.092
2010	0.330	0.137	0.302	0.110	0.190	0.095	0.167	0.072
2011	0.341	0.149	0.307	0.119	0.218	0.108	0.191	0.081
2012	0.373	0.174	0.335	0.137	0.215	0.105	0.185	0.081
2013	0.406	0.204	0.359	0.160	0.246	0.138	0.221	0.108
2014	0.469	0.257	0.430	0.212	0.273	0.148	0.245	0.117

Source: developed by the authors

It is worth noting that the CNMM model generates markedly smaller *ex-post* errors (in terms of *MSE* or *MAD* measures) than the LC model, which is visible especially for last years of prediction. For instance, for the prediction years 2010, 2011, 2012, 2013 and 2014 the *ex-post* errors obtained with the CNMM model are less than half of what was obtained with the LC model.

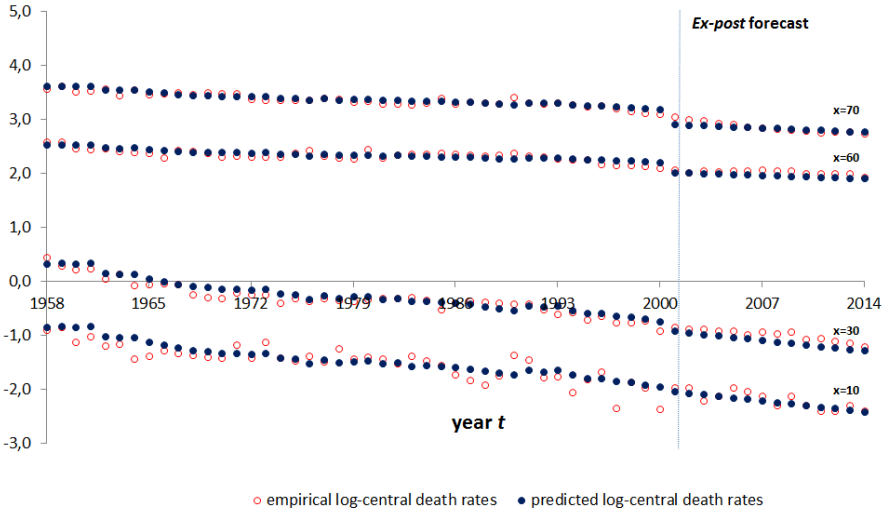
A comparison between empirical log-central death rates and those rates obtained from the CNMM model for some age groups is illustrated in Figures 7 and 8. It is worth noting that the models' parameters were estimated using the 1958–2000 data, therefore the log-central death rates estimated for the years 2001–2014 represent the *ex-post* forecasting.

Figure 7. Real and predicted log-central death rates for some age groups (males)



Source: developed by the authors.

Figure 8. Real and predicted log-central death rates for some age groups (females)



Source: developed by the authors.

6. Final remarks

We should explain to the reader why we have applied the exponential functions while building the theoretical function space as a basis of our new mortality model.

This approach has theoretical and practical advantages. Practical ones are delivered in the paper of Nasibov and Peker (2011), where an easy and useful fitting algorithm is proposed. Based on this algorithm it is possible to fit an exponential functions to the empirical distributions of the observed data, or – as in our case – to the normalized frequencies of residuals in the regression model.

The theoretical advantage of applying exponential membership functions lies in the desirable theoretical properties, because such functions can be transformed into the Hilbert spaces of quaternion valued functions. It is possible that other functions offer better fit to the observed data. This approach will be the subject of further research.

Acknowledgements

Both authors gratefully acknowledge that their research was supported by a grant from the National Science Centre under contract UMO-2015/17/B/HS4/00927.

REFERENCES

BOOTH, H., (2006). Demographic forecasting: 1980 to 2005 in review, International Journal of Forecasting, 22, pp. 547–581.

ISHIKAWA, S., (1997a). Fuzzy inferences by algebraic method, Fuzzy Sets and Systems, 87, pp. 181–200.

ISHIKAWA, S., (1997b). A quantum mechanical approach to a fuzzy theory, Fuzzy Sets and Systems, 90, pp. 277–306.

KOISSI, M.-C., SHAPIRO, A. F., (2006). Fuzzy formulation of the Lee-Carter model for mortality forecasting, Insurance: Mathematics and Economics, 39, pp. 287–309.

KOSIŃSKI, W., PROKOPOWICZ, P., ŚLĘZAK, D., (2003). Ordered Fuzzy Numbers, Bull. Polish Acad. Sci. Math., 51, pp. 327–338.

LEE R. D., CARTER, L., (1992). Modeling and forecasting the time series of U.S. mortality, Journal of the American Statistical Association, 87, pp. 659–671.

NASIBOV, E., PEKER, S., (2011), Exponential Membership Function Evaluation based on Frequency, Asian Journal of Mathematics and Statistics, 4 (1), pp. 8–20.

WILMOTH, J. R., HORIUCHI, S., (1999). Rectangularization revised: variability of age at death within human populations, Demography, 36 (4), pp. 475–495.

ROSSA, A., SOCHA, L., SZYMAŃSKI, A., (2015). Hybrydowe modelowanie umieralności za pomocą przełączających układów dynamicznych i modeli rozmytych, (in Polish), University of Lodz Press.

SZYMAŃSKI, A., ROSSA, A., (2014). Fuzzy mortality model based on Banach algebra, International Journal of Intelligent Technologies and Applied Statistics, 7, pp. 241–265.

TABEAU, E., BERG JETHS, A., HEATHCOTE, CH., (eds.), (2001). Forecasting of mortality in developed countries: insights form a statistical, demographic and epidemiological perspective, Kluwer Academic Publishers, London.

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 725–742, DOI 10.21307/stattrans-2017-009

ESTIMATION OF SMALL AREA CHARACTERISTICS USING MULTIVARIATE RAO-YU MODEL

Alina Jędrzejczak [1,2]**, Jan Kubacki** [2]

ABSTRACT

The growing demand for high-quality statistical data for small areas coming from both the public and private sector makes it necessary to develop appropriate estimation methods. The techniques based on small area models that combine time series and cross-sectional data allow for efficient "borrowing strength" from the entire population and they can also take into account changes over time. In this context, the EBLUP estimation based on multivariate Rao-Yu model, involving both autocorrelated random effects between areas and sampling errors, can be useful. The efficiency of this approach involves the degree of correlation between dependent variables considered in the model. In the paper we take up the subject of the estimation of incomes and expenditure in Poland by means of the multivariate Rao-Yu model based on the sample data coming from the Polish Household Budget Survey and administrative registers. In particular, the advantages and limitations of bivariate models have been discussed. The calculations were performed using the *sae* and *sae2* packages for R-project environment. Direct estimates were performed using the WesVAR software, and the precision of the direct estimates was determined using a balanced repeated replication (BRR) method.

Key words: small area estimation, EBLUP estimator, Rao-Yu model, multivariate analysis.

1. Introduction

The motivation for the paper is twofold. First, the growing demand for high-quality statistical data at low levels of aggregation, observed over the last few decades, has attracted much attention and concern amongst survey statisticians, but only a few works have been devoted to the small area estimation involving the combination of cross-sectional and time-series data. Second, the evidence on income distribution and poverty gathered for OECD countries in the latter part of

[1] Institute of Statistics and Demography, Faculty of Economics and Sociology, University of Łódź. E-mail: jedrzej@uni.lodz.pl.
[2] Centre of Mathematical Statistics, Statistical Office in Łódź. E-mail: j.kubacki@stat.gov.pl.

the first decade of the 2000s confirms that there has been an significant increase in income inequality, which has grown since at least the mid-1980 and there are still substantial differences in regional income levels (see: *Growing Unequal?*, OECD 2008; *Divided We Stand*. *Why Inequality Keeps Rising*. OECD 2011). Due the problem of high disparities between regions it is becoming crucial to provide reliable estimates of income distribution characteristics for small areas. The task is rather difficult as heavy-tailed and extremely asymmetrical income distributions can yield many estimation problems even for large domains. For some population divisions (by age, occupation, family type or geographical area) the problem becomes more severe and estimators of income distribution characteristics can be seriously biased and their standard errors far beyond the values that can be accepted by social policy-makers for making reliable policy decisions. That latter case is the area of applications for small area estimation.

