STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

CONTENTS

Invited papers

Research articles

Other articles:

Research Communicates and Letters

Volume 20, Number 4, December 2019

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, e-mail: m.cierpial-wolan@stat.gov.pl
Secretary
Patryk Barszcz, e-mail: p.barszcz@stat.gov.pl, phone number + 48 22 — 608 33 66
Technical Assistant
Rajmund Litkowiec, e-mail: r.litkowiec@stat.gov.pl

Address for correspondence

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

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. I−IV

FROM THE EDITOR

With this issue of *Statistics in Transition new series*, we are closing the 26th year of our journal's existence. It has been yet another year of the successful continuation of our efforts towards increasing the journal's recognisability, prestige and popularity, especially among the scientific community. This edition, as usual, provides an opportunity for us to thank our reviewers for their invaluable contribution to ensuring the high quality of articles published in *SiTns* – their names are listed in the acknowledgments.

Statistics in Transition new series is present in over two dozen international indexing databases (including Scopus), and in virtually all of them, its ranking has been improving year by year – only recently, Index Copernicus has raised our rating to 121.11 points on the ICI Journals Master List. I would like to take this opportunity to thank all our collaborators and stakeholders, including our readers, for supporting us in various ways and thus making our progress possible on a regular and lasting basis.

The issue consists of eleven articles, arranged under four categories. According to the previously-announced and partly-implemented practice of opening an issue with an article written by a distinguished author specially invited by the Editor, as was the case with the issue featuring a paper by Pfeffermann et al. (Vol. 20.2) – this issue opens with a paper by Jacek Wesołowski. The three remaining groups of articles encompass original research papers, other articles, and research communicates.

Jacek Wesołowski's *Multi-domain Neyman-Tchuprov optimal allocation* identifies the eigenproblem solution of the multi-domain efficient allocation as a direct generalization of the classical Neyman-Tchuprov optimal allocation in stratified SRSWOR. This is achieved through the analysis of eigenvalues and eigenvectors of a suitable population-based matrix D. The object of this article's interest lies in the structure of the optimal allocation vector and relative variance rather than in purely numerical tools (even though the eigenproblem solution provides also numerical solutions). The domain-wise optimal allocation and the respective optimal variance of the estimator are determined by the unique direction (defined in terms of the positive eigenvector of matrix D) in the space RI, where I is the number of domains in the population.

The paper by **Erik Šoltés, Silvia Zelinová** and **Mária Bilíková** entitled ***General linear model – an effective tool for analysis of claim severity in motor third party liability insurance*** comes as first under the Research articles section. It focuses on the analysis of claim severity in motor third party liability insurance under the general linear model. The study was based on a set of anonymized data of an insurance company operating in Slovakia, and yielded informative results, which, however, cannot be applied universally to all insurance companies.

Hybrid multiple imputation in a large-scale complex survey by **Humera Razzak** and **Christian Heumann** discusses the problem of missing data in large-scale complex surveys. The paper introduces the 3-stage Hybrid Multiple Imputation (HMI) approach, computationally efficient and easy to implement, to impute complex survey data sets with both continuous and categorical variables. The proposed approach seems to be a good alternative to the existing ones, as it yields lower root mean square errors, empirical standard errors and standard errors than other approaches. In particular, the HMI method has proven to be superior to the existing MI methods in terms of computational efficiency.

In the paper entitled ***Linear Cholesky decomposition of covariance matrices in mixed models with correlated random effects****,* authors **Anasu Rabe**, **D. K. Shangodoyin** and **K. Thaga** present a modelling approach to the covariance matrix in linear mixed models, which facilitates making inference about subject-specific effects. This concerns especially the analysis of repeated measurement data where time-ordering of the responses induces significant correlation. However, as several drawbacks tend to arise when adopting the modelling approach, the authors propose a solution – a linear Cholesky decomposition of the random effects in a covariance matrix – that neutralizes them. The proposed decomposition proves particularly useful in parameter estimation using the maximum likelihood and restricted/residual maximum likelihood procedures.

Modelling language extinction using Susceptible-Infectious-Removed (SIR) model by **Ikoba N. A.** and **Jolayemi E. T.** presents a very interesting approach where a stochastic epidemic model has been applied to the model of indigenous language extinction. The Susceptible-Infectious-Removed (SIR) categorization of an endemic disease has been reformulated to capture the dynamics of indigenous language decline. On the basis of the time in which a language is expected to be extinct, determined by a modified SIR model, several of the surveyed languages appeared to be in danger of becoming extinct comparatively soon, while others were doing reasonably well, thanks to intensified, inter-family language transfers.

Rajesh Singh's and **Madhulika Mishra's** article entitled ***Estimating population coefficient of variation using a single auxiliary variable in simple random sampling*** proposes an improved estimation method for the population coefficient of variation, which uses information on a single auxiliary variable. The authors demonstrate that their estimators are more efficient than the existing ones, and verify these results with both empirical and simulation studies.

Other articles section opens with the paper by **Alina Jędrzejczak** and **Jan Kubacki** – ***Estimation of income characteristics for regions in Poland using spatio-temporal small area models*****.** This study uses income-related and explanatory variables from two separate sources. It compares the properties of EBLUPs based on spatiotemporal models with EBLUPs based on spatial models obtained separately for each year and with EBLUPs based on the Rao-Yu model. The outcome of the analysis demonstrates that spatiotemporal small-area models yield more precise results than the spatial ones. In the computations, it was possible to perform a decomposition with respect to spatial and temporal parts, thus establishing an original, novel solution.

In the paper ***The impact of the applied typology on the statistical picture of population ageing in urban areas in Poland – a comparative analysis***, authors **Tomasz Klimanek** and **Sylwia Filas-Przybył** examine the process of population ageing in Polish urban areas using methods based on the National Official Register of the Territorial Division of the Country (TERYT) classification and on the classification for urban areas (LAU 2 units) – Degree of Urbanisation (DEGURBA). Several traditional demographic measures for population ageing were applied. The comparison of the outcomes of both the above-mentioned ways of measuring the phenomenon of population ageing showed some discrepancies due to different typologies used (DEGURBA and TERYT).

Dominika Polko-Zając's paper ***On permutation location–scale tests*** presents the advantage of permutation tests over classical parametric tests while performing statistical inference. Permutation tests are comparably powerful to parametric tests but at the same time require meeting fewer assumptions. As the study demonstrated, they work well even when applied to small-size samples, where other types of tests usually fail.

Włodzimierz Okrasa and **Dominik Rozkrut,** in their paper ***Subjective and community well-being interaction in multilevel spatial modelling framework***, present the problem of modelling cross-level interaction between the individual and community well-being, taking into consideration geographic membership and spatial variation. The authors develop an explicitly-spatial, multilevel model in order to identify both the space- and place-related effects for the smallest administrative units (LAU2 - gminas). In their analysis, two methods for measuring well-being were employed: (i) individual (subjective) well-being measure, and (ii) multidimensional index of local deprivation composed of eleven domain-scales. The multilevel modelling was finally extended by the authors' attempt to assess the spatial variation effect on the cross-level relationships.

Mir Subzar, S. Maqbool, T. A. Raja, and **Prayas Sharma** seek for more precise and efficient ratio estimators for the estimation of a population mean or a population variance in their article entitled ***A new ratio estimator: an alternative to regression estimator in survey sampling using auxiliary information****,* featured in the last section of the journal – Research Communicates and Letters. Since using auxiliary information makes it possible to improve sampling designs and to enhance the precision and efficiency of the estimators, the presented strategy proposes a class of ratio estimators, modified accordingly, to estimate the population mean. The study demonstrates that the presented estimator is as efficient as the regression estimator, and more efficient than other examined estimators.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. V

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and Statistics 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/Statistics Poland,
Al. Niepodległości 208, R. 296, 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/

EDITORIAL POLICY

The broad objective of *Statistics in Transition new series* is to advance the statistical and associated methods used primarily by statistical agencies and other research institutions. To meet that objective, the journal encompasses a wide range of topics in statistical design and analysis, including survey methodology and survey sampling, census methodology, statistical uses of administrative data sources, estimation methods, economic and demographic studies, and novel methods of analysis of socio-economic and population data. With its focus on innovative methods that address practical problems, the journal favours papers that report new methods accompanied by real-life applications. Authoritative review papers on important problems faced by statisticians in agencies and academia also fall within the journal's scope.

***

ABSTRACTING AND INDEXING DATABASES

Statistics in Transition new series is currently covered in:

- **The New Scimago Journal & Country Rank**
- BASE – Bielefeld Academic Search Engine
- CEEOL
- CEJSH
- CNKI Scholar
- CIS
- Dimensions
- EconPapers
- Elsevier – Scopus
- ERIH Plus
- Google Scholar
- Index Copernicus
- J-Gate
- JournalGuide
- JournalTOCs
- Keepers Registry
- MIAR
- OpenAIRE
- ProQuest – Summon
- Publons
- RePec
- Wanfang Data
- WorldCat
- Zenodo

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 1–12, DOI 10.21307/stattrans-2019-031
Submitted – 09.09.2019; Paper ready for publication – 12.11.2019

MULTI-DOMAIN NEYMAN-TCHUPROV OPTIMAL ALLOCATION

Jacek Wesołowski[1]

ABSTRACT

The eigenproblem solution of the multi-domain efficient allocation is identified as a direct generalization of the classical Neyman-Tchuprov optimal allocation in stratified SRSWOR. This is achieved through analysis of eigenvalues and eigenvectors of a suitable population-based matrix $\mathbf{D}$. Such a solution is an analytical companion to NLP approaches, which are often used in applications, see, e.g. Choudhry, Rao and Hidiroglou (2012). In this paper we are interested rather in the structure of the optimal allocation vector and relative variance than in such purely numerical tools (although the eigenproblem solution provides also numerical solutions, see, e.g. Wesołowski and Wieczorkowski (2017)). The domain-wise optimal allocation and the respective optimal variance of the estimator are determined by the unique *direction* (defined in terms of the *positive* eigenvector of matrix $\mathbf{D}$) in the space $\mathbb{R}^I$, where I is the number of domains in the population.

Key words: Neyman-Tchuprov allocation, multi-domain allocation, eigenproblem, stratified SRSWOR.

MSC2010 Classification: 62D05

1. Introduction

Consider a stratified SRSWOR in a population U of size N with strata $W_1,\ldots,W_H$, which form a partition of U and let N_h denote the size of the stratum W_h, $h=1,\ldots,H$. For a variable $\mathscr{Y}$ defined on U we denote $y_k=\mathscr{Y}(k)$, $k\in U$. The standard estimator of the total $\tau=\sum_{k\in U} y_k$ has the form $\hat{\tau}_{st}=\sum_{h=1}^H N_h\bar{y}_h$, where $\bar{y}_h=\frac{1}{n_h}\sum_{k\in\mathscr{S}_h} y_k$ with n_h denoting the size of the sample $\mathscr{S}_h$ drawn from W_h, $h=1,\ldots,H$. The variance of $\hat{\tau}_{st}$ is $D^2=\sum_{h=1}^H\left(\frac{1}{n_h}-\frac{1}{N_h}\right)N_h^2S_h^2$, where $S_h^2=\frac{1}{N_h-1}\sum_{k\in W_h}(y_k-\bar{y}_{W_h})^2$ is the population variance in W_h, $h=1,\ldots,H$.

In such a setting one of the main issues is the optimal allocation, $\underline{n}=(n_1,\ldots,n_H)$, of the sample among the strata. To this end one may assign a given (relative) variance of the estimator $\hat{\tau}_{st}$ and minimize the costs expressed, e.g. by the total sample size $\sum_{h=1}^H n_h$. An alternative approach is by fixing the total sample size $n=\sum_{h=1}^H n_h$ and minimize the (relative) variance of $\hat{\tau}_{st}$. Both cases are solved through the classical Neyman-Tchuprov optimal allocation procedure (see, e.g. Särndal, Swensson and Wretman, 1992). In particular, it is well known that under the constraint

[1] Statistics Poland and Warsaw University of Technology. E-mail: wesolo@mini.pw.edu.pl.