Within the framework of survey methodology and small area estimation one can apply several methods to improve the estimation quality. Making use of auxiliary data coming from administrative registers or censuses within the traditional framework of survey methodology (ratio and regression estimators) can obviously improve the quality of estimates. However, the most important issue is the synthetic estimation that moves away from the design-based estimation of conventional direct estimates to indirect (and usually model-dependent) estimates that „borrow strength" from other small areas or other sources in time and/or in space. The term „borrowing strength" means increasing the effective sample size and is related to using additional information from larger areas, which can be applied for both interest (Y) and auxiliary variables (X). A large variety of small-area techniques, including small area models, have been described in Rao (2003), Rao, Molina (2015). In the paper we are especially interested in the multivariate case of the Rao-Yu model, the extension of the Fay-Herriot model, which "borrows strength" from other domains and over time.

Multivariate models can account for the correlation between several dependent variables and can specifically be applied to the situations when correlated income characteristics are involved. Multivariate models, being extensions of basic small area models, have been studied in some papers within the framework of small area estimation literature. In particular, interesting studies concerning multivariate linear mixed models can be found in the papers by Fay (1987) and Datta et al. (1991). In Datta et al. (1996) one can find the application of multivariate Fay-Herriot model in the context of hierarchical model with the application to estimating the median income of four-person families in the USA. Recently, some papers have been published where the multivariate linear mixed models were employed, including the works by Benavent and Morales (2016), Porter et al. (2015). The interesting applications related to the victimization surveys in the USA can be found in Fay and Diallo (2012), in Fay and in Li, Diallo and Fay (2012). Also, some applications of Rao-Yu model have been published. Here, we can mention the works by Janicki (2016) and Gershunskaya (2015). One of the applications for the univariate case of the Rao-Yu model can

be found in the previous paper of the authors (Jędrzejczak, Kubacki (2016)). The increase in the number of applications in this area can also be related to the recently published package sae2 for R-project environment (Fay, Diallo (2015)).

The aim of this paper is to present the method for estimating small area means on the basis of sample and auxiliary data coming from other areas and different periods of time. The authors' proposition is to use two-dimensional models which can be applied to simultaneously estimate correlated income variables. The example of the application is based on the micro data coming from the 2003–2011 Polish Household Budget Survey on income and expenditure assumed as dependent variables, and administrative registers. In the application two-dimensional Rao-Yu model is compared with simpler estimation techniques.

2. Univariate and multivariate Rao-Yu model

Various small area models can be utilized in order to improve the quality of estimation in the presence of insufficient sample sizes. They can account for between-area variability beyond that explained by traditional regression models and thus make it possible to adjust for specific domains. Most of these models are special cases of the general linear mixed model.

General linear mixed model is a statistical linear model containing both fixed and random effects, which can be described as follows (see e.g.: Rao (2003), Chapter 6.2):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} \tag{1}$$

In the equation given above $\mathbf{y}$ is a $n \times 1$ vector of the observations that can come from a sample survey, $\mathbf{X}$ and $\mathbf{Z}$ are known $n \times p$ and $n \times h$ matrices that can represent auxiliary data, $\mathbf{v}$ and $\mathbf{e}$ are independently distributed random variables with covariance matrices $\mathbf{G}$ and $\mathbf{R}$ respectively, related to the model variance components. Depending on the variance-covariance structure many variants of the model (1) can be specified, among them the model with block-diagonal covariance structure, which has been the basis for many small area models, including the popular Fay-Herriot model or the Rao-Yu model. They are the examples of area-level model in contrast to the unit-level models that are not considered in the paper.

Univariate model

Rao-Yu small area model, which incorporates time series and cross-sectional data, is a special case of the general linear mixed model with block diagonal covariance structure as described in Rao and Yu (1994) and in Rao (2003). A linear mixed model for the population values, θ_{it}, for the domain i ($i=1,\ldots m$) in time t ($t=1,\ldots,T$) is the following

$$\theta_{\mathrm{it}} = \mathbf{x}_i^T\boldsymbol{\beta} + \mathrm{v_i} + \mathrm{u}_{it} \tag{2}$$

where:

x_i^T is a row vector of known auxiliary variables,

β is a vector of fixed effects,

v_i is a random effect for the area i, $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$,

u_{it} is a random effect for the area i and time t, representing the stationary time-series described by AR(1) process

$$u_{it} = \rho u_{i,t-1} + \epsilon_{it}$$

with constraint $|\rho|<1$ and $\varepsilon_{it} \overset{iid}{\sim} N(0, \sigma^2)$.

Based on the model (2) we can obtain the corresponding model for the observed sample values, y_{it}, which takes the form:

$$y_{it} = \theta_{it} + e_{it} = \mathbf{x}_i^T \boldsymbol{\beta} + v_i + u_{it} + e_{it} \tag{3}$$

where:

e_{it} is a random sampling error for the area i and time t, with

$$\mathbf{e}_i = (e_{i1}, \dots, e_{iT})^T$$

following T-variate normal distribution with the mean 0 and known covariance matrix $\boldsymbol{\Sigma}$.

It is worth noting that the random variables v_i, ε_i and $\mathbf{e}_i$ are mutually independent and the matrix $\boldsymbol{\Sigma}$ with diagonal elements equal to sampling variances for the domain i corresponds to the matrix $\mathbf{R}$ from the model (1).

The crucial role in the model is played by the random terms v and u. They are two components constituting the total random effect of the Rao-Yu model. The first one (v) accounts for the between-area variability while the second one (u) accounts for the variability across time. In particular: v_i's are independent and identically distributed random effects that describe time-independent differences between areas; the u_i's follow the autoregressive process with ρ being temporal correlation parameter for all the areas of interest.

Multivariate model

Assume $\boldsymbol{\theta}_{it} = (\theta_{it,1}, \dots, \theta_{it,r})^T$ as a vector of unknown population parameters. Let $\mathbf{y}_{it}$ be a vector of direct estimators of r parameters of interest related to sample observations which can be expressed as $\mathbf{y}_{it} = (y_{it,1}, \dots, y_{it,r})^T$. The multivariate population model for the j-th variable of interest (j=1,...,r) takes the following form (similar model can be found in Fay et al. (2012)):

$$\theta_{it,j} = \mathbf{x}_{it,j}^T \boldsymbol{\beta}_j + v_{i,j} + u_{\mathrm{it,j}} \tag{4}$$

where:

$\mathbf{v}_i = (v_{i,1}, \dots, v_{i,r})^T \overset{iid}{\sim} N_r(0, \boldsymbol{\sigma}_v^2)$ is a vector of random effects for the area i,

u_{it} is a random effect for the area i and time t, representing the stationary time-series described by AR(1) process

$$u_{it,k} = \rho u_{i,t-1,k} + \epsilon_{it,k}$$

with constraint $|\rho|<1$ and $\boldsymbol{\varepsilon}_{it} = (\varepsilon_{it,1}, \dots, \varepsilon_{it,r})^T \overset{iid}{\sim} \mathrm{N}_r(0, \boldsymbol{\sigma^2})$.

It is worth noting that the model (4) also posits a single autoregression parameter ρ and the random variables $\mathbf{v_i}$, $\boldsymbol{\varepsilon}_i$ and

$$\mathbf{e_i} = (e_{i1,1}, e_{i2,1}, \dots, e_{iT,1}, \dots, e_{i1,r}, e_{i2,r} \dots, e_{iT,r})^T$$

are mutually independent.

The sampling model corresponding to the formula (4) can be written as

$$y_{it,j} = \theta_{it,j} + e_{it,j} = \mathbf{x}_{it,j}^T \boldsymbol{\beta}_j + v_{i,j} + u_{\mathrm{it,j}} + e_{it,j} \tag{5}$$

with the covariance matrix of random effects, linking the matrices $\boldsymbol{\sigma}^2$ and $\boldsymbol{\sigma}$, equal

$$\mathbf{G} = \mathbf{M} \otimes \left[\left(\left(\boldsymbol{\sigma}\boldsymbol{\sigma}^{\boldsymbol{T}} \right) \mathbf{u}_c \right) \otimes \boldsymbol{\Gamma}_u + \left(\left(\boldsymbol{\sigma}_{\boldsymbol{v}} \boldsymbol{\sigma}_{\boldsymbol{v}}^{\boldsymbol{T}} \right) \mathbf{u}_c \right) \otimes \boldsymbol{\Gamma}_v \right],$$

where: $\boldsymbol{\Gamma}_u$ is covariance matrix of $\mathbf{u}_i = (u_{i1}, \dots, u_{iT})^T$ with the elements equal to $\rho^{|t-s|}/(1-\rho^2)$ for an entry (t,s) that represent the AR(1) model for $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$, with constraint $|\rho|<1$. Vectors $\mathbf{v}_\mathrm{i}$ represent the random effects, reflecting time-independent differences between areas. The vectors $\boldsymbol{\sigma}_v$ and $\boldsymbol{\sigma}$ represent the model errors connected with the random effects $\mathbf{u}$ and $\mathbf{v}$, respectively, and have r elements each. The matrix $\mathbf{u_c}$ is $r \times r$ matrix of $\rho_{u,jk}$ values with the diagonal elements equal to 1 and for the remaining elements ($j \neq k$), related to the correlation of the random effects $\mathbf{u}$ with respect to the multidimensional structure specified within the model. $\mathbf{M}$ is $m \times m$ diagonal matrix with elements equal to 1.

Using the multivariate Rao-Yu model given by (5) we can formulate the **best linear unbiased predictor (BLUP**) **estimator** of a small area parameter θ_{it} as a linear combination of fixed and random effects:

$$\tilde{\theta}_{iT} = \mathbf{x}_{iT}^T \tilde{\boldsymbol{\beta}} + \mathbf{m}_i^T \mathbf{G_i} \mathbf{V}_i^{-1} (\mathbf{y_i} - \mathbf{X_i} \tilde{\boldsymbol{\beta}}) \tag{6}$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X^T V^{-1} X})^{-1} \mathbf{X^T V^{-1} y}$ is the generalized least squares estimator of $\boldsymbol{\beta}$ and $\mathbf{m}_\mathrm{i}$ is a vector with values equal to 1 for the area i for j-th variable and T-th period of time and zeroes for the other elements and $\mathbf{V_i} = \mathbf{R_i} + \mathbf{Z_i G_i Z_i^T}$. Note that in the multidimensional case, the i subscript is connected with r-dimensional vectors, where r is the number of dependent variables in the multidimensional model.

The procedure of obtaining EBLUP (Empirical BLUP) estimates is involved in the replacement of several variance components by their consistent estimators using Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) procedures (see e.g.: Rao and Molina (2015), pp.102–105).

Assuming that the vector of the estimators of the model variance parameters is $\tilde{\boldsymbol{\delta}} = (\tilde{\sigma}^2, \tilde{\sigma}_v^2, \tilde{\rho})$, the second-order approximation of **mean square error (MSE) of the EBLUP estimator** can be obtained using the following general formula (see e.g.: Rao, 2003, eq.(6.3.15)):

$$\widehat{MSE}\left(\tilde{\theta}_{it}(\tilde{\boldsymbol{\delta}})\right) = g_{1iT}(\tilde{\boldsymbol{\delta}}) + g_{2iT}(\tilde{\boldsymbol{\delta}}) + 2g_{3iT}(\tilde{\boldsymbol{\delta}})$$

where

$$g_{1it}(\tilde{\boldsymbol{\delta}}) = \mathbf{m_i^T}(\mathbf{G_i} - \mathbf{G_i V_i^{-1} G_i})\mathbf{m_i}$$

$$g_{2iT}(\tilde{\delta}) = \mathbf{d_i^T}\left(\sum_{i=1}^{m} \mathbf{X_i^T V_i^{-1} X_i}\right)^{-1} \mathbf{d_i}$$

$$g_{3iT}(\tilde{\delta}) = tr\left[\left(\frac{\partial \mathbf{b}_{iT}^T}{\partial \boldsymbol{\delta}}\right)\mathbf{V}_i\left(\frac{\partial \mathbf{b}_{iT}^T}{\partial \boldsymbol{\delta}}\right)^T \bar{V}(\hat{\boldsymbol{\delta}})\right]$$

where

$$\mathbf{d_i^T} = \mathbf{x_{iT}^T} - \mathbf{b_i^T X_i^T}$$

$$\mathbf{b_i^T} = \mathbf{m_i^T G_i V_i^{-1}}$$

The detailed expressions of the derivatives $\mathbf{b}_i$ can be found in Diallo (2014) and in Fay and Diallo (2012). For the multidimensional case one can also check the sae2 source code (Fay and Diallo (2015)) available at http://cran.r-project.org .

3. Results and discussion

In the application we were interested in the simultaneous estimation of *per capita income* (Y_1) and *expenditure* (Y_2) in Poland by region NUTS2, based on the sample data coming from the Polish Household Budget Survey. Multivariate models can fit to this kind of situations as they account for the correlation between several dependent variables. To improve the estimation quality we decided to formulate a bivariate small area model where the explanatory variables (X_1, X_2) were GDP per capita for regions coming from administrative registers. To obtain better estimates for the year 2011, we decided to utilize historical data coming from the years 2003-2011, which enabled "borrowing strength" not only across areas but also over time. This was possible by using the multivariate Rao-Yu model (5) based on cross-sectional and time-series data and obviously making use of the correlation between the predicted variables. The results obtained on the basis of these model were compared to the ones obtained from the respective univariate models for each response variable and to the classical Fay-Herriot model. The basis for the calculations was the micro data coming from the Polish Household Budget Survey and regional data from the GUS Local Data Bank.

At the first stage, direct estimates of both parameters of interest for 16 regions were calculated from the HBS sample together with their standard errors obtained by means of the Balanced Repeated Replication (BRR) technique. At the second

stage the models were formulated and estimated from the data and finally EBLUP estimates were obtained as well as their MSE estimates. In order to evaluate the possible advantages of the estimators obtained by means of the bivariate Rao-Yu model (5) for j=1,2, we also estimated the parameters of simpler small area models and their corresponding EBLUPs. In particular, we additionally estimated the parameters of:

- the traditional Fay-Herriot model, "borrowing strength" only from other areas,
- univariate Rao-Yu model (eq. 3), "borrowing strength" from areas and over time.

In the computations conducted in R-project environment the packages *sae* and *sae2* have been applied. The sae2 package includes the implementation of the estimation procedure for the Rao-Yu model, which provides an extension of the basic type A model to handle time series and cross-sectional data (Rao (2003)). A special R macro has been developed that simplifies the reading of the input data from Excel spreadsheets, performing calculations for ordinary EBLUP models and Rao-Yu models for both uni- and two-dimensional cases. This macro has been helpful in obtaining the following: the diagnostics for EBLUP models, diagnostic charts for relative estimation errors (REE), relative estimation error reduction (REE reduction) and REE reduction due to time relationships. The macro presented in the appendix describes simple calculations for 3-dimensional Rao-Yu model using sae2 package and eblupRY function.

In Table 1 we show estimation results obtained for the two-dimensional model (5). For each dependent variable the estimates of fixed effects and the parameters of variance-covariance structure of the model, σ^2, σ_v^2 and ρ, are presented.

Table 1. Diagnostics of Rao-Yu two-dimensional model of *available income* and *expenditure* based on sample and administrative data

Variable	Coefficient estimates	Standard error	t-Statistics	P value
Submodel 1: **Y_1- Avail. Income 2003-2011**	**σ_1^2= 1309.49**	**σ_{1v}^2=0.002**	**ρ=0.959**	**LogL=-1415.140**
Intercept	76.455	49.170	1.555	0.120
X_1 GDP per capita	0.030	0.001	21.293	0.000
Submodel 2: **Y_2- Expenditure 2003-2011**	**σ_2^2= 620.050**	**σ_{2v}^2=0.001**	**ρ=0.959**	**LogL=-1415.140**
Intercept	226.620	34.046	6.656	0.000
X_2 GDP per capita	0.021	0.001	21.131	0.000

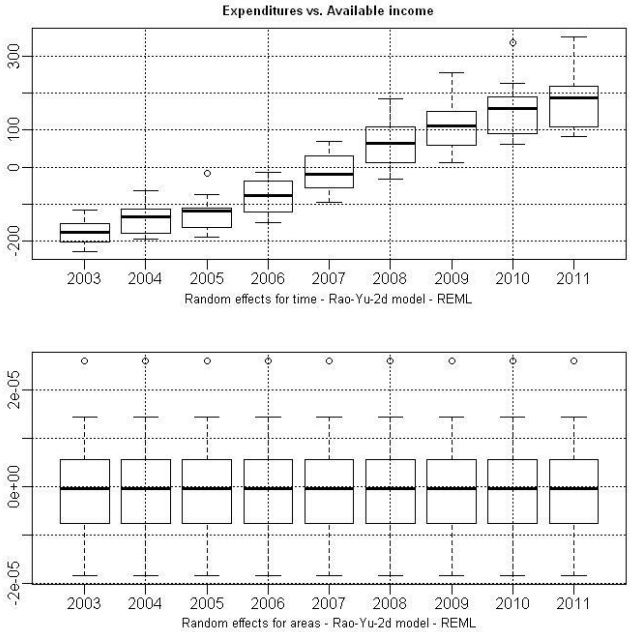


Figure 1. Distributions of random effects obtained for *available income*/*expenditure* 2-dimensional Rao-Yu model (top- **time effects**, bottom-**area effects**)

Source: Own calculations.