$n = n_1 + \ldots + n_H$ the Neyman-Tchuprov optimal allocation is

$$n_h = n \frac{N_h S_h}{\sum_{g=1}^{H} N_g S_g}, \qquad h = 1, \ldots, H. \tag{1}$$

Then, the optimal relative variance assumes the form

$$D_{opt}^2 = \frac{1}{\tau^2} \left[\frac{1}{n} \left(\sum_{h=1}^{H} N_h S_h \right)^2 - \sum_{h=1}^{H} N_h S_h^2 \right]. \tag{2}$$

Note that in order for (1) to be a valid solution it is necessary that

$$n < \frac{\left(\sum_{h=1}^{H} N_h S_h\right)^2}{\sum_{h=1}^{H} N_h S_h^2}. \tag{3}$$

Otherwise, (2) gives a non-positive value which is forbidden.

On the other hand, we may want to minimize $\sum_{h=1}^{H} n_h$ under the constraint imposed on the variance of $\hat{\tau}_{st}$ of the form

$$\sum_{h=1}^{H} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) N_h^2 S_h^2 = T,$$

where T is given. Then, it is well known that the optimal allocation is given by

$$n_h = \frac{N_h S_h}{T + \sum_{g=1}^{H} N_g S_g^2} \sum_{g=1}^{H} N_g S_g, \quad h = 1, \ldots, H. \tag{4}$$

The optimal size of the sample is

$$n_{opt} = \frac{\left(\sum_{h=1}^{H} N_h S_h\right)^2}{T + \sum_{h=1}^{H} N_h S_h^2}. \tag{5}$$

Note that these two solutions are dual in the following sense: If we insert $n := n_{opt}$ as given in (5) in the formula (1) we obtain (4). Similarly, if we insert $T := \tau^2 D_{opt}^2$ as given in (2) in the formula (4) we obtain (1).

However, even if (3) is satisfied the Neyman solution may still not be satisfactory: it may happen that the formula (1) yields $n_h > N_h$ for some $h \in \{1, \ldots, H\}$. Moreover, n_h as given in (1) typically is not integer-valued. Therefore, in recent years there has been a growing interest in more refined allocation methods, mostly based on nonlinear programming (NLP), see, e.g. the monograph Valliant, Dever and Kreuter (2013) and references given therein (actually, the literature on the subject is more than abundant). Such procedures give remedies for the basic drawbacks of the Neyman allocation, by imposing block constraints of the form $0 < m_h \leq n_h \leq N_h$, $h = 1, \ldots, H$, on entries of the allocation vector $\underline{n}$. Recently numerical procedures for optimal positive integer solutions also have appeared in the literature, see, e.g. Friedrich, Münnich, de Vries and Wagner (2015) or Wright (2017). Nevertheless, the Neyman-Tchuprov solution remains the only one which gives insight into the analytic structure of the optimal allocation and the optimal variance. For example, it is obvious from (2) that, up to a constant additive term (which is typically small), the optimal (relative) variance is of order $1/n$.

The situation becomes much more complex in the case of multi-domain efficient allocation. In such a setting the population is partitioned into disjoint domains (eventually, domains are further partitioned into strata). The task is to allocate the sample in the domains (eventually in the strata in each domain) in such a way that, simultaneously, the estimators of the total value of a given variable in every domain and in the whole population have minimal variances or relative variances (the precise formulation of the problem is given at the beginning of Section 2). Apparently, such a statement of the allocation problem is natural in many surveys when the goal is to estimate the parameter of interest not only for the whole population but for all the domains the population is partitioned into (e.g. admistration regions in a given country).

NLP procedures are often relatively easily adjustable to multi-domain efficient allocation. One example of such an adjustment is the procedure proposed in Choudhry, Rao and Hidiroglou (2012) (referred to as CRH in the sequel), which is explained in detail later on in this section. A respective useful adjustment of the Neyman-Tchuprov approach seems to be far more challenging.

One example of such an approach is provided by Longford (2006), where the author suggested to minimize (under a constraint given by the total sample size) the objective function

$$\sum_{i=1}^{I} P_i D^2(\bar{y}_i) + GP_+ D^2(\bar{y}_{st}), \tag{6}$$

where P_i, $i = 1, \ldots, I$ are relative preassigned weights which describe ”importance” of domains, $P_+ = \sum_{i=1}^{I} P_i$ and G is a weight responsible for a priority for the variance of the population mean estimator. Mathematically, this approach reduces to the Neyman allocation scheme. The weights $(P_i, i = 1, \ldots, I)$ are designed in order to cover, at least to some extent, jointly the optimality issue for domains and for the whole population. As pointed out in Friedrich, Münnich and Rupp (2018), the approach of Gabler, Ganninger and Münnich (2012), in which additional box constraints on the strata (or domain) sample sizes are imposed, can be used also in this context. However, within such multi-domain adjustment it is not clear how to assess the impact of values of weights P_i, $i = 1, \ldots, I$, and GP_+ on variances $D^2(\bar{y}_i)$, $i = 1, \ldots, I$, and $D^2(\bar{y}_{st})$. In a numerical example given in the Appendix of Khan and Wesolowski (2019) it is visible that the control on the domain-wise efficiency within this kind of approach is rather problematic.

On contrary, the eigenproblem approach to the domain-optimal allocation gives a full control of the domain-wise efficiency. Moreover, the optimal allocation is given through explicit formulas, not just numerically. This is the essence of the present paper, in which we describe the eigenproblem approach as a generalization of the classical Neyman-Tchuprov methodology to the case of multi-domain optimal allocation. Such eigenproblem setting in the context of the domain-wise efficient allocation originally was proposed in Niemiro and Wesołowski (2001), and developed more recently in Wesołowski and Wieczorkowski (2017), and Khan and Wesołowski (2019). In the first of these papers the authors considered two-stage sampling schemes with SRSWOR and stratification either at the first or at the second stage. The setting considered there imposed jointly two sample size constraints: one on the sample size at the first stage (either in terms of the number of PSUs or SSUs) and one on the sample size at the second stage. Such constraints setting was studied also in the second paper, but for a wider family of sampling schemes: SRSWOR with stratification at both the first and the second stage and the Hartley-Rao scheme at the first stage and stratified SRSWOR at the second stage. Each of these schemes was also considered with additional constraints of equal SSU sample sizes within each of PSUs. The last of three papers dealt with the problem under a single sample size constraint, which was formulated in terms of the expected overall cost. Except of two-stage stratified SRSWOR sampling schemes taken into account in earlier papers, here a combination of *pps* sampling and stratified SRSOWR either at the first or at the second stage was also considered. Finally, the eigenproblem approach was applied in the three-stage sampling scheme with SRSWOR (with no stratification) at each stage. Survey applications and some additional refinements of the eigenproblem approach were given, e.g. in Kozak (2004), Kozak and Zieliński (2005) and Kozak, Zieliński and Singh (2008).

Before we move to a detailed description of the eigenproblem approach, we will first analyze the setting of CRH. These authors consider a population U partitioned into disjoint domains U_i, $i=1,\ldots,I$. In each domain U_i the sample of size n_i is drawn independently according to the SRSWOR, $i=1,\ldots,I$. The aim is to minimize the total sample size

$$g(\underline{n}) = n_1 + \ldots + n_I$$

under the constraints for relative variances of estimators of the domain totals

$$T_i := \tfrac{1}{\tau_i^2}\left(\tfrac{1}{n_i}-\tfrac{1}{N_i}\right)N_i^2S_i^2 \le RV_{oi}, \quad i=1,\ldots,I, \tag{7}$$

where $\tau_i = \sum_{k\in U_i} y_k$ is the total for the ith domain, $i=1,\ldots,I$, and the constraint on the relative variance of the estimator of population total

$$\mathrm{S} := \tfrac{1}{\tau^2}\sum_{i=1}^{I}\left(\tfrac{1}{n_i}-\tfrac{1}{N_i}\right)N_i^2S_i^2 \le RV_o. \tag{8}$$

Note that in this approach one specifies conditions for each of domains and for the whole population separately by assigning (given) upper bounds RV_{oi}, $i=1,\ldots,I$ and RV_o. The problem was solved in CRH under additional box constraints of the form $0<n_i\le N_i$, $i=1,\ldots,I$, by the NLP method involving the popular Newton-Raphson algorithm. An extension of this approach to the case of stratified SRSWOR in each of the domains is rather straightforward.

Actually, in the case of the problem considered in CRH with constraints restricted to (7), i.e. to those imposed on the relative variances of the estimators of domain totals, the overall sample size is minimized by the trivial solution

$$n_i = \left\lceil \frac{N_i S_i^2}{\tau_i^2 T_i + N_i S_i^2} \right\rceil \in (0, N_i], \quad i = 1, \ldots, I.$$

Of course, it may happen that for such values of n_i's, $i = 1, \ldots, I$, condition (8) may not hold and only then the numerical procedure is needed.

NLP solutions, as the one described in CRH, typically are efficient and rather universal tools for optimal allocation in real surveys, when the practitioners need just numerical values for allocation of the sample in the particular survey. Nevertheless, they have forms of *black boxes*, that is, they are fed with population data (or estimates) and their output gives numbers responsible for allocation. Consequently, such numerical methods do not provide any information on the structure of optimal solutions (the allocation vector and the optimal relative variance), while such structural knowledge is important at the stage of survey design, e.g. for assigning proper efficiency priorities or for strata and/or domains construction.

To shed more light on the structure of optimal solutions we will analyze the eigenproblem approach. As it has been already mentioned, this methodology was developed recently in Wesołowski and Wieczorkowski (2017), referred to as WW in the sequel. To large extent, the results of the present paper depend on a correct interpretation of introductory Th. 2.3 of WW, where stratified SRSWOR in each of domains was analyzed. In comparison with WW, the formulas for domain optimal allocation, which are given in terms of an eigenvector of certain population dependent matrix, will be slightly modified here due to (known) priority weights assigned to each of domains. More importantly, a new analytic formula for the optimal relative variance in terms of this eigenvector will be derived. Combined together, these formulas allow one to conclude that the eigenvector solution is a direct generalization of the classical Neyman-Tchuprov allocation. This is the main message of the present article. In particular, we will see that in the case when there are no domains (i.e. when $I = 1$), the new formulas are reduced directly to (1) and (2). Moreover, in the situation when there are no strata in the domains the eigenvector solution is an analytic alternative to the NLP solution of CRH. Last but not least, let us mention that the analytic formulas we obtain can be also used for computing particular values of the optimal allocation vector (procedures for eigenvectors and eigenvalues are available in many computer packages, e.g. procedure *eigen* in the R package). Typically, numerical values obtained in this way, agree with NLP solutions.

Finally, let us mention that while being attractive at the analytical and theoretical level, the eigenproblem apporach has its limitations: e.g. it may give the allocation values which exceed the strata sizes. The NLP black box methods do not have this deficiency. Therefore, it would be plausible to overcome this drawback of the eigenproblem approach. In particular, it would be interesting to study the question whether a recursive version of the proposed methodology, similar to the recursive Neyman approach (see, e.g. Rem. 12.7.1 in Särndal, Swensson and Wretman

(1992)), gives the domain-wise efficient allocation with sample strata sizes within the strata size ranges. At present, this problem is under study. It would be also interesting to investigate possibilites of multivariate extensions of the eigenproblem methodology, since in many applications one would like to allocate the sample taking under account optimality with respect to more than one variable. A step in this direction was made in Kozak (2004).

2. Minimization of domain-wise relative variances

In the case of stratified domains, $U_i = \bigcup_{h=1}^{H_i} W_{i,h},\ i = 1,\ldots,I$, the domain relative variances are

$$T_i = \frac{1}{\tau_i^2}\sum_{h=1}^{H_i}\left(\frac{1}{n_{i,h}} - \frac{1}{N_{i,h}}\right)N_{i,h}^2 S_{i,h}^2, \quad i = 1,\ldots,I, \tag{9}$$

where $N_{i,h} = \#(W_{i,h})$, $S_{i,h}^2 = \frac{1}{N_{i,h}-1}\sum_{k\in W_{i,h}}(y_k - \bar{y}_{i,h})^2$, with $\bar{y}_{i,h} = \frac{1}{N_{i,h}}\sum_{k\in W_{i,h}} y_k$, $\tau_i = \sum_{k\in U_i} y_k$ and $n_{i,h}$ being the size of the sample in hth stratum of ith domain, $i = 1,\ldots,I$. The relative total variance is

$$\mathrm{S} = \frac{1}{\tau^2}\sum_{i=1}^{I}\sum_{h=1}^{H_i}\left(\frac{1}{n_{i,h}} - \frac{1}{N_{i,h}}\right)N_{i,h}^2 S_{i,h}^2. \tag{10}$$

We will minimize simultaneously all T_i, $i = 1,\ldots,I$, as well as S under the constraint on the total sample size. To this end to each domain U_i a (known) priority weight $\kappa_i > 0$ will be assigned. These weights, describing domain-wise efficiency priorities can be read out e.g. from CRH assignment of the domain-wise relative variance boundary values RV_{oi}, $i = 1,\ldots,I$. That is, for any $i = 1,\ldots,I$, the priority weight κ_i can be taken as $\kappa_i = \frac{RV_{oi}}{RV}$, where $RV = \sum_{i=1}^{I} RV_{oi}$.