The model diagnostics indicate that the parameter σ_v^2 has only a small contribution to the variability of the model, which is mostly determined by time-related component. Figure 1 additionally shows the decomposition of random effects of the model (5) into two components: area effects (v_i) and time-area effects (u_{it}). In the figure it is possible to observe the impact and distribution of these effects over time. The random effects are consumed by time-related component while the influence of time-independent ones remains negligible.

Tables 2 and 3 show estimation results obtained for 16 NUTS2 regions in Poland. To assess the average relative efficiency and efficiency gains for each pair of estimators we utilized the following formulas (see: Rao (2003)):

$$\overline{EFF}_{est1/est2} = \frac{\overline{REE}(EST_1)}{\overline{REE}(EST_2)}, \quad \text{where:} \quad \overline{REE}(EST) = \sum_{i=1}^{m} REE_i$$

Table 2 comprises the estimates of both variables of interest: per-capita *available income* and *expenditure* for regions, obtained using direct estimator, Rao-Yu EBLUP and Rao-Yu two-dimensional EBLUP. Each estimate is accompanied by its estimated precision: relative estimation error (REE) defined as the relative root MSE. The results obtained for income are in general better than the corresponding ones obtained for expenditure, which can be explained by

higher dispersion of income. The improvement is also more evident for regions with poor direct estimates.

Table 2. Estimation results for *available income* and *expenditure* by region in the year 2011 (direct estimates and Rao-Yu EBLUPs – uni- and two-dimensional in PLN)

Region	Direct		Rao-Yu model		2d Rao-Yu model		Efficiency gains due to time effects [%]	
	Para-meter estimate	REE [%]	Para-meter estimate	REE [%]	Para-meter estimate	REE [%]	1D model	2D model
	Available income							
Dolnośląskie	1282.93	2.68	1321.88	1.99	1305.90	1.87	125.7	133.7
Kujawsko-Pomor.	1108.94	2.17	1111.89	1.95	1114.76	1.63	107.3	128.1
Lubelskie	1025.80	2.07	1027.82	1.81	1017.38	1.65	111.6	121.9
Lubuskie	1189.89	1.55	1192.57	1.38	1182.99	1.31	110.1	116.0
Łódzkie	1203.19	2.62	1224.93	2.00	1219.33	1.77	123.1	138.8
Małopolskie	1156.79	2.53	1167.22	2.02	1165.11	1.85	118.7	129.3
Mazowieckie	1622.96	2.02	1669.56	1.59	1649.08	1.42	126.0	141.3
Opolskie	1181.90	1.88	1178.66	1.64	1182.39	1.55	111.6	117.7
Podkarpackie	937.85	2.52	945.67	2.11	946.37	1.77	114.5	136.2
Podlaskie	1224.92	1.45	1208.41	1.34	1202.31	1.33	107.1	108.1
Pomorskie	1286.94	3.09	1298.67	2.20	1298.66	1.84	129.0	154.0
Śląskie	1215.44	0.95	1220.96	0.91	1222.23	0.84	104.3	112.1
Świętokrzyskie	1062.78	2.37	1057.54	2.05	1045.38	1.79	111.0	126.8
Warmińsko-Maz.	1096.87	2.63	1111.93	2.17	1099.61	2.01	115.4	124.8
Wielkopolskie	1135.02	2.73	1170.17	2.09	1148.01	1.84	121.0	137.3
Zachodniopomor.	1231.10	3.16	1226.36	2.27	1210.95	2.08	128.1	140.2
Average	**1185,21**	**2,28**	**1195,89**	**1,85**	**1188,15**	**1,66**	**117.4**	**130.6**
	Expenditure							
Dolnośląskie	1057.49	2.91	1086.02	2.06	1077.05	1.71	127.3	153.2
Kujawsko-Pomor.	922.75	1.16	924.81	1.10	924.16	1.03	104.1	111.7
Lubelskie	856.17	2.03	860.19	1.79	868.50	1.49	110.1	132.1
Lubuskie	975.64	2.13	983.78	1.77	996.11	1.39	115.7	146.9
Łódzkie	1042.70	1.96	1055.32	1.62	1049.65	1.41	116.4	133.3
Małopolskie	982.59	2.62	989.86	1.99	986.73	1.63	123.0	150.3
Mazowieckie	1308.35	1.62	1339.86	1.35	1331.49	1.19	119.4	135.6
Opolskie	1048.57	2.63	1048.66	1.99	1043.37	1.55	124.2	159.7
Podkarpackie	843.00	1.44	845.14	1.33	842.73	1.20	107.0	118.1
Podlaskie	903.42	4.58	889.59	2.70	947.43	1.71	142.4	124.6
Pomorskie	1061.25	1.85	1058.78	1.58	1058.49	1.42	113.2	125.9
Śląskie	1039.73	0.95	1043.57	0.89	1037.58	0.79	104.9	118.2
Świętokrzyskie	848.58	1.84	851.34	1.63	859.73	1.41	109.8	126.3
Warmińsko-Mazur.	870.30	3.06	880.69	2.42	888.67	1.92	116.8	146.7
Wielkopolskie	913.66	2.03	930.58	1.70	928.85	1.49	113.4	129.1
Zachodniopomor.	972.04	2.78	979.81	2.08	992.95	1.77	123.6	145.4
Average	**977,89**	**2,22**	**985,50**	**1,75**	**989,59**	**1,44**	**118.9**	**144.7**

Source: Own calculations.

The last two columns of Table 2 demonstrate "*efficiency gains due to time effects*" obtained as REE reduction for the Rao-Yu models with respect the ordinary EBLUP estimators based on Fay-Herriot model. It can be noticed that the proposed method overwhelms the classical approach by 30.6% for *available income* and by 44.7% for *expenditure*. This improvement was possible due to time relationships incorporated into Rao-Yu models which are not included into the classical Fay-Herriot ones.

As it can be noticed in Table 2, the average efficiency gains coming from time-correlation between random effects are on average doubled when the bivariate Rao-Yu model is taken into account - for *available income* they exceed 30 %, for *expenditure* are almost 45% (the corresponding values for the univariate Rao-Yu model were 14.4% and 18.9%). This improvement comes from the bivariate approach making use of the correlation between several dependent variables.

Table 3 presents in detail the efficiency gains coming from the application of 2d Rao-Yu model for both variables of interest. The EBLUPs based on this model were compared to the direct approach and to the EBLUPs obtained on the basis of simpler model-based approaches. Even with respect to the univariate Rao-Yu model one can observe substantial increase in precision (for income by 11.2% and for expenditure by 21.2%). Figures 2 and 4 present the empirical distributions of REEs for different small area estimators applied in the study while the distributions of REE reduction by means of the proposed model are presented in Figures 3 and 5. As it can be seen in the illustrations the bivariate approach can significantly improve the precision of the estimates.