Then, (9) can be written as

$$\frac{1}{\tau_i^2}\sum_{h=1}^{H_i}\left(\frac{1}{n_{i,h}} - \frac{1}{N_{i,h}}\right)N_{i,h}^2 S_{i,h}^2 = \kappa_i T, \quad i = 1,\ldots,I, \tag{11}$$

where T is an unknown positive constant. Under (11) the parameter T controls both the relative variances in domains and the overall relative variance S of the estimator of the population mean. To see the latter, note that (10) implies

$$\mathrm{S} = \left(\frac{1}{\tau^2}\sum_{i=1}^{I}\rho_i^2\right)T, \tag{12}$$

where $\rho_i = \tau_i\sqrt{\kappa_i}$, $i = 1,\ldots,I$. Therefore, T will be called the *base* of the relative variance.

To formulate the main result we need to introduce and analyze properties of a population $I \times I$ matrix

$$\mathbf{D} = \frac{1}{n}\underline{a}\,\underline{a}^T - \mathrm{diag}(\underline{c}), \tag{13}$$

where

$$\underline{a} = (a_1, \ldots, a_I)^T = \left(\frac{1}{\rho_i} \sum_{h=1}^{H_i} N_{i,h} S_{i,h},\, i = 1, \ldots, I \right)^T, \tag{14}$$

$$\underline{c} = (c_1, \ldots., c_I)^T = \left(\frac{1}{\rho_i^2} \sum_{h=1}^{H_i} N_{i,h} S_{i,h}^2,\, i = 1, \ldots, I \right)^T \tag{15}$$

and $\mathrm{diag}(\underline{c})$ is a diagonal matrix with the vector $\underline{c}$ being its diagonal.

Proposition 2.1 *Assume that*

$$n < \sum_{i=1}^{I} \frac{\left(\sum_{h=1}^{H_i} N_{i,h} S_{i,h} \right)^2}{\sum_{h=1}^{H_i} N_{i,h} S_{i,h}^2}. \tag{16}$$

Then, $\mathbf{D}$ has the unique, simple and positive eigenvalue λ^* and the unique unit eigenvector $\underline{v}^* \in \mathbb{R}^I$ associated to λ^*, which has all coordinates positive.

The proof of this proposition is given in Section 3.

It appears that the eigenvalue λ^* and the eigenvector $\underline{v}^*$ from Prop. 2.1 are crucial for the multi-domain version of the classical Neyman-Tchuprov allocation, which is the main result of this paper.

Theorem 2.2 *Consider stratified SRSWOR in all domains (as described above) with the total sample size*

$$n = \sum_{i=1}^{I} \sum_{h=1}^{H_i} n_{i,h} \tag{17}$$

and assume that (16) *holds. Let λ^* and $\underline{v}^*$ be as in Prop. 2.1.*

Then, the multi-domain optimal allocation (with priority weights κ_i, $i = 1, \ldots, I$), that is the allocation satisfying (11) *with the minimal base of relative variance under the sample size constraint* (17) *has the form*

$$n_{i,h} = n \frac{v_i^* N_{i,h} S_{i,h} / \rho_i}{\sum_{r=1}^{I} v_r^* \sum_{g=1}^{H_r} N_{r,g} S_{r,g} / \rho_r}, \quad h = 1, \ldots, H_i,\ i = 1, \ldots, I. \tag{18}$$

For the optimal base of the relative variance T_{opt} we have $T_{opt} = \lambda^*$. Moreover,

$$T_{opt} = \frac{1}{\sum_{i=1}^{I} \rho_i^2} \left[\frac{1}{n} \left(\sum_{i=1}^{I} \frac{\rho_i}{v_i^*} \sum_{h=1}^{H_i} N_{i,h} S_{i,h} \right) \left(\sum_{i=1}^{I} \frac{v_i^*}{\rho_i} \sum_{h=1}^{H_i} N_{i,h} S_{i,h} \right) - \sum_{i=1}^{I} \sum_{h=1}^{H_i} N_{i,h} S_{i,h}^2 \right]. \tag{19}$$

Remark 2.1 *Note that* (18)*, while inserted into* (9)*, implies*

$$T_{i,opt} = \frac{\rho_i}{n \tau_i^2 v_i^*} \sum_{h=1}^{H_i} N_{h,i} S_{h,i} \sum_{r=1}^{I} \frac{v_r^*}{\rho_r} \sum_{g=1}^{H_i} N_{r,g} S_{r,g} - \frac{1}{\tau_i^2} \sum_{h=1}^{H_i} N_{i,h} S_{i,h}^2. \tag{20}$$

The proof of Theorem 2.2 is given in Section 3.

Note that (19) together with (12) implies that, similarly as in the classical Neyman-Tchuprov case, the overall relative variance is of order $1/n$ up to the additive (typically small) constant.

In the boundary case of $I=1$, that is, when there are no domains in U, $\frac{\rho_1}{v_1^*}$ cancels out in (18) and (19). Consequently, these formulas are transformed into the original Neyman-Tchuprov formulas (1) and (2), respectively. Also, (16) becomes (3).

Another boundary case is when there are no strata in domains. Then, from Th. 2.2 we obtain an analytic solution which can be viewed as an alternative to the NLP approach of CRH. In this case (no strata in domains) the matrix $\mathbf{D}$, as defined in (13), has a simple form since then

$$\underline{a}=\left(\tfrac{N_iS_i}{\rho_i},\, i=1,\ldots,I\right)^T, \qquad \underline{c}=\left(\tfrac{N_iS_i^2}{\rho_i^2},\, i=1,\ldots,I\right)^T.$$

Since $H_i=1$, $i=1,\ldots,I$, the inequality (16) is a consequence of the natural assumption $n<N$, where $N=\sum_{i=1}^I N_i$. Let $\underline{v}^*$ be the unique unit eigenvector with positive coordinates for the simplified $\mathbf{D}$ matrix given above (by Prop. 2.1 we know that such vector $\underline{v}^*$ exists).

Corollary 2.3 *In the case of SRSWOR in each of domains (no strata) the optimal domain-wise efficient allocation (with priority weights κ_i, $i=1,\ldots,I$) under the sample size constraint*

$$\sum_{i=1}^I n_i = n < N \tag{21}$$

has the form

$$n_i = n\,\frac{v_i^* N_iS_i/\rho_i}{\sum_{j=1}^I v_j^* N_jS_j/\rho_j}, \quad i=1,\ldots,I. \tag{22}$$

Then, the optimal base of the relative variance assumes the form

$$T_{opt}=\frac{1}{\sum_{i=1}^I \rho_i^2}\left[\frac{1}{n}\left(\sum_{i=1}^I \frac{\rho_i}{v_i^*}N_iS_i\right)\left(\sum_{i=1}^I \frac{v_i^*}{\rho_i}N_iS_i\right)-\sum_{i=1}^I N_iS_i^2\right]. \tag{23}$$

On the other hand, we may want to minimize the sample size $n=\sum_{i=1}^I\sum_{h=1}^{H_i} n_{i,h}$ under the constraints (9) with given T_i, $i=1,\ldots,I$. A straightforward application of the Lagrange multipliers gives the analog of (4) of the form

$$n_{i,h}=N_{i,h}S_{i,h}\,\frac{\sum_{g=1}^{H_i}N_{i,g}S_{i,g}}{\tau_i^2T_i+\sum_{g=1}^{H_i}N_{i,g}S_{i,g}^2}, \quad h=1,\ldots,H_i,\, i=1,\ldots,I. \tag{24}$$

Therefore,

$$n_{opt}=\sum_{i=1}^I \frac{\left(\sum_{h=1}^{H_i}N_{i,h}S_{i,h}\right)^2}{\tau_i^2T_i+\sum_{h=1}^{H_i}N_{i,h}S_{i,h}^2}. \tag{25}$$

Similarly as in the original Neyman-Tchuprov case the two approaches are dual in the following sense: (18) follows by inserting $T_i := T_{i,opt}$ as given in (20) into (24); dually, we note that (20) (again with $T_i := T_{i,opt}$) can be rewritten as the following relation between elements of the eigenvector $\underline{v}^*$

$$\frac{n v_i^*}{\sum_{j=1}^{I} v_j^* \sum_{h=1}^{H_i} \frac{N_{j,h} S_{j,h}}{\rho_j}} = \frac{\rho_i \sum_{h=1}^{H_i} N_{h,i} S_{h,i}}{\tau_i^2 T_i + \sum_{h=1}^{H_i} N_{i,h} S_{i,h}^2}, \quad i = 1, \ldots, I$$

and this formula gives (24) when combined with (18).

3. Proofs

The proofs, to some extent, can be read out from Sec. 2 of WW. Nevertheless, to make this article more self-contained we provide most of the arguments referring only to a rather technical Prop. 2.2 from WW. The main new aspect of the argument is related to the formula (19) for the base of relative variances.

[Proof of Prop. 2.1] We first refer to Prop. 2.2 of WW, the proof of which was based on the Weyl inequalities (relating eigenvalue of the sum of two matrices to eigenvalues of the summands). Then, see Rem. 2.1 in WW, it follows that there exists a unique, positive eigenvalue of the matrix $\mathbf{D}$, denoted here by λ^*. Moreover, the eigenvalue λ^* is simple, i.e. its eigenspace is one-dimensional.

To show that there exists a unit length eigenvector $\underline{v}^*$ (associated with λ^*) with all coordinates positive we use the celebrated Perron-Frobenius theorem: *If $\mathbf{A}$ is a matrix with all strictly positive entries then there exists a unique positive eigenvalue v of $\mathbf{A}$, it is simple and such that $v > |\lambda|$ for any other eigenvalue λ of $\mathbf{A}$. The respective eigenvector (attached to v) has all entries strictly positive (up to scalar multiplication)* - see, e.g. Kato (1981), Th. 7.3 in Ch. 1.

Fix a number $\alpha > \max_{1 \leq i \leq I} c_i > 0$. Note that the matrix $\mathbf{D}_\alpha := \mathbf{D} + \alpha \mathbf{Id}$, where $\mathbf{Id}$ is an $I \times I$ identity matrix, has all entries strictly positive. For any eigenvalue λ of $\mathbf{D}$ and the respective eigenvector $\underline{w}$ we have

$$\mathbf{D}_\alpha \underline{w} = (\lambda + \alpha) \underline{w}, \tag{26}$$

that is, $\mu = \lambda + \alpha$ and $\underline{w}$ are eigenvalue and associated eigenvector of $\mathbf{D}_\alpha$, respectively. By the Perron-Frobenius theorem, there exists an eigenvalue μ^* of $\mathbf{D}_\alpha$ such that $\mu^* > |\lambda + \alpha| > \lambda + \alpha$ for any other eigenvalue $\lambda + \alpha$ of $\mathbf{D}_\alpha$. Moreover, the unit eigenvector $\underline{v}^*$ associated with μ^* has all coordinates positive.

We will show that $\lambda^* = \mu^* - \alpha$. Assume not. Then, there exists an eigenvalue $\mu_0 < \mu^*$ of $\mathbf{D}_\alpha$ such that $\lambda^* = \mu_0 - \alpha$. Thus, $\lambda^* < \mu^* - \alpha = \tilde{\lambda}$, where $\tilde{\lambda}$ is an eigenvalue of $\mathbf{D}$. Since λ^* is the unique positive eigenvalue of $\mathbf{D}$, we obtained a contradiction. Therefore, $\lambda^* = \mu^* - \alpha$ and $\mathbf{D}\underline{v}^* = \lambda \underline{v}^*$.

Consequently, λ^* is the unique simple positive eigenvalue of the matrix $\mathbf{D}$ and the respective eigenspace is spanned by the unit vector $\underline{v}^*$ with all components positive.

Now we are ready to prove the main result.

[Proof of Theorem 2.2] With $A_{i,h}=\frac{N_{i,h}S_{i,h}}{\rho_i}$ and c_i's defined in (15), equation (11) can be written as

$$\sum_{h=1}^{H_i}\frac{A_{i,h}^2}{n_{i,h}}-c_i=T,\quad i=1,\ldots,I. \tag{27}$$

Consequently, the Lagrange function for the minimization problem assumes the form

$$F(T,\underline{n})=T+\sum_{i=1}^{I}\mu_i\left(\sum_{h=1}^{H_i}\frac{A_{i,h}^2}{n_{i,h}}-c_i\right)+\mu\sum_{i=1}^{I}\sum_{h=1}^{H_i}n_{i,h}.$$

Upon differentiating with respect to $n_{i,h}$ we obtain

$$\frac{\partial F}{\partial n_{i,h}}=\mu-\mu_i\frac{A_{i,h}^2}{n_{i,h}^2}=0$$

which yields $v_i^2:=\mu_i/\mu>0$ and $n_{i,h}=v_iA_{i,h}$, $h=1,\ldots,H_i$, $i=1,\ldots,I$.

Since $a_i=\sum_{h=1}^{H_i}A_{i,h}$, see (14), the constraint (27) assumes the form

$$a_i-c_iv_i=Tv_i,\quad i=1,\ldots,I. \tag{28}$$

Moreover, (17) yields $\frac{1}{n}\sum_{j=1}^{I}v_ja_j=1$. Therefore, (28) can be written in the form

$$\frac{1}{n}\left(\sum_{j=1}^{I}v_ja_j\right)a_i-c_iv_i=Tv_i,\quad i=1,\ldots,I.$$

Equivalently, $\mathbf{D}\underline{v}=T\underline{v}$ with $\mathbf{D}=\frac{1}{n}\underline{a}\underline{a}^T-\mathrm{diag}(\underline{c})$, and $\underline{v}=(v_1,\ldots,v_I)^T$. That is, $\underline{v}$ which is a vector with positive components, is an eigenvector of $\mathbf{D}$ and T is the eigenvalue associated to $\underline{v}$. According to Prop. 2.1, the unique unit vector $\underline{v}$ satisfying positivity requirement is $\underline{v}=\underline{v}^*$ and then $T=\lambda^*$. Consequently,

$$n_{i,h}\propto A_{i,h}v_i^*,\quad h=1,\ldots,H_i,\ i=1,\ldots,I.$$

Using again the constraint (17) we obtain (18).

On the other hand, we plug $n_{i,h}$, as given in (18), into the formula for the total relative variance (10). Upon cancelations we get (19).

4. Conclusion

The minimization of the common base T of the relative variances in the domains under domain-wise stratified SRSWOR can be achieved analytically through the eigenproblem approach. The formulas for the allocation as well as for the optimal relative variance are explicit in terms of the unique unit eigenvector with positive coordinates of a properly designed population matrix $\mathbf{D}$. Consequently, a direct (but not straightforward) generalization of the classical Neyman-Tchuprov optimal allocation is obtained. Although it has similar drawbacks to those of the Neyman-

Tchuprov allocation, it has its rather unique advantage: it reveals structural properties of the domain-wise optimal allocation. Additionally, in typical situations, the eigenproblem approach gives also numerical solutions which are either identical or close to those obtained through NLP tools. Of course, the NLP procedures allow one to obtain optimal sample strata sizes not exceeding actual strata sizes. The eigenproblem approach may give optimal allocations which do not satisfy such requirements. The proper adjustment of the eigenproblem methodology remains a challenging issue.

REFERENCES

CHOUDHRY, G. H., RAO, J. N. K., HIDIROGLOU, M. A., (2012).On sample allocation for efficient domain estimation, Survey Meth. 38(1) , pp. 23–29.

FRIEDRICH, U., MÜNNICH, R., DE VRIES, S., WAGNER, M., (2015). Fast integer-valued algorithm for optimal allocations under constraints in stratified sampling, Comp. Statist. Data Anal., 92 , pp. 1–12.

FRIEDRICH, U., MÜNNICH, R., RUPP, M., (2018). Multivariate optimal allocation with box-constraints, Austrian J. Statist., 47 , pp. 33–52.

GABLER, S., GANNINGER, M., MÜNNICH, R., (2012). Optimal allocation of the sample size to strata under box constraints, Metrika, 75(2) , pp. 151–161.

KATO, T., (1981), A Short Introduction to Perturbation Theory for Linear Operators, Springer, New York.

KHAN, M. G. M., WESOŁOWSKI, J., (2019). Neyman-type sample allocation for domains-efficient estimation in multistage sampling. Adv. Stat. Anal., 103 , pp. 563–592.

KOZAK, M., (2004). Method of multivariate sample allocation in agricultural surveys. Biom. Collq., 34 , pp. 241–250.

KOZAK, M., ZIELIŃSKI, A., (2005). Sample allocation between domains and strata. Int. J. Appl. Math. Stat, 3 , pp. 19–40.

KOZAK, M., ZIELIŃSKI, A., SINGH, S., (2008). Stratified two-stage sampling in domains: sample allocation between domains, strata and sample stages. Statist. Probab. Lett., 78 , pp. 970–974.

LONGFORD, N. T., (2006). Sample size calculation for small-area estimation. Survey Meth., 32 , pp. 87–96.

NIEMIRO, W., WESOŁOWSKI, J., (2001). Fixed precision allocation in two-stage sampling. Appl. Math., 28 , pp. 73–82.

SÄRNDAL, C.-E., SWENSSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling, Springer, New York .

VALLIANT, R., DEVER, J.A., KREUTER, F., (2013). Practical Tools for Designing and Weighting Sample Surveys, Springer.

WESOŁOWSKI, J., WIECZORKOWSKI, R., (2017). An eigenproblem approach to optimal equal-precision sample allocation in subpopulations. Comm. Statist. Theory Meth., 46(5) , pp. 2212–2231.

WRIGHT, T., 2017). Exact optimal sample allocation: More efficient than Neyman. Statist. Probab. Lett., 129 (, pp. 50–57.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 13–31, DOI 10.21307/stattrans-2019-032
Submitted – 02.07.2019; Paper ready for publication – 12.11.2019

GENERAL LINEAR MODEL: AN EFFECTIVE TOOL FOR ANALYSIS OF CLAIM SEVERITY IN MOTOR THIRD PARTY LIABILITY INSURANCE

Erik Šoltés[1], **Silvia Zelinová**[2], **Mária Bilíková**[3]

ABSTRACT

The paper focuses on the analysis of claim severity in motor third party liability insurance under the general linear model. The general linear model combines the analyses of variance and regression and makes it possible to measure the influence of categorical factors as well as the numerical explanatory variables on the target variable. In the paper, simple, main and interaction effects of relevant factors have been quantified using estimated regression coefficients and least squares means. Statistical inferences about least-squares means are essential in creating tariff classes and uncovering the impact of categorical factors, so the authors used the LSMEANS, CONTRAST and ESTIMATE statements in the GLM procedure of the Statistical Analysis Software (SAS). The study was based on a set of anonymised data of an insurance company operating in Slovakia; however, because each insurance company has its own portfolio subject to changes over time, the results of this research will not apply to all insurance companies. In this context, the authors feel that what is most valuable in their work, is the demonstration of practical applications that could be used by actuaries to estimate both the claim severity and the claim frequency, and, consequently, to determine net premiums for motor insurance (regardless of whether for motor third party liability insurance or casco insurance

Key words: general linear model, claim severity, motor third party liability insurance, least squares means.

1. Introduction

In general, two approaches are used to determine net premiums in non-life insurance. Either the target variable is equal to the net premium (euros of loss per exposure) or it is separately modelled the claims frequency (number of claims per

[1] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Statistics. E-mail: erik.soltes@euba.sk. ORCID ID: https://orcid.org/0000-0001-8570-6536.

[2] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Mathematics and Actuarial Science. E-mail: silvia.zelinova@euba.sk. ORCID ID: https://orcid.org/0000-0002-9932-6857.

[3] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Mathematics and Actuarial Science. E-mail: maria.bilikova@euba.sk.

exposure) and the claim severity (average loss per claim). Goldburd et al. (2016) mention that special modelling of frequency and severity is more stable and leads to a lower variance of the error term compared to when the net premium is directly modelled. In addition, in the case of a separate analysis of frequency and severity we can detect effects in the data that we otherwise would not. On the other hand, the standard techniques of net premium determination based on specific modelling of frequency and severity assume independence between the number and the size of claims. Methods that are appropriate in the case of correlation between frequency and severity components are dealt with by, e.g. Shi et al. (2015). The above facts motivated us to consider a separate modelling, so the paper focuses only on the claim severity in motor third party liability (MTPL) insurance. Since severity refers to the cost of a claim, through this metric we can identify those tariff classes in MTPL insurance which are more expensive and those which are cheaper for an insurance company.

For the calculation of auto insurance premiums, many actuaries use techniques based on regression analysis and analysis of variance in their scientific work. Very popular models include generalized linear models, which are used by, e.g. (De Azevedo et al., 2016), (Kafková and Křivánková, 2014), (Jong and Heller, 2008) and (Frees et al., 2016). The Poisson regression model is frequently used to model claim frequency and the Gamma regression model is used to model claim costs (see, e.g. (David, 2015) and (Duan et al., 2018)). As David (2015) indicates, generalized linear models allow for the modelling of a non-linear behaviour and a non-Gaussian distribution of residuals, which is very useful for the analysis of non-life insurance, where claim frequency and claim cost follow an asymmetric density, which is clearly non-Gaussian. A special case of generalized linear model (GzLM) is the general linear model (GLM), which we use in the article to assess the impact of relevant factors on claim severity. GLM and GzLM are two commonly used families of statistical methods to relate some number of continuous and/or categorical predictors to a single outcome variable. The main difference between the two approaches is that GLM strictly assumes that the residuals will follow a conditionally normal distribution, while GzLM loosens this assumption and allows for a variety of other distributions from the exponential family for the residuals (see, e.g. (Agresti, 2015), (Fox, J., 2015), (Kim and Timm, 2006) and (Littell, et al., 2010)).

GLM includes the t-test, analysis of variance (ANOVA), multiple regression, descriptive discriminant analysis (DDA), multivariate analysis of variance (MANOVA), canonical correlation analysis (CCA) and structural equation modelling (SEM). Therefore, Graham (2008) indicates that the vast majority of parametric statistical procedures in common use are part of the general linear model. Thompson (2015) discusses GLM as a unifying conceptual framework that helps students and researchers understand common features of analyses included in GLM.

The aim of the article is to provide a presentation of the possibility of using general linear models for claim severity analysis in motor third party liability insurance for the purpose of tariffication. The article does not limit itself to an illustration of general linear models by means of a demonstrational example but provides the analysis of an actual data set from an unnamed insurance company operating in Slovakia.

In the past, actuaries often relied on a one-way analysis of pricing. However, one-way analyses do not consider interdependencies between factors in the way they affect claim experience, which is why multivariate methods are more effective (Anderson et al., 2007). For this reason, in this paper we use multivariate methods included in general linear models, which correct the correlation between factors and allow for the investigation of interaction effects.

2. Research methods

The general linear model, which will be the subject of interest in our paper, can be simplified as follows:

$$y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\mu_{ij}} + \varepsilon_{ijk} \tag{1}$$

where y_{ijk} is *k*-th observation of the target (explained) variable *Y* in cell ij , i.e. at the *i*-th level of factor *A* and at the same time the *j*-th level of factor *B*. We assume that the random errors ε_{ijk} are independent of each other and identically distributed with the normal distribution $N(0,\sigma^2)$.

Let us denote by μ_{ij} the mean of the target variable for the *i*-th variation of factor *A* and the *j*-th variation of factor *B*. This mean is called the *cell mean* for cell ij and is defined as the sum of the constant μ (intercept), α_i - factor *A* effect, β_j - factor *B* effect and $(\alpha\beta)_{ij}$ – the interaction effect between factors *A* and *B*. Note that more than two factors will be taken into account in the application part of this paper, some will be in the form of quantitative variables and others in the form of categorical variables.