Table 3. Relative efficiency [in%] for *available income* and *expenditure* in 2011

Region	$EFF_{direct/Rao-Yu2d}$		$EFF_{EBLUP/Rao-Yu2d}$		$EFF_{RaoYu/Rao-Yu2d}$	
	Available income	Expen-diture	Available income	Expen-diture	Available income	Expen-diture
Dolnośląskie	143.7	169.9	133.7	153.2	106.3	120.3
Kujawsko-Pomorskie	133.2	113.1	128.1	111.7	119.4	107.3
Lubelskie	125.4	136.5	121.9	132.1	109.2	119.9
Lubuskie	118.2	153.3	116.0	146.9	105.4	126.9
Łódzkie	147.7	138.7	138.8	133.3	112.8	114.5
Małopolskie	136.6	161.1	129.3	150.3	108.9	122.2
Mazowieckie	142.0	136.0	141.3	135.6	112.2	113.5
Opolskie	120.9	169.5	117.7	159.7	105.5	128.5
Podkarpackie	142.4	120.1	136.2	118.1	118.9	110.4
Podlaskie	109.4	268.2	108.1	124.6	101.0	157.7
Pomorskie	167.8	130.5	154.0	125.9	119.4	111.2
Śląskie	113.1	119.4	112.1	118.2	107.5	112.6
Świętokrzyskie	132.2	130.3	126.8	126.3	114.2	115.1
Warmińsko-Mazurskie	130.9	159.2	124.8	146.7	108.2	125.6
Wielkopolskie	148.2	136.0	137.3	129.1	113.5	113.9
Zachodniopomorskie	152.0	157.0	140.2	45.4	109.5	117.6
Average efficiency gain	**135,2**	**149,9**	**130.6**	**144.7**	**111.2**	**121.2**

Source: Own calculations.

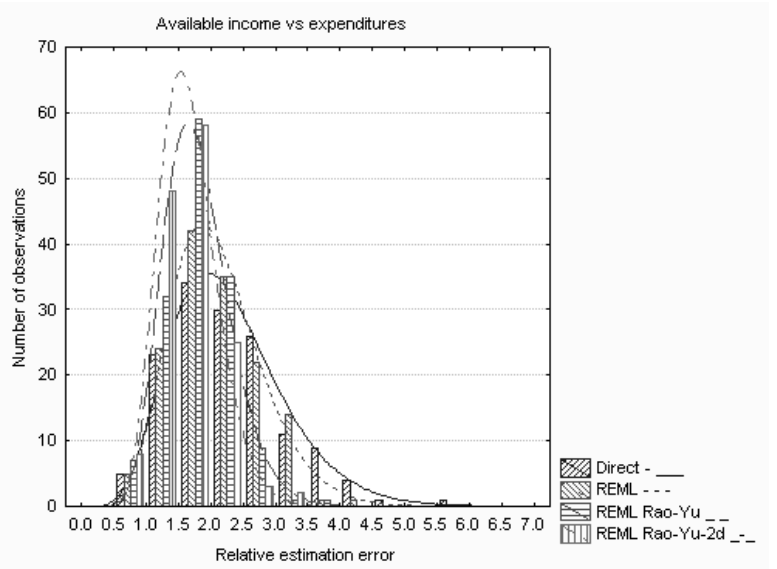


Figure 2. Distribution of REE for *available income* estimates in % in the years 2003-2011 (direct estimator and EBLUPs: ordinary and using Rao-Yu model – both 1 and 2-dimensional)

Source: Own calculations.

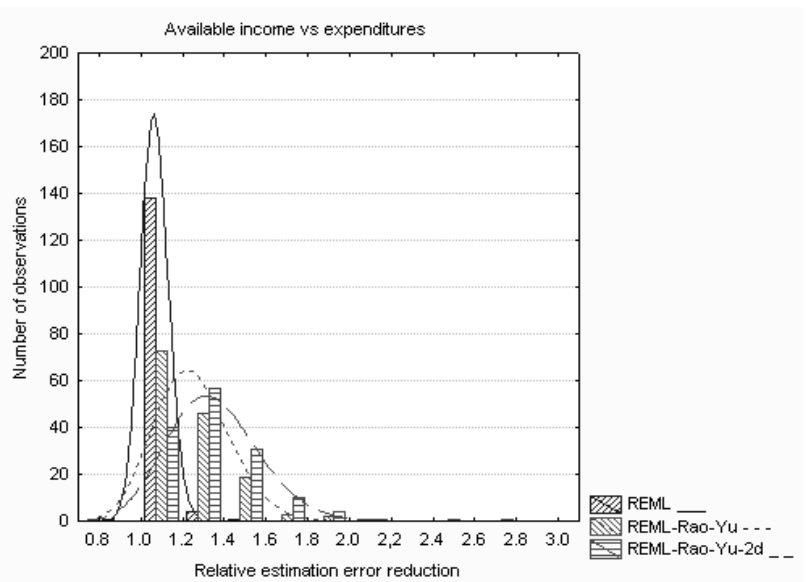


Figure 3. Distribution of REE reduction for *available income* estimates in the years 2003-2011 (direct estimator and EBLUPs: ordinary and using Rao-Yu model – both 1 and 2-dimensional)

Source: Own calculations.

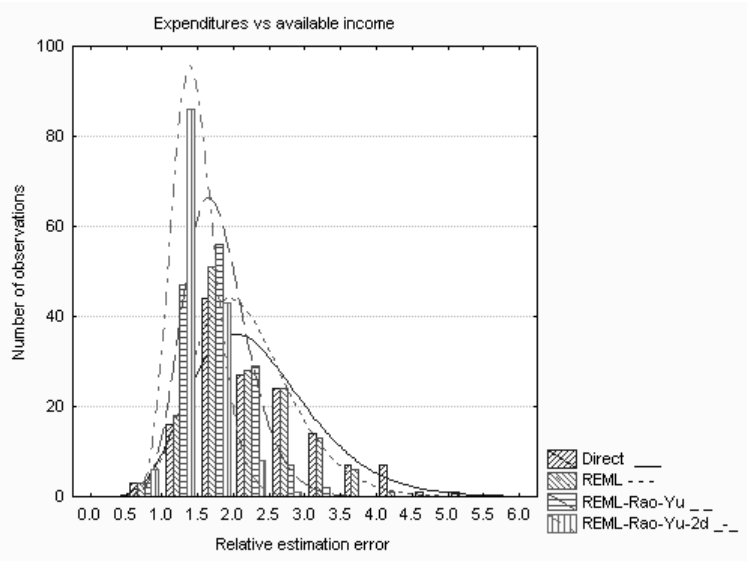


Figure 4. Distribution of REE for *expenditure* estimates in % in the years 2003-2011 (direct estimator and EBLUPs: ordinary and using Rao-Yu model – both 1 and 2-dimensional)

Source: Own calculations.

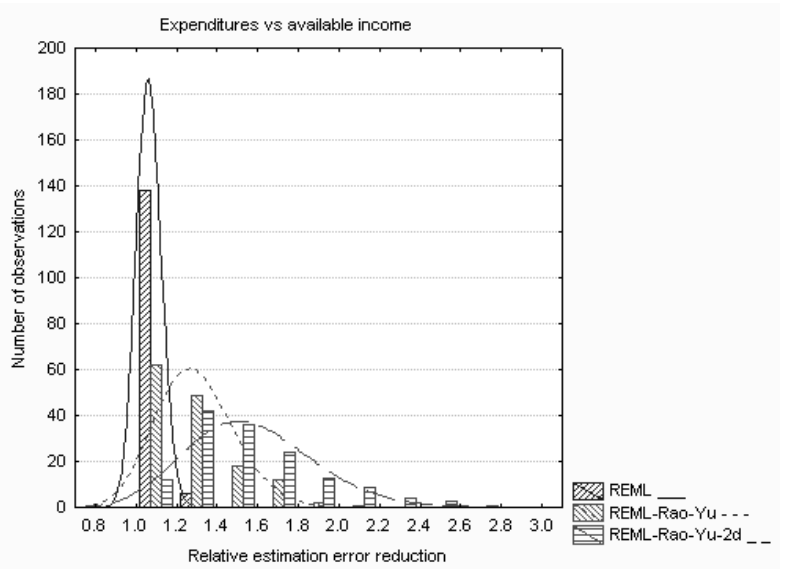


Figure 5. Distribution of REE reduction for *expenditure* estimates in the years 2003-2011 (direct estimator and EBLUPs: ordinary and using Rao-Yu model – both 1 and 2-dimensional)

Source: Own calculations.