The general linear model can be used to examine several types of effects, such as:

- *simple effects*, which indicate that one factor level affects the target variable, while other factors remain constant at that level;
- *interactions,* which characterize how levels of one factor affect the target variable across levels of another factor. If, at all levels of 2nd factor, 1st factor affects the target variable equally, it is a non-interaction model. If, at different levels of 2nd factor, 1st factor affects the target variable differently, it is an interaction model;
- *main effects,* which reflect the overall differences between the levels of each factor averaged across all levels of another factor.

The focus should be on interaction and then on simple or main effects. If a significant interaction is confirmed, it is appropriate to compare simple effects. One way to compare the means of the target variable at different levels of one factor specifically for different levels of the second factor is to carry out the analysis of variance or general linear model separately for different levels of the second factor. However, by this separate analysis, we discard some of the information from other levels of the second factor, and this unused information manifests itself in a low number of degrees of freedom for SS (ERROR), which is central to statistical tests associated with the analysis of variance (Littell et al., 2010). This inefficient solution would waste a lot of data, which will severely reduce the strength of the tests. With the tools in the GLM procedure (PROC GLM) of the SAS statistical software, which we use in the paper, it is possible to avoid such a problem.

PROC GLM has options within the LSMEANS statement that allow you to test each factor at a particular level of another factor. The LSMEANS statement calculates the estimate of the so-called least squares mean (LS mean), also referred to as the marginal mean. In unbalanced, multi-way designs, the LS means estimation is often assumed to be closer to reality. LS means correct the design's imbalance. In balanced designs, or in unbalanced one-way ANOVA designs, observed means and least squares means are the same ((Lenth, 2016) and (Cai, 2014)).

The general linear model can be written in the form of a multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \tag{2}$$

PROC GLM for estimating the parameters of such a model, therefore, uses the least squares method, which results in the formula

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{3}$$

In view of the fact that in the GLM procedure generally considered with the classification explanatory variables, which are converted to dummy variables, the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not of full rank and therefore has no unique inverse. For such a situation, PROC GLM computes a general inverse $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}$ and the parameters of the regression model (2) are estimated according the formula

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{4}$$

where the estimated parameter vector $\hat{\boldsymbol{\beta}}$ has zero values at the locations that correspond to the zero rows in the matrix $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}$. The estimate $\hat{\boldsymbol{\beta}}$ thus obtained is not unique. However, there is a set of linear functions $\mathbf{L}\hat{\boldsymbol{\beta}}$ where $\mathbf{L}$ is a linear combination of rows of the matrix $\mathbf{X}$, which are called estimable functions (more detail in (Agresti, 2015, pp. 14–15) and (Littell et al., 2010, pp. 194–203)) and have these features:

- $\mathbf{L}\hat{\boldsymbol{\beta}}$ and its covariance matrix $Var\left(\mathbf{L}\hat{\boldsymbol{\beta}}\right)$ are unique,
- $\mathbf{L}\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\mathbf{L}\boldsymbol{\beta}$.

As with the full rank, covariance matrix $\mathbf{L}\hat{\boldsymbol{\beta}}$ is given by the formula

$$Var\left(\mathbf{L}\hat{\boldsymbol{\beta}}\right) = \sigma_{\varepsilon}^{2}\left[\mathbf{L}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{L}^{\mathrm{T}}\right] \tag{5}$$

wherein the estimate of the variance of the random error σ_{ε}^{2} is the residual variance MSE, which is calculated similarly to the multiple regression analysis, while the sum of squared error SSE (also known as the sum of squared residuals – SSR) is no longer dependent on the general inverse $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}$.

3. Preparation of input variables, selection of regressors and verification of assumptions about the error term

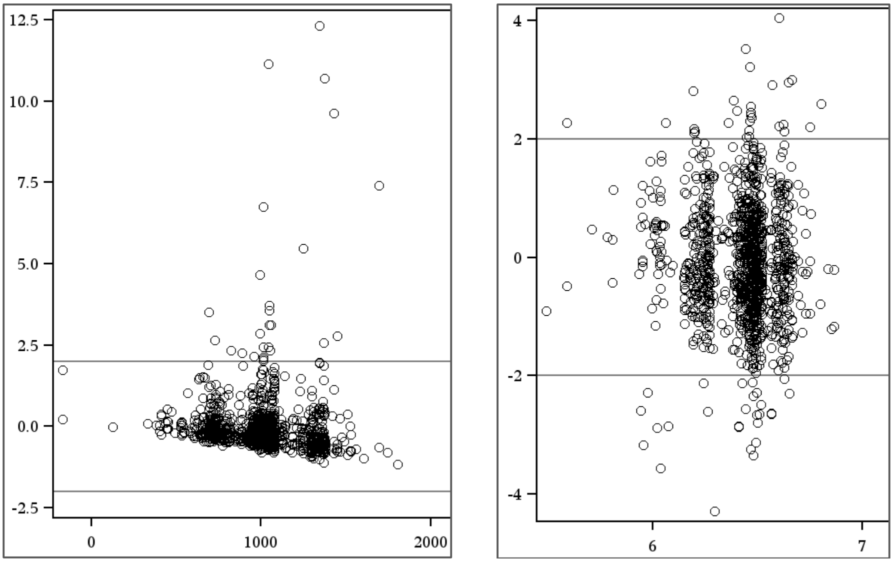
Our analysis focuses on the target variable – the claim severity (average costs per claim) of passenger cars in MTPL insurance. We modelled this variable depending on the following factors:

- relating to the insured vehicle such as Engine Power (kW, abbr. EP), Engine Volume (cm^3), Weight (kg), Age (years) and Car Make,
- relating to the vehicle owner such as Age (years) and Residence.

We categorized the vehicle owner's age and created the Age_group variable, which has six groups: the vehicle owners aged up to 30, aged 30–40, 40–50, 50–60, 60–70 and over 70 (upper limits of the indicated intervals are closed).

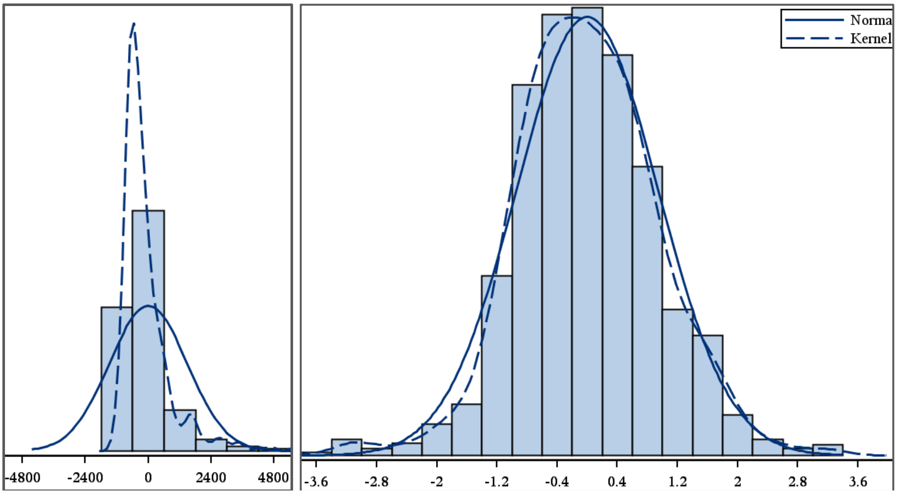
Since the residuals showed heteroscedasticity (Figure 1, on the left) and were markedly right-skewed (Figure 2, on the left) while modelling claim severity, we decided to use the logarithmic transformation of the explained variable. In the log-linear model, which modelled the dependence on the factors considered, the residuals had approximately a normal distribution with zero mean (see (Figure 1, on the right) and (Figure 2, on the right)).

Figure 1. Studentized residuals vs predicted for claim severity (on the left) and vs predicted for logarithm of claim severity (on the right)



Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

Figure 2. Distribution of residuals for claim severity (on the left) and for logarithm of claim severity (on the right)



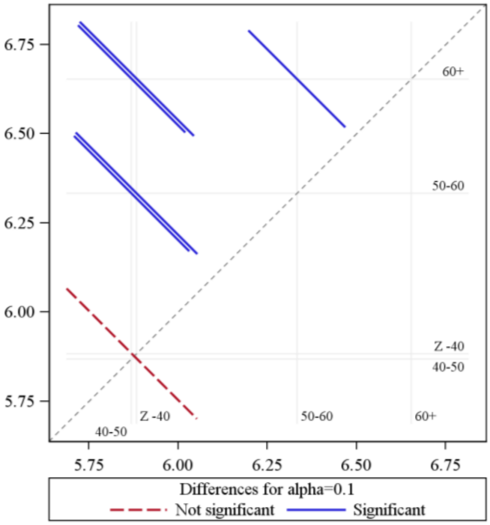
Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

We verified the homoscedasticity of the error term by using the White test. This test uses a model of the second squares of residuals depending on the predicted values and their squares, while the original test is based on the model of squared residuals depending on the original explanatory variables, the squares and cross products of independent variables (see (Wooldridge, 2013, pp. 279–280)). Based on calculated test statistics $\chi^2 = 3.3376$, which had 2 degrees of freedom, we quantified $p - value = 0.1885$. Since $p - value$ is greater than any commonly used confidence level, we do not reject the null hypothesis of homoscedasticity.

Based on the average amount of claims incurred in fixing the other considered factors, we transformed some of the original explanatory variables during the modelling process. We created three groups of vehicle makes and we call the resulting variable in other analyses Vehicle_group. This categorical variable has values 1, 2 and 3, with category 1 being the makes of vehicles with the highest average costs per claim, and category 3 including vehicle makes, where we quantified the lowest average costs per claim (in eliminating the influence of other factors). Similarly, we developed new categories of the Residence variable, using the LS means tests below. This process created a classification variable with 3 values (A, B and C), with category A (including the regional cities of Košice and Trenčín), where we detected the highest average costs per claim, category B (including villages, small towns, all district towns as well as the regional cities of Bratislava and Nitra) and category C (including the regional cities of Banská Bystrica, Prešov, Trnava and Žilina), where we quantified the lowest average costs per claim. We have to emphasize that statistically significant differences in average costs per claim were not confirmed among the owners' residence that fall within the same category.

We included the variables of Engine Power (EP), Engine Volume, Weight, Age, Vehicle_group, Age_group and Residence, as well as the polynomials of numerical explanatory variables, but also the interaction between the considered variables. By the method of backward elimination (Agresti, 2015), factors that did not have a significant effect on the explained variable at the confidence level 0.1 were excluded from the model. At the same time, the equality of the marginal means (LS-means) were tested using the Tukey-Kramer test (adjusted Tukey's test appropriate for unbalanced data, see (Wilcox, 2003)). In the case of insignificant differences between the marginal means of the target variable on two levels of one particular factor and after taking into account the logical context we finally merged the original categories of that factor.

Figure 3. Comparison of LS means of logarithm of claim severity for factor Age_group



Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

In short, we will explain this procedure with the factor of Age_group. This factor originally contained 6 categories, but because of the insignificant differences in average severity between insured aged 70+ and 60–70, we merged these categories to form a category 60+. Similarly, we proceeded in the same way in the case of age categories 30-40 years and up to 30 years. However, we must remark that in the 70+ and under 30 age groups, the insurance company had a low number of claims and, therefore, based on the input database, we cannot persuasively claim that young or old vehicle owners (over 70) do not have higher or lower average severity compared to the other age categories. In the case of the Age_group factor, the next analysis found that insured persons aged 40 to 50 and under 40 report the smallest average severity, with no significant differences between these two age categories, as shown in Figure 3.

We merged these two categories and present the category of those aged up to 50 in the following results.

4. Empirical results

In this section of the paper, we will provide the results of the analysis obtained from the PROC GLM of the SAS statistical software. In Section 4.1 we will focus on assessing the differences between the individual levels of the competent relevant factors and quantifying the impact of these factors on the average claim severity of vehicles in MTPL insurance. Section 4.2 offers examples of the application of the CONTRAST and ESTIMATE statements that actuaries and statisticians can use for further analyses of the impact of the factors on the target variable.

4.1. Estimating the model and quantifying the impact of relevant factors

As we mentioned in the previous section, the method of backward elimination was used to select regressors, in which the statistical significance of a particular factor was assessed by the F-test, which uses the partial sum of squares, called Type II SS in regression analysis, but Type III SS in the GLM procedure (see more in (Kuznetsova et al., 2017), (LaMotte, 2019) and (Littell et al., 2010)). This sum of squares for the particular variable represents an increase in SSM due to the addition of this variable to the model. This type of sum of squares does not depend on the sequence in which the independent variable is loaded into the model and is useful to verify the statistical significance of the effect of the analysed explanatory variable on the target variable *Y*. Table 1 confirms the significance of the influence of the factors left in the resulting model.