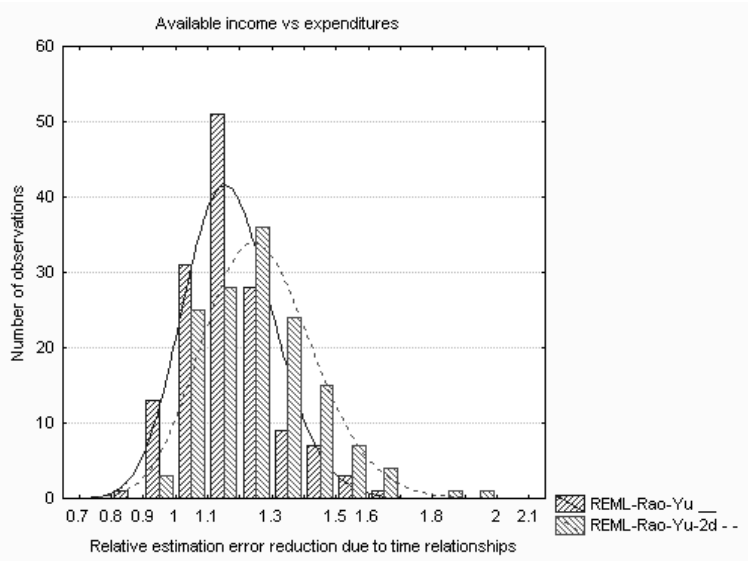


Figure 6. Distribution of REE reduction for *available income* using Rao-Yu EBLUP estimators due to time-related effects (referenced to the ordinary EBLUPs for one and two-dimensional models)

Source: Own calculations.

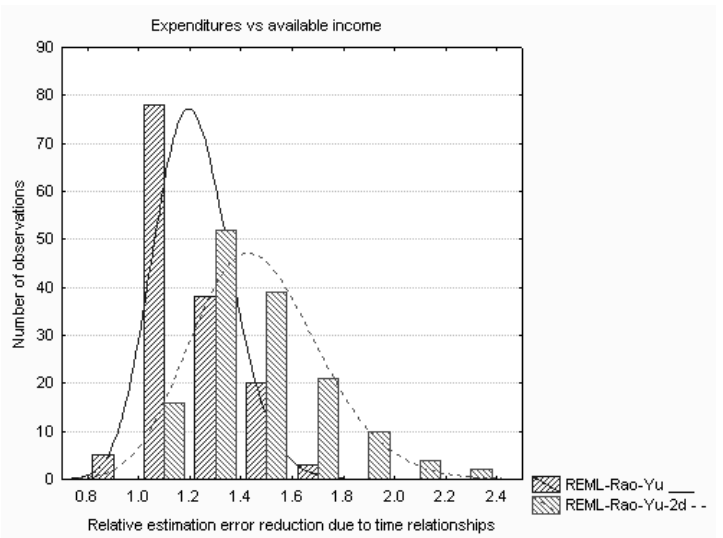


Figure 7. Distribution of REE reduction for *expenditure* using Rao-Yu EBLUP estimators due to time-related effects (referenced to the ordinary EBLUPs for one and two-dimensional models).

Source: Own calculations.

Table 4. Selected diagnostics for 2d Rao-Yu estimators referenced to the ordinary EBLUPs for different categories of income by region in the years 2003-2011

First dependent variable Y_1	Second dependent variable Y_2	$u_{c,(1,2)}$	$\rho_{(Y_1,Y_2)}$	$\frac{\mathrm{REE}_{EBLUP}}{\mathrm{REE}_{\mathrm{R-Y2d}}}$ for Y_1	$\frac{\mathrm{REE}_{EBLUP}}{\mathrm{REE}_{\mathrm{R-Y2d}}}$ for Y_2
Available	Expenditures	0.9464	0.9751	1.080	1.207
Available	Hired work	0.9800	0.9769	1.172	1.238
Available	Self-empl.	0.9321	0.8643	1.033	1.126
Available	Social benef.	0.6379	0.8067	1.001	1.098
Available	Retirm. pays	0.6261	0.8462	1.002	1.077
Available	Disabil. pens.	0.1912	-0.5435	0.999	1.048
Available	Family pens.	-0.0561	0.2464	0.996	1.057
Available	Other social	0.2896	-0.3227	0.997	1.032
Available	Unem.benef.	0.7253	-0.2659	1.010	1.092

Source: Own calculations.

Table 4 summarizes efficiency gains due to the application of two-dimensional models with respect to the classical Fay-Herriot one, which are especially visible for the cases of remarkable correlation between dependent variables Y_1 and Y_2. For the pairs presenting the Pearson correlation exceeding 0.9: *available income* and *expenditure* or *available income* and *income from hired work*, the relative estimation errors are significantly reduced. For example, the average REEs of EBLUPs of *income from hired work* are by 20% higher than the corresponding values obtained by means of the two-dimensional Rao-Yu model. It is worth noting that similar dependencies were observed for the univariate case of the Rao-Yu model (see e.g.: Jędrzejczak, Kubacki (2016)).

4. Conclusions

Multivariate small area models which make use of auxiliary information coming from repeated surveys can lead to significant quality improvements as they borrow information from time and space and additionally exploit the correlation between the considered parameters. In the paper, the advantages and limitations of bivariate small-area models for income distribution characteristics have been discussed. To assess the possible quality improvements, the multivariate Rao-Yu and Fay-Herriot models have been implemented and utilized to the estimation of income characteristics for the Polish households by region. Significant estimation error reductions have been observed for the variables that were evidently time-dependent and strictly correlated with each other and for the domains with relatively poor direct estimators. In the preliminary analysis of the models incorporating larger number of dependent variables also three- and four-

dimensional Rao-Yu models have been specified but the gains from introducing additional dependent variables turned out to be rather ambiguous.

It would be advisable to check this method also for counties (poviats) and determine whether similar time-related relationships, which are observed for regions, could be observed for counties. The analysis presented here may also indicate that further comparisons between the Rao-Yu method and dynamic models, panel econometric models and nonlinear models should be conducted.

REFERENCES

BENAVENT, R., MORALES, D., (2015). Multivariate Fay–Herriot models for small area estimation, Computational Statistics & Data Analysis, Vol. 94, February 2016, pp. 372–390, http://www.sciencedirect.com/science/article/pii/S016794731500170X.

DATTA, G. S., FAY, R. E., GHOSH, M., (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. In: Proceedings of Bureau of the Census 1991 Annual Research Conference, US Bureau of the Census, Washington, DC, pp. 63–79.

DATTA, G. S., GHOSH, M., NANGIA, N., NATARAJAN, K., (1996). Estimation of median income of four-person families: a Bayesian approach. In: Berry, D. A., Chaloner, K. M., Geweke, J. M. (Eds.), Bayesian Analysis in Statistics and Econometrics. Wiley, New York, pp. 129–140.

DATTA, G. S., DAY, B. MAITI, T., (1998). Multivariate Bayesian Small Area Estimation: An Application to Survey and Satellite Data, Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), Vol. 60, No. 3, Bayesian Analysis (Oct., 1998), pp. 344–362, http://sankhya.isical.ac.in/search/60a3/60a3ga.html.

DIALLO, M. S., (2014). Small Area Estimation under Skew-Normal Nested Error Models, A thesis submitted to the Faculty of the Graduate and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Carleton University, Ottawa, Canada.

FABRIZI, E., FERRANTE, M. R., PACEI, S., (2005). Estimation of poverty indicators at sub-national level using multivariate small area models, Statistics in Transition, December 2005, Vol. 7, No. 3, pp. 587–608.

FAY, R. E., (1987). Application of multivariate regression to small domain estimation, Small Area Statistics, Eds.: R. Platek, J. N. K., Rao, C. E., Sarndal, M. P., Singh. Wiley, New York, pp. 91–102.

FAY, R. E., DIALLO, M., (2012). Small Area Estimation Alternatives for the National Crime Victimization Survey, [in:] Proc. Survey Research Methods Section of the American Statistical Association, pp. 3742–3756, https://ww2.amstat.org/sections/SRMS/Proceedings/y2012/Files/304438_73111.pdf.

FAY, R. E, DIALLO, M., PLANTY, M., (2013). Small Area Estimates from the National Crime Victimization Survey, [in:] Proc. Survey Research Methods Section of the American Statistical Association, pp. 1544–1557, http://ww2.amstat.org/sections/srms/Proceedings/y2013/Files/308383_80758.pdf.

FAY, R. E., DIALLO, M., (2015). sae2: Small Area Estimation: Time-series Models, package version 0.1-1, https://cran.r-project.org/web/packages/sae2/index.html.

FAY, R. E., HERRIOT, R. A., (1979). Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data, Journal of the American Statistical Association, 74, pp. 269–277, http://www.jstor.org/stable/2286322.