Table 1. Verifying the impact of considered factors on claim severity

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EP	1	0.86485345	0.86485345	6.72	0.0096
EP * EP	1	1.03197187	1.03197187	8.02	0.0047
EP * EP * EP	1	1.28920833	1.28920833	10.01	0.0016
Age_group	2	2.87135126	1.43567563	11.15	<.0001
Vehicle group	2	1.82409264	0.91204632	7.08	0.0009
Residence	2	3.26944802	1.63472401	12.70	<.0001
Age_group*Residence	4	2.43942634	0.60985659	4.74	0.0008

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

The regression coefficients (Table 2) of the dummy variables, which encode the categories of Age_group, Vehicle group and Residence, are statistically significant at the 0.1 confidence level. In the above categories, the average severity of insurance claim is significantly different from the reference category of the relevant factor (at the level of confidence of 0.1). Figures 4 and 5 confirm that not only in comparison with the reference category, but among all pairs of particular factor categories there are significantly different LS means of the target variable at the 0.1 confidence level. The highest average severity when fixing other factors was found for the oldest vehicle owners (in our case over the age of 60), then in the owners of vehicles from the regional cities of Trenčín and Košice and in the makes of vehicles belonging to group 1. On the contrary, we found the lowest average severity under ceteris paribus conditions in the group of vehicle owners aged under 50, as well as in vehicles of the group 3 and for vehicles from the regional cities of Banská Bystrica, Prešov, Trnava and Žilina.

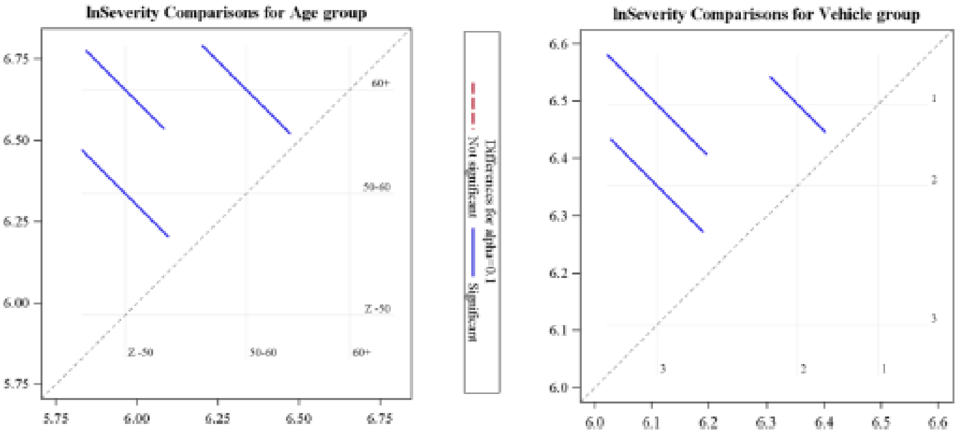
Table 2. Estimate of the parameters of general model for natural logarithm of claim severity

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	3.1170	B	0.8205	3.80	0.0002
EP	0.0907		0.0350	2.59	0.0096
EP * EP	-0.0014		0.0005	-2.83	0.0047
EP * EP * EP	6.7E-6		0.0000	3.16	0.0016
Age_group 60+	1.0729	B	0.2958	3.63	0.0003
Age_group 50-60	0.8390	B	0.3001	2.80	0.0052
Age_group -50	0.0000	B	.	.	.
Vehicle group 1	0.3854	B	0.1065	3.62	0.0003
Vehicle group 2	0.2439	B	0.0993	2.46	0.0141
Vehicle group 3	0.0000	B	.	.	.
Residence A	1.0997	B	0.3044	3.61	0.0003
Residence B	1.1402	B	0.2380	4.79	<.0001
Residence C	0.0000	B	.	.	.
Age_group*Residence 60+ A	-0.1156	B	0.4361	-0.27	0.7910
Age_group*Residence 60+ B	-1.0213	B	0.3029	-3.37	0.0008
Age_group*Residence 60+ C	0.0000	B	.	.	.
Age_group*Residence 50-60 A	-0.6337	B	0.4865	-1.30	0.1929
Age_group*Residence 50-60 B	-0.7596	B	0.3060	-2.48	0.0132
Age_group*Residence 50-60 C	0.0000	B	.	.	.
Age_group*Residence -50 A	0.0000	B	.	.	.
Age_group*Residence -50 B	0.0000	B	.	.	.
Age_group*Residence -50 C	0.0000	B	.	.	.

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

The interaction between the factors Age_group and Residence showed to be statistically significant. Based on the LS means tests (Figure 5, on the right) for pairs of vehicle owner groups that arose from breaking down by the two mentioned factors, we found that not all pairs report different average severities. It is clear that because of the interaction of Age_group and Residence factors it is significantly the highest claim severity in the case of vehicle owners who live in the villages falling into category A and at the same time are aged 60+. On the other hand, the general linear model quantified that the lowest average severity is among the group of vehicle owners under the age of 50 who live in the regional cities of Banská Bystrica, Prešov, Trnava and Žilina.

Figure 4. Comparison of LS means for factor Age_group (on the left) and for factor Vehicle group (on the right)

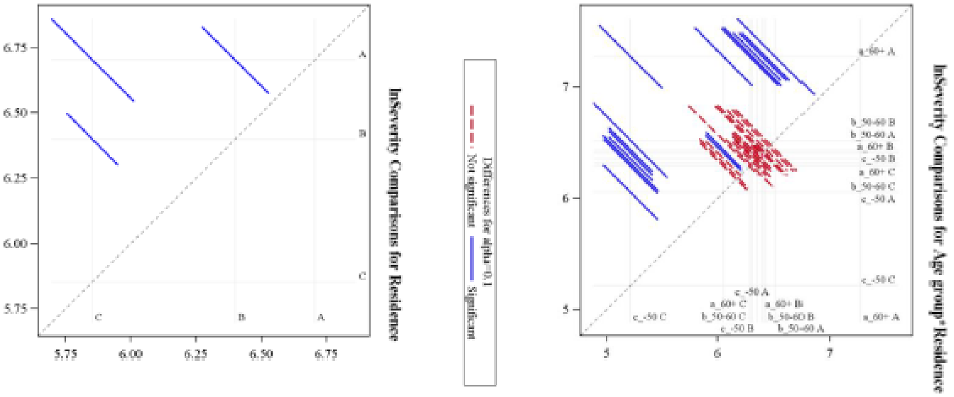


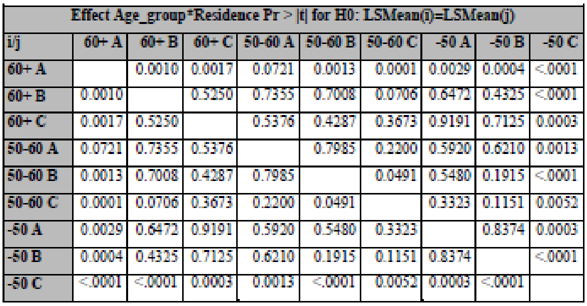
Least Squares Means for effect Age_group Pr > |t| for H0: LSMean(i)=LSMean(j)			
i/j	60+	50-60	-50
60+		0.0542	<.0001
50-60	0.0542		0.0221
-50	<.0001	0.0221	

Least Squares Means for effect Vehicle group Pr > |t| for H0: LSMean(i)=LSMean(j)			
i/j	1	2	3
1		0.0174	0.0003
2	0.0174		0.0141
3	0.0003	0.0141	

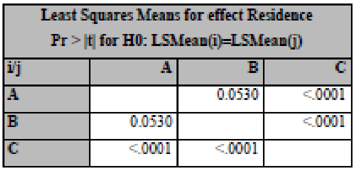
Source: Unnamed insurance company, self-processed in SAS Enterprise Guide

Figure 5. Comparison of LS means for factor Residence (on the left) and for interaction Age_group×Residence (on the right)





Effect Age_group*Residence Pr > |t| for H0: LSMean(i)=LSMean(j)									
i/j	60+ A	60+ B	60+ C	50-60 A	50-60 B	50-60 C	-50 A	-50 B	-50 C
60+ A		0.0010	0.0017	0.0721	0.0013	0.0001	0.0029	0.0004	<.0001
60+ B	0.0010		0.5250	0.7355	0.7008	0.0706	0.6472	0.4325	<.0001
60+ C	0.0017	0.5250		0.5376	0.4287	0.3673	0.9191	0.7125	0.0003
50-60 A	0.0721	0.7355	0.5376		0.7985	0.2200	0.5920	0.6210	0.0013
50-60 B	0.0013	0.7008	0.4287	0.7985		0.0491	0.5480	0.1915	<.0001
50-60 C	0.0001	0.0706	0.3673	0.2200	0.0491		0.3323	0.1151	0.0052
-50 A	0.0029	0.6472	0.9191	0.5920	0.5480	0.3323		0.8374	0.0003
-50 B	0.0004	0.4325	0.7125	0.6210	0.1915	0.1151	0.8374		<.0001
-50 C	<.0001	<.0001	0.0003	0.0013	<.0001	0.0052	0.0003	<.0001	



Least Squares Means for effect Residence Pr > |t| for H0: LSMean(i)=LSMean(j)			
i/j	A	B	C
A		0.0530	<.0001
B	0.0530		<.0001
C	<.0001	<.0001	

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

In order to quantify the impact of various factors on the average severity it is necessary to convert the estimate of the model $\ln \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$ shown in Table 2 into the form $y_i = e^{\hat{\beta}_0} \cdot \left(e^{\hat{\beta}_1}\right)^{x_{i1}} \cdot \left(e^{\hat{\beta}_2}\right)^{x_{i2}} \cdot \ldots \cdot \left(e^{\hat{\beta}_k}\right)^{x_{ik}}$. Naturally, in the additive model, the influence of reference categories is at the "0" level, which is transformed into a value $e^0 = 1$ in the multiplicative model. Based on the above transformation, using the parameter estimates from Table 2, we get

$$\hat{y}_i = 22.579 \cdot (1.095)^{EP} \cdot 0.9986^{EP^2} \cdot 1.000067^{EP^3} \cdot 2.924^{Age\ Group=60+} \cdot 2.314^{Age\ Group=50-60} \cdot$$

$$\cdot 1.470^{Vehicle\ Group=1} \cdot 1.276^{Vehicle\ Group=2} \cdot 3.003^{Residence=A} \cdot 3.127^{Residence=B} \cdot$$

$$\cdot 0.891^{Age\ Group=60+\wedge Residence=A} \cdot 0.360^{Age\ Group=60+\wedge Residence=B} \cdot$$

$$\cdot 0.531^{Age\ Group=50-60\wedge Residence=A} \cdot 0.468^{Age\ Group=50-60\wedge Residence=B}$$

The shape of the function with EP (Engin Power) as an explanatory variable shows that with a normal engine power of between 50 and 100 kW, the claim severity is approximately constant while fixing other factors and starts to rise more quickly for vehicles with engine power over 100 kW. In the case of vehicle makes falling under category 1, we estimate an almost 1.5 times higher average severity than for the 3rd category of vehicles and about 15% higher (1.152 = 1.470 / 1.276) than for the 2nd category of vehicles.

Since there is an interaction between the factors Age_group and Residence, the influence of these factors can be quantified from the exponential bases of the dummy variables belonging to the variables Age_group, Residence and their interactions Age_group × Residence.

Table 3. Multiplier estimates for vehicle owners broken down by-Age_group and Residence factors

Age_group	Residence		
	A	B	C
60+	7.822	3.293	2.924
50-60	3.688	3.386	2.314
-50	3.003	3.127	1.000

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

It is clear from Table 3 that the highest average loss per claim is for vehicle owners over the age of 60 who live in the municipalities of group A (regional towns of Trenčín and Košice). The fact that it is the riskiest group from the point of view of claim severity was already confirmed in Figure 5 (on the right). Now, we have found that their average severity is up to 7.8 times higher than in the case of the least risky group, which is vehicle owners under the age of 50 living in villages in category C (regional towns Banská Bystrica, Prešov, Trnava and Žilina). Similarly, the other multipliers estimated in Table 3 could also be interpreted as compared to the "-50 C" reference category.