FAY, R. E., LI, J., (2012). Rethinking the NCVS: Subnational Goals through Direct Estimation, presented at the 2012 Federal Committee on Statistical Methodology Conference, Washington, DC, Jan. 10–12, 2012, https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/Fay_2012FCSM_I-B.pdf.

GERSHUNSKAYA, J., (2015). Combining Time Series and Cross-sectional Data for Current Employment Statistics Estimates, Proceedings of the Joint Statistical Meetings 2015 Survey Research Methods Section, Seattle, Washington, August 8 13, 2015, http://ww2.amstat.org/sections/srms/Proceedings/y2015/files/233962.pdf.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2005). Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model. Working Paper 05-49 (10), Statistics and Econometrics Series 061, Departamento de Estadística, Universidad Carlos III de Madrid, https://e-archivo.uc3m.es/bitstream/handle/10016/230/ws054910.pdf.

JANICKI, R., (2016). Estimation of the Difference of Small Area Parameters from Different Time Periods. Center for Statistical Research & Methodology Research Report Series (Statistics #RRS2016-01). U.S. Census Bureau, https://www.census.gov/srd/papers/pdf/RRS2016-01.pdf.

JĘDRZEJCZAK, A., KUBACKI, J., (2016). Estimation of Mean Income for Small Areas in Poland Using Rao-Yu Model, Acta Universitatis Lodziensis, Folia Oeconomica, 3 (322), pp. 37–53.

LI, J., DIALLO, M. S., FAY, R. E., (2012). Rethinking the NCVS: Small Area Approaches to Estimating Crime, presented at the Federal Committee on Statistical Methodology Conference, Washington, DC, Jan. 10–12, 2012, https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/Li_2012FCSM_I-B.pdf.

MOLINA, I., MARHUENDA, Y., (2015). sae: An R Package for Small Area Estimation, The R Journal, Vol. 7, No. 1, pp. 81–98, http://journal.r-project.org/archive/2015-1/molina-marhuenda.pdf.

OECD, (2008). Growing Unequal? Income Distribution and Poverty in OECD Countries, http://www.oecd-ilibrary.org/social-issues-migration-health/growing-unequal_9789264044197-en.

OECD, (2011). Divided We Stand: Why Inequality Keeps Rising, OECD Publishing, http://dx.doi.org/10.1787/9789264119536-en.

PORTER, A. T., HOLAN, S. H., WIKLE, C. K., (2015). Multivariate spatial hierarchical Bayesian empirical likelihood methods for small area estimation. STAT, 4, 108–116, DOI: 10.1002/sta4.81, http://onlinelibrary.wiley.com/doi/10.1002/sta4.81/abstract.

RAO, J. N. K., (2003). Small Area Estimation, Wiley Interscience, Hoboken, New Jersey.

RAO, J. N. K., MOLINA, I., (2015). Small Area Estimation (2nd edition). John Wiley & Sons, Inc., Hoboken, New Jersey.

RAO, J. N. K., YU, M., (1992). Small area estimation combining time series and cross-sectional data. Proc. Survey Research Methods Section. Amer. Statist. Assoc., pp.1–9, https://ww2.amstat.org/sections/SRMS/Proceedings/papers/1992_001.pdf.

RAO, J. N. K., YU, M., (1994). Small-Area Estimation by Combining Time-Series and Cross-Sectional Data, The Canadian Journal of Statistics, Vol. 22, No. 4, pp. 511–528, http://www.jstor.org/stable/3315407.

R CORE TEAM, (2015). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

WESTAT, (2007). WesVar® 4.3 User's Guide.

YU, M., (1993). Nested error regression model and small area estimation combining cress-sectional and time series data, A thesis submitted to the Faculty of the Graduate and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Carleton University, Ottawa, Canada.

APPENDIX

The macro presented below describes simple calculations for 3-dimensional Rao-Yu model using sae2 package and eblupRY function.

```
library(sae2)
library(RODBC)
channel1 <- odbcConnectExcel("Input.xls")
command <- paste("select * from [Sheet1$]", sep="")
base <- sqlQuery(channel1, command)
data <- c(base$DOCHG_SD, base$D901_SD, base$D905_SD)
D <- 16
T <- 9
n_var <- 3
var_ptr <- vector(mode = "integer", length = D*T*n_var)
for(i in 1:D) {
  for(j in 1:n_var) {
    for(k in 1:T) {
      var_ptr[(i-1)*(T*n_var)+(j-1)*T+k] <- (j-1)*(D*T)+(i-1)*T+k
    }
  }
}
errmat <- diag((data[var_ptr])^2)
resultT.RY <- eblupRY(list(DOCHG_AVG ~ PKBPC_ABS, D901_AVG ~ PKBPC_ABS, D905_AVG ~ PKB_PC), D=D, T=T, vardir = errmat,data=base, ids=base$WOJ, MAXITER = 500)
```

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 743

On behalf of the President of the Statistics Poland (GUS),
Dominik Rozkrut, and the President of the Polish Statistical Association,
Czeslaw Domanski, we are pleased to announce the 2nd Congress of the Polish Statistics to be held on the occasion of the
100th Anniversary of the Central Statistical Office (GUS)
on July 10–12, 2018, in Warsaw.
One of the Congress' section will be devoted to the 25th Anniversary of the *Statistics in Transition new series* and international activities on building statistical capacities.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 745–747

ABOUT THE AUTHORS

Das Radhakanta is an Assistant Professor in Department of Statistics, Presidency University, Kolkata-700073, India. His fields of interest include parametric and non-parametric statistical inference, Bayesian inference.

Gorey Swarangi M. is a research scholar in School of Studies in Statistics, Vikram University, Ujjain, Madhya Pradesh, India. Her main area of research is sampling techniques.

Jędrzejczak Alina is an Associate Professor at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. Simultaneously she holds the position of an expert in the Center of Mathematical Statistics at the regional Statistical Office in Lodz. Her main areas of interest include income distribution, income inequality measurement and decomposition as well as small area estimation. Currently, she is a member of two editorial boards: Statistica & Applicazioni and Acta Universitatis Lodziensis Folia Oeconomica.

Kalbarczyk Małgorzata graduated in economics from the Faculty of Economic Sciences at the University of Warsaw. She works in the Faculty of Economic Sciences at the University of Warsaw, Department of Econometrics and Statistics. Her main fields of current research interest are aging, financial and nonfinancial inter-vivos transfers, last year of life, measuring the costs of children.

Kubacki Jan is a Chief Specialist at the Centre of Mathematical Statistics, Statistical Office in Lodz. In 2009, he received his PhD from Warsaw School of Economics in the area of social statistics, in particular in small area estimation under the supervision of Prof. Jan Kordos. His interests include small area estimation, sample survey methods, classification methods, data processing and statistical software development. He is a member of the Editorial Board of "Statistical News" (Wiadomości Statystyczne).

Kumar Devendra is an Assistant Professor at the Department of Statistics in Central University of Haryana, Mahendergarh, India. His research interests include order statistics, generalized order statistics, record values, distribution theory and statistical inference. Dr Kumar has published more than 100 research papers in international/national journals. He has also published two books/monographs. Dr Kumar is an active member of many scientific professional bodies.

Malik Mansoor Rashid is a research scholar at the Department of Statistics in Amity University, Noida, India. His research interests include order statistics, generalized order statistics and record values. Dr Malik is an active member of many scientific professional bodies.

Miazga Agata graduated in economics from the Faculty of Economic Sciences at the University of Warsaw. She works in the Institute for Structural Research. As an analyst she focused on topics connected with inequalities (especially defining and measuring energy poverty in Poland), economics of education and measuring the costs of children in Poland. Nowadays she uses her scientific experience to promote science.

Mittal Richa is an Assistant Professor in Department of Mathematics, National Institute of Technology, Calicut, Kerala, India and has research experience in Major Research Project under the supervision of Dr Kumari Priyanka in Shivaji College, University of Delhi, India. Her research interests include sampling theory and statistical inference. She is also a reviewer of many international peer-reviewed journals.

Młodak Andrzej is a consultant in the Centre for Small Area Statistics of the Statistical Office in Poznań and an Associate Professor at the Inter–Faculty Department of Mathematics and Statistics at President Stanisław Wojciechowski State University of Applied Sciences in Kalisz. His main areas of interest include multivariate data analysis, taxonomy, cluster analysis, mathematical economics and statistical education. Currently, he is a member of two editorial boards: Statistical News (where he is Deputy Editor–in–Chief) and Statistics in Transition new series.