4.1. Use of the CONTRAST and ESTIMATE statements for a deeper analysis of the impact factors

According to Table 3 and the estimated LS means, it appears that in the age group of vehicle owners aged 50 to 60, residence has little impact on average severity. In the age group 50-60 years, between the residence categories A and B, based on the LS means test (Figure 5, $p-value=0.7895$), it did not confirm the significant difference and thus we can assume equality $H_0:\mu_{2A}=\mu_{2B}$. Note that for ease of writing, we will use index 2 to denote the second variation of the Age_group variable (50-60 years). In order to verify the equality of the corresponding 3 means $H_0:\mu_{2A}=\mu_{2B}=\mu_{2C}$, we will test the hypothesis

$$H_0:\left(\mu_{2A}+\mu_{2B}\right)/2=\mu_{2C} \quad \text{or equivalently} \quad H_0:0.5\mu_{2A}+0.5\mu_{2B}-\mu_{2C}=0$$

We will verify this hypothesis in the SAS software with the CONTRAST statement, using Table 4 to determine the coefficients in this statement.

Table 4. Coefficients to the CONTRAST statement to verify the hypothesis

$$H_0 : 0.5\mu_{2A} + 0.5\mu_{2B} - \mu_{2C} = 0$$

Age_group	**Residence**			**Σ**
	A	**B**	**C**	
1=60+	0	0	0	**0**
2=50-60	0.5	0.5	-1	**0**
3=-50	0	0	0	**0**
Σ	**0.5**	**0.5**	**-1**	**0**

Source: Self-processed.

Then the statement has a syntax

```
contrast 'Age_group*Residence 2A 2B vs 2C' Residence 0.5 0.5 -1
Age_group*Residence 0 0 0 0.5 0.5 -1;
```

The result of the test is given in the first row in the body of Table 5. Depending on the level of confidence, we reject or do not reject the null hypothesis. If we take into account the level of confidence of 0.05, we do not reject the null hypothesis that the average severity of vehicle owners aged 50-60 in categories A and B is the same as that of the residents aged 50-60 in category C. However, at a confidence level of 0.1, we reject this null hypothesis.

A more correct way to verify the hypothesis $H_0 : \mu_{2A} = \mu_{2B} = \mu_{2C}$ is by simultaneously testing hypotheses

$$H_0 : \mu_{2A} = \mu_{2B} \quad \text{and} \quad H_0 : (\mu_{2A} + \mu_{2B})/2 = \mu_{2C}$$

To verify these two hypotheses, we use the CONTRAST statement in the form

```
contrast 'Age_group*Residence 2A=2B=2C'
Residence 1 -1  Age_group*Residence 0 0 0 1 -1,
Residence 0.5 0.5 -1  Age_group*Residence 0 0 0 0.5 0.5 -1;
```

The result of the simultaneous testing of the two mentioned null hypotheses is an F-test statistic with degrees of freedom 2 for the nominator, which is also shown in 2nd row of the body of Table 5. Remember that degrees of freedom for the denominator correspond to the degrees of freedom SSE.

Table 5. Results of the CONTRAST statement

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Age_group*Residence 2A 2B vs 2C	1	0.36714636	0.36714636	2.85	0.0915
Age_group *Residence 2A=2B=2C	2	0.51349173	0.25674587	1.99	0.1365

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

Based on simultaneous testing, we find that even at the confidence level of 0.1 residence has no significant impact on the average severity in the age category of people aged 50 to 60. Given the insignificant differences, an insurance company may be interested in the degree of impact on the average severity when the insured person is aged 50-60 (on average over all residences). We can estimate this by using the ESTIMATE statement using Table 6.

Table 6. The coefficients for the ESTIMATE statement to estimate the mean $E\left(\mu_{2A}, \mu_{2B}, \mu_{2C}\right)$

Age_group	Residence			Σ
	A	B	C	
1=60+	0	0	0	**0**
2=50-60	1	1	1	**3**
3=-50	0	0	0	**0**
Σ	**1**	**1**	**1**	**3**

Source: Self-processed.

The values in the body of Table 6 correspond to the coefficients of the means μ_{2A}, μ_{2B} and μ_{2C} in the required formula $\left(\mu_{2A}+\mu_{2B}+\mu_{2C}\right)/3$. These coefficients are then taken as coefficients for interaction. The values in the sum column and the sum row are used as coefficients for the effects of factors A and B, and the sum value in the lower right corner represents the coefficient for the intercept. In order to obtain the required average of the three means, we must use the option DIVISOR = 3 to divide by the value of 3. The required statement then has the form

```
estimate 'Age_group*Residence mean 2A 2B 2C'
intercept 3  Age_group 0 3 0 Residence 1 1 1
Age_group*Residence 0 0 0 1 1 1 / divisor=3;
```

In addition to the point estimate LS mean for vehicle owners aged 50 to 60 (across all residences), the first row of the body of Table 8 provides also the test of significance, i.e. the test result $H_0:(\mu_{2A}+\mu_{2B}+\mu_{2C})/3=0$. In our case, this test is not of great importance, but thanks to the standard error estimate (0.7956) we can easily calculate the interval estimate and possibly verify the hypotheses that may be of interest to the insurance company. Based on the estimated value (4.4479) and its transformation $e^{4.4479}=85.45$, we get a multiplier for those policyholders aged 50 to 60 (including the intercept). Of course, a part of the estimated regression function, which includes the influence of other factors, has to be used to estimate the average severity. In our case it is the factors Engine power and Vehicle_group, whose impact on the average severity we quantified by the expression

$$(1.095)^{EP}\cdot 0.9986^{EP^2}\cdot 1.000067^{EP^3}\cdot 1.470^{Vehicle\ Group=1}\cdot 1.276^{Vehicle\ Group=2}$$

The estimate of the average severity for vehicle owners aged 50 to 60 is calculated so that the value of the above expression is in addition multiplied by the multiplier 85.45. After adjusting for the intercept, i.e. $e^{4.4479}/e^{3.1170}=e^{4.4479-3.1170}=3.7844$, the value $e^{4.4479}$ indicates that policyholders aged between 50 and 60 have an average severity, which is 3.7844 times higher than the reference category, which in our case consists of policyholders under the age of 50 from the regional cities of Banská Bystrica, Prešov, Trnava and Žilina.

If the insurance portfolio of policyholders aged 50 to 60 is 20% residence group A, 50% residence group B and residence group C the reminder, then it is necessary to use the weighted average of μ_{2A}, μ_{2B} and μ_{2C} to calculate the overall mean in the group of policyholders aged 50-60. Therefore, the interaction coefficients in the ESTIMATE statement follow the 2:5:3 ratio, which is captured also in Table 7.

Table 7. The coefficients for the ESTIMATE statement to estimate the mean $E(\mu_{2A},\mu_{2B},\mu_{2C})$ with weights in the ratio 2:5:3

Age_group	**Residence**			**Σ**
	A	**B**	**C**	
1=60+	0	0	0	**0**
2=50-60	0.2	0.5	0.3	**1**
3=-50	0	0	0	**0**
Σ	**0.2**	**0.5**	**0.3**	**1**

Source: Self-processed.

The statement ESTIMATE for the required weight mean has the form

```
estimate 'Age_group*Residence w_mean 2A 2B 2C'
intercept 1 Age_group 0 1 0 Residence 0.2 0.5 0.3
Age_group*Residence 0 0 0 0.2 0.5 0.3;
```

and it generates the output shown in row 2 of the body of Table 8.

Table 8. ESTIMATE statements results

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Age_group*Residence mean 2A 2B 2C	4.4479	0.7956	5.59	<.0001
Age_group*Residence w_mean 2A 2B 2C	4.4492	0.7896	5.63	<.0001

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

Given the fact that between the means μ_{2A}, μ_{2B} and μ_{2C} significant differences were not confirmed (Table 5), the selected weights have a minimal impact on the overall mean μ_2 as indicated by the negligible differences in the point estimates of the LS means, as shown in Table 8.

5. Conclusions

The paper points to the possibilities of using the general linear model to analyse claim severity in motor third party liability insurance. In order to make adequate use of GLM, it was necessary to apply the logarithmic transformation of the explained variable, thereby eliminating the problem of heavy-tailed distribution and heteroscedasticity of error terms. Thus, the analyses presented in the paper are based on a log-linear model, in which the individual components are in an additive formula, which, however, is converted to a multiplicative formula after the backward exponential transformation. This fact needs to be taken into account when interpreting the results.

Our analyses confirmed that engine power and engine volume are strongly correlated, and their impact on claim severity overlaps significantly. By using the backward elimination method, only the engine power was retained from the two variables in the model, which avoided strong multicollinearity that could lead to problems with the interpretation of the results. Including this variable, categorical variables such as the age group and the owner's residence, as well as their interaction and the Vehicle group factor, were left in the model from the set of explanatory variables (listed in Section 3). Due to the fact that our base is an unbalanced multi-factor model, we could not use arithmetic means to compare the differences in claim severity at different levels of the relevant factors, so we used least squares means. By the gradual merging of categories in which comparable LS means of claim severity were estimated, among which there were no statistically significant differences, we created 3 groups of vehicle makes, 3 age categories of vehicle owners and 3 groups of residential cites. The results of our research reveal the impact of the relevant factors on claim severity, which is quantified by multipliers for each category of relevant factor through the exponential transformation of the respective regression coefficients. Since a significant interaction between the Age group and the Residence factors was confirmed, the paper also quantifies the multipliers for the categories that were created by combining the categories of the mentioned two factors.

Our empirical study shows that the claim severity does not change significantly in vehicles with 50 to 100 kW engine power, and a substantial increase occurs only in vehicles with higher power. The highest average severity was found in owners aged over 60 and in the owners from the regional cities Trenčín and Košice. Vehicle owners who were aged over 60 and had permanent residence in Trenčín and Košice showed 7.8-fold higher average severity, with other variables fixed, as compared to owners under the age of 50 living in the regional towns of Banska Bystrica, Prešov, Trnava and Žilina. That age category (up to 50 years) and the category of residence mentioned (Banská Bystrica, Prešov, Trnava and Žilina) are the least risky in terms of claim severity and their combination reduces the risk.

The benefit of the paper is not only empirical results, but the paper also points to the application of the general linear model to create tariff classes, to estimate average severities for these tariff classes and to detect simple and interaction effects of relevant factors. The general linear model provides such findings through model parameter estimates and least squares means, which are directly available in SAS software or which can be quantified using the CONTRAST and ESTIMATE statements.

The paper shows that the general linear model is an effective tool for the modelling of claim severity because it allows us to use quantitative and categorical regressors and their interactions as well. Unlike other methods, GLM provides estimation of the least square means (besides the arithmetic means) of the target variable. Moreover, PROC GLM in software SAS offers the CONTRAST statement, which is very useful to confirm significant differences between tariff classes in motor insurance. The values of these differences can be estimated using the ESTIMATE statement. Due to the possibility of testing several individual statistical hypotheses for LS means and the possibility of simultaneous testing, the GLM procedure is very flexible and proper for the purpose of tariffication in motor insurance. One disadvantage of the general linear model is the assumptions put on the random error. The error term often does not fulfil the assumption about homoscedasticity. In such a case, a researcher can try to use a logarithmic transformation as we did in our analysis presented in the article. If it does not work, we suggest applying generalized linear models, which are more flexible in this aspect.

Tools of the general linear model applied in the paper can be used by actuaries not only in claims severity, but also in claims frequency, and then for the determination of net premiums in motor insurance.

Acknowledgements

The publication was financed from the fund VEGA number 1/0618/17 *Modern tools for modeling and managing life insurance risks* granted by the Ministry of Education, Science, Research and Sport of the Slovak Republic.

REFERENCES

AGRESTI, A., (2015). Foundations of linear and generalized linear models, John Wiley & Sons.

ANDERSON, D., FELDBLUM, S., MODLIN, C., SCHIRMACHER, D., SCHIRMACHER, E., THANDI, N., (2007). A Practitioner's Guide to Generalized Linear Models: A foundation for theory, interpretation and application (3rd ed.), Towers Watson.