Nicińska Anna graduated in economics from the Faculty of Economic Sciences at the University of Warsaw. Her main fields of interest are behavioural economics, population economics and socio-economics of ageing. Currently she investigates how private transfers of time and money are shaped by the size and composition of a recipient's social network and by the proximity between a recipient and a donor.

Priyanka Kumari is the Principal Investigator of Major Research Projects sponsored by SERB, India and UGC, New Delhi, India. She is also an Assistant Professor in Department of Mathematics, Shivaji College, University of Delhi, India. Her research interests include sampling theory, statistical inference and industrial statistics. She is also a reviewer of many international peer reviewed journals.

Rossa Agnieszka is an Associate Professor at the Department of Demography and Social Gerontology, Faculty of Economics and Sociology, University of Lodz. She is an author or co-author of approximately 50 papers and a few monographs in the field of event history analysis, multivariate statistical analysis, demographic forecasting, data mining and other research topics.

Sedory Stephen A. is a Professor at Department of Mathematics, Texas A&M University-Kingsville, TX, USA.

Singh Housila P. is a Professor in School of Studies in Statistics, Vikram University, Ujjain, Madhya Pradesh, India. His areas of interest include sampling techniques, inference and probability distribution. He has rich experience in teaching graduate and postgraduate classes. He has published more than 400 research papers in national and international journals.

Singh Sarjinder is a Professor at Department of Mathematics, Texas A&M University-Kingsville, TX, USA.

Szymański Andrzej has graduated from the Electrical Faculty at the Technical University of Lodz, the section of automatic control, and began to work at the Department of Cybernetics at the Mathematical Faculty of the Lodz University. Simultaneously he began to study mathematics at the section of general cybernetics at the Faculty of Mathematics. After finishing the mathematical studies he concentrated his efforts on the regression analysis in Hilbert spaces. Some of his results contained in the PhD thesis have been used in the improved mortality model being the subject of the published article. During a short collaboration with prof. Witold Kosiński and under his influence he began the treatment of the fuzzy variables as the elements of Hilbert space. A few of common papers and books written jointly with Agnieszka Rossa are examples of a new approach in mortality models based on Banach algebras, and quaternion-valued functions giving raise to Hilbert spaces and von Neumann algebras.

Szymkowiak Marcin is an Assistant Professor at the Department of Statistics at Poznań University of Economics and Business (Faculty of Informatics and Electronic Economy) and consultant at the Centre for Small Area Estimation at the Statistical Office in Poznań, with many years of experience in the field of statistical data analysis. He specializes in small area estimation, methods of dealing with nonresponse (imputation and calibration), survey sampling, statistical methods of data integration (probabilistic record linking, statistical matching), and multivariate data analysis.

Vadlamudi Kalyan Rao has recently graduated with MS degree in Statistical Analytics, Computing and Modeling, from Texas A&M University-Kingsville.

Wawrowski Łukasz is an Assistant Professor at the Department of Statistics, Faculty of Informatics and Electronic Economy, Poznań University of Economics and Business. Simultaneously he holds the position of a senior specialist in the Centre for Small Area Estimation at the Statistical Office in Poznań. His main areas of interest include poverty measurement, small area estimation, data visualization and multivariate data analysis.

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 749–752

ACKNOWLEDGEMENTS TO REVIEWERS

The Editor and Editorial Board of Statistics in Transition new series wish to thank the following persons who served from 31 December 2016 to 31 December 2017 as peer-reviewers of manuscripts for the ***Statistics in Transition new series – Volume 18, Numbers 1–4***; the authors` work has benefited from their feedback**.**

Alho Juha, Department of Social Research, University of Helsinki, Helsinki, Finland

Alpu Ozlem, Department of Statistics, University of Eskisehir Osmangazi, Eskişehir, Turkey

Arayal T. R., Department of Statistics Tribhuvan University of Nepal, Nepal

Beltadze Diana, Statistics Estonia, Tallin, Estonia

Berlin Wu, Department of Mathematical Sciences, National Chengchi University Taipei, Taiwan

Betti Gianni, Department of Economy, Politics and Statistics, University of Siena, Siena, Italy

Bouza Carlos N., Department of Applied Mathematics, University of Havana, Havana, Cuba

Dehnel Grażyna, Department of Statistics, Poznan University of Economics and Business, Poznan, Poland

Demkowicz Leszek F., Institute for Computational and Engineering Sciences (ICES), University of Texas at Austin, Texas, USA

Dmytrów Krzysztof, Department of Econometrics and Statistics, University of Szczecin, Szczecin, Poland

Domański Czesław, Department of Statistical Methods, University of Lodz and Polish Statistical Association, Poland

Dudek Hanna, Department of Econometrics and Statistics, Warsaw University of Life Sciences SGGW, Warsaw, Poland

Eliseeva Irina, Department of Statistics and Econometrics, European University at St Petersburg, St Petersburg, Russia

Frenger Monika, Sportwissenschaftliches Institut, Saarland University, Saarbrücken, Germany

Gamrot Wojciech, Department of Statistics, University of Economics in Katowice, Katowice, Poland

García Luengo, Amelia Victoria, Department of Mathematics, University of Almeria – Universidad de Almeria – UAL, Almeria, Spain

Gołata Elżbieta, Department of Statistics, Poznan University of Economics and Business, Poznan, Poland

Gómez Rubio Virgilio, Department of Mathematics, University of Castilla-La Mancha, Ciudad Real, Spain

Hastenteufel Jessica, Lehrstuhl für Betriebswirtschaftslehre insb. Bankbetriebslehre, Saarland University, Saarbrücken, Germany

Heilpern Stanisław, Department of Statistics, Wrocław University of Economics, Wroclaw, Poland

Helenowski Irene B., Department of Preventive Medicine, Feinberg School of Medicine Northwestern University, Chicago, USA

Hussain Zawar, Department of Statistics, King Abdulaziz University, Jeddah, Makkah, Saudi Arabia

Jajuga Krzysztof, Department of Financial Investment and Risk Management, Wrocław University of Economics, Wroclaw, Poland

Jong-Min Kim, Division of Science and Mathematics, University of Minnesota-Morris, Morris, USA

Jozani Mohammad Jafari, Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada

Kalton Graham, WESTAT, and University of Maryland, USA

Khan Rafiqullah, Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh, India

Kharin Yuriy, Department of Mathematical Modeling and Data Analysis, Belarusian State University, Minsk, Belarus

Kowalczyk Barbara, Department of Econometrics, SGH Warsaw School of Economics, Warsaw, Poland

Kończak Grzegorz, Department of Statistics, University of Economics in Katowice, Poland

Krzyśko Mirosław, Department of Probability and Mathematical Statistics, Adam Mickiewicz University, Poznan, Poland

Kumar Sunil, Aliance University, Bangalore, India

Preweda Edward, Department of Geomatics, AGH University of Science and Technology, Cracow, Poland

Pupovac David, Central European University, Budapest, Hungary

Rai Piyush Kant, Department of Mathematics and Statistics, Banasthali University, India

Roths Scott, Department of Statistics, Penn State University, USA

Rusnak Zofia, Department of Statistics, Wrocław University of Economics, Wroclaw, Poland

Sarkar Nityananda, Indian Statistical Institute, Kolkata, India

Sinha Bikas K., Indian Statistical Institute, Kolkata, India

Sinha Samiran, Department of Statistics, Texas A&M University, Texas, USA

Schubert Torben, Fraunhofer Institute for Systems and Innovation Research ISI, Lund University, Lund, Sweden

Schwaab Marcio, Departamento de Engenharia Química da UFRGS, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Suter Christian, Department of Sociology, University of Neuchâtel, Neuchâtel, Switzerland

Szymkowiak Marcin, Department of Statistics, Poznan University of Economics and Business, Poznan, Poland

Vishwakarma Gajendra K., Department of Applied Mathematics, Indian Institute of Technology (ISM) Dhanbad, India

Witkovský Viktor, Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia

Wiśniewski Jerzy W, Faculty of Econometrics and Statistics, Nicolaus Copernicus University, Torun, Poland

Wywiał Janusz, Department of Statistics, University of Economics in Katowice, Poland

Zimková Emília, Department of Economy, Matej Bel University in Banská Bystrica, Banska Bystrica, Slovakia

STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 753–756

INDEX OF AUTHORS, VOLUME 18, 2017

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).