CAI, W., (2014). Making Comparisons Fair: How LS-Means Unify the Analysis of Linear Models. SAS Institute Inc. Paper. SA, S060-2014. [online] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.644.7680&rep=rep1&type=pdf [Accessed on 30 April 2019].

DAVID, M., (2015). Auto insurance premium calculation using generalized linear models, Procedia Economics and Finance, 20, pp. 147–156, DOI: https://doi.org/10.1016/S2212-5671(15)00059-3.

DE AZEVEDO, F. C., OLIVEIRA, T. A., OLIVEIRA, A., (2016). Modeling non-life insurance price for risk without historical information, REVSTAT–Statistical Journal, 14(2), pp. 171–192.
Available at: https://www.ine.pt/revstat/pdf/rs160205.pdf [Accessed on 30 April 2019].

DUAN, Z., CHANG, Y., WANG, Q., CHEN, T., ZHAO, Q., (2018). A Logistic Regression Based Auto Insurance Rate-Making Model Designed for the Insurance Rate Reform. International Journal of Financial Studies, 6(1), 18, DOI: https://doi.org/10.3390/ijfs6010018.

FOX, J., (2015. Applied regression analysis and generalized linear models. Sage Publications.

FREES, E., LEE, G., YANG, L., (2016). Multivariate frequency-severity regression models in insurance, Risks, 4(1), 4,
DOI: https://doi.org/10.3390/risks4010004.

GOLDBURD, M., KHARE, A., TEVET, D., (2016). Generalized linear models for insurance rating, Casualty Actuarial Society, CAS Monographs Series, (5).

GRAHAM, J. M., 2008). The general linear model as structural equation modeling, Journal of Educational and Behavioral Statistics, 33(4), pp. 485–506, DOI: https://doi.org/10.3102/1076998607306151.

JONG, DE P., HELLER, G. Z., (2008). Generalized linear models for insurance data, Cambridge University Press.

KAFKOVÁ, S., KŘIVÁNKOVÁ, L., (2014). Generalized Linear Models in Vehicle Insurance. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 62(2), pp. 383–388,
DOI: https://doi.org/10.11118/actaun201462020383.

KIM, K., TIMM, N., (2006). Univariate and multivariate general linear models: theory and applications with SAS. Chapman and Hall/CRC.

KUZNETSOVA, A., BROCKHOFF, P. B., CHRISTENSEN. R. H. B., (2017). lmerTest package: tests in linear mixed effects models. Journal of Statistical Software, 82(13), DOI: https://doi.org/10.18637/jss.v082.i13.

LAMOTTE, L. R., (2019). A formula for Type III sums of squares. Communications in Statistics-Theory and Methods, pp. 1–11,
DOI: https://doi.org/10.1080/03610926.2019.1586933.

LENTH, R. V., (2016). Least-squares means: the R package lsmeans. Journal of statistical software. 69(1), pp. 1–33,
DOI: https://doi.org/10.18637/jss.v069.i01.

LITTELL, C. L., STROUP, W. W., FREUND, R. J., (2010). SAS for Linear Models (4th Revised ed.). North Carolina, USA: SAS Institute.

SHI, P., FENG, X., IVANTSOVA, A., (2015). Dependent frequency–severity modeling of insurance claims. Insurance: Mathematics and Economics, 64, pp. 417–428. DOI: https://doi.org/10.1016/j.insmatheco.2015.07.006.

THOMPSON, B., (2015). The Case for Using the General Linear Model as a Unifying Conceptual Framework for Teaching Statistics and Psychometric Theory. Journal of Methods and Measurement in the Social Sciences, 6(2), pp. 30–41, DOI: https://doi.org/10.2458/azu_jmmss_v6i2_thompson.

WILCOX, R. R., (2003). Applying contemporary statistical techniques, Elsevier.

WOOLDRIDGE, J. M., (2013). Introductory econometrics: A modern approach (5th ed.), Mason: South-Western.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 33–58, DOI 10.21307/stattrans-2019-033
Submitted – 25.02. 2019; Paper ready for publication – 19.10.2019

HYBRID MULTIPLE IMPUTATION IN A LARGE SCALE COMPLEX SURVEY

Humera Razzak[1], **Christian Heumann**[2]

ABSTRACT

Large-scale complex surveys typically contain a large number of variables measured on an even larger number of respondents. Missing data is a common problem in such surveys. Since usually most of the variables in a survey are categorical, multiple imputation requires robust methods for modelling high-dimensional categorical data distributions. This paper introduces the 3-stage Hybrid Multiple Imputation (HMI) approach, computationally efficient and easy to implement, to impute complex survey data sets that contain both continuous and categorical variables. The proposed HMI approach involves the application of sequential regression MI techniques to impute the continuous variables by using information from the categorical variables, already imputed by a non-parametric Bayesian MI approach. The proposed approach seems to be a good alternative to the existing approaches, frequently yielding lower root mean square errors, empirical standard errors and standard errors than the others. The HMI method has proven to be markedly superior to the existing MI methods in terms of computational efficiency. The authors illustrate repeated sampling properties of the hybrid approach using simulated data. The results are also illustrated by child data from the multiple indicator survey (MICS) in Punjab 2014.

Key words: complex surveys, high-dimensional data, missing data, multiple imputation.

1. Introduction

Large scale complex surveys contain high-dimensional data with a large number of variables measured on an even larger number of respondents. The Multiple Indicator Cluster Surveys (MICS) is such a popular large scale international household survey. Like other cross-sectional surveys, the data sets from MICS contain complex survey features (e.g. many categorical variables). Missing values are also a problem in MICS surveys. Missing data problems arise when a sampled unit does not respond to the entire survey (also called unit non response) or to a particular question (also called item non response). For example, the MICS Punjab 2014 child data set contains more than 200 child health background variables on 31083 children under the age of 5. Among all

[1] humera.razzak@stat.uni-muenchen.de.
[2] christian.heumann@stat.uni-muenchen.de.

these variables, the missing data rates per variable range from 10% to 95% and 26 variables have more than 50% missing values. Questions related to a child cleaning utensils or washing clothes and physical punishment, etc. may make participants reluctant to provide full information, which results in incomplete data (Akmatov (2011)) (Cappa and Khan (2011)).

In recent decades, considerable efforts have been made into the development of statistical methods to treat the problem of missing data. Complete-case or available-case analysis, or single imputation methods such as mean and regression imputation, often result in potentially biased estimates when applied to incomplete data (Anderson et al. (1983)). Rubin (1987) proposed multiple imputation (MI) as an appropriate alternative under certain assumptions. Predictive distributions are used to draw repeated samples in order to simulate values for missing data. *M>1* complete data sets are generated and point and variance estimates of interest are estimated and combined using the formulas developed by Rubin (1987). One advantage of MI is the decoupling of the imputation task and the analysis task although one has to be careful in choosing the imputation and the analysis model (Xie et al. (2017)).

In this paper, we propose a computationally efficient and an easy to implement 3-stage Hybrid Multiple Imputation (HMI) approach to impute complex survey data sets that contain both continuous and categorical variables. The HMI approach applies sequential regression MI techniques to impute continuous variables by using information of categorical variables already imputed by a non-parametric Bayesian MI approach. This blended version of joint and sequential modelling MI techniques makes it possible to obtain complete datasets with both types of variables. This approach is motivated by missing values in background variables related to children's life and health in MICS. In order to get valid and accurate results, it becomes important to impute all types of variables in MICS. As we noted earlier, handling mixed continuous and categorical data in high dimensions presents unique challenges to MI. Existing MI methods can be difficult to implement in the presence of complex dependence structures among categorical variables, whereas some recently developed methods focus on missing values of few variables (Zhao and Long (2016)). Moreover, various MI techniques are limited to categorical variables or require transformations (or other tricks) for continuous variables (Si and Reiter (2013)).

The reminder of the paper is organized as follows. We begin in Section 2 by describing missing data mechanisms. In Section 3, we review imputation methods dedicated to categorical, continuous and mixed data in high dimensions. In section 4 we illustrate Rubin's inference and various estimates used for comparing the performance of the imputation algorithms. Section 5 presents the proposed hybrid architecture. In Section 6 we present the simulation studies and relevant results to evaluate our proposed approach. Section 7 presents the imputation of the MICS Child Data. We conclude with a discussion at the end.

2. Missing data mechanisms

There are three missing data mechanisms. Missing values in any data can be missing completely at random (MCAR), or missing at random (MAR), or missing not at random (MNAR) (Rubin (1987)), (Little and Rubin (2002)). Let Y denote the n × p data matrix with n rows (cases) and p variables. Let y_{ij} refer to the value in row i and column j of Y, where i=1,…,n and j=1,…,p. Further, suppose that there are two components of the data set $Y = \{Y^{miss}, Y^{obs}\}$ where the first component denotes the observed part of the data and the second component is the missing data. Let U be a response indictor matrix with the same dimensions as Y indicating whether an element of Y is observed or missing:

$$U_{ij} = \begin{cases} 0 \ \ if\ y_{ij}\ is\ \ missing, \\ 1 \ \ if\ y_{ij}\ is\ \ observed. \end{cases}$$

Data is MCAR when $Pr(U|Y^{miss}, Y^{obs})$=$Pr(U)$, MAR when $Pr(U|Y^{miss}, Y^{obs})$=$Pr(U|Y^{obs})$ and MNAR when $Pr(U|Y^{miss}, Y^{obs}) \neq \ Pr(U|Y^{obs})$ (Little and Rubin (2002)). MNAR is also called “non-ignorable” (NI).

3. Imputation methods for large scale complex surveys

A complete overview of the state of the art MI methods for accommodating nonlinear relationships and best ways to impute categorical and non-normal continuous variables is given in Vermunt et al. (2008), Yucel et al. (2011), Lee et al. (2012), Seaman et al. (2012) and Lee and Carlin (2017). Information on missing categorical data can be obtained by log-linear models (Schafer (1997)).

Imputation of large scale survey data can become challenging due to the presence of irregular missing patterns, interdependent logical constraints and data inconsistencies. There exist several approaches for MI for high-dimensional data. For example, in hot-deck imputation, which replaces missing values with observed values of pre-defined “donor” cells (Marker et al. (2002)), the probabilities of donor selection can be modified by respondent sampling weights (Andridge (2009)), or a k nearest neighbours (KNN) MI approach based on the distance metric for high-dimensional data (Holder (2015)) may be used or a principal component method to impute missing values (Audigier et al. (2016)). But most of the existing methods are not designed to handle mixed data (quantitative and categorical), become difficult to implement in the situation of large dimensions and are extremely time-consuming (Erosheva et al. (2002)). Moreover, the presence of complex dependence structures can also lead to biased estimates (Wirth and Tchetgen (2014)).

Sequential regression models (Raghunathan et al. (2001)) or fully conditional specification (FCS) (Su et al. (2011)), (van Buuren and Oudshoorn (1999)) is another general approach for MI. It is an iterative process. It specifies univariate conditional distributions on a variable-by-variable basis, and it draws missing values iteratively from the specified conditional distributions. FCS is also known as MI by chained equations (MICE) (Raghunathan et al. (2001)), (van Buuren and Groothuis-Oudshoorn (2011)), (White et al. (2011)), (Su et al. (2011)). The researcher can choose a suitable regression model for each incomplete variable where all the other variables are included as predictor variables, and a suitable imputation method, e.g. predictive mean matching (PMM) (Morris et al. (2014)). Examples are a linear regression model for a continuous variable or a logistic regression model for a binary variable. Also, classification and regression trees (CART; Breiman (2001)) can be used. Vermunt et al. (2008) and van Buuren (2007) applied FCS to impute a small number of categorical and continuous variables. The theoretical implementation of this approach may become challenging when specified conditional densities become incompatible due to high dimensions (White et al. (2011)). Chained equations, when implemented by default settings (i.e. ignoring interaction effects in the conditional models) can also result in biased estimates. Moreover, standard MICE methods cannot handle high-dimensional data (Deng et al. (2016)). Sometimes problems of convergence and incompatibility arise when MICE is used to specify univariate conditional distributions (Arnold and Press (1989)), (Gelman and Speed (1993)) and due to the presence of complex dependencies, implementation of MICE may fail. Similar to log-linear models, conditional models in MICE suffer from model selection and estimation problems in high dimensions, which makes the regression imputations very time-consuming.

Random forest imputation is a method for handling missing data (Stekhoven et al. (2012)). Random forest imputation is a machine learning technique for nonlinearity and interaction problems and does not require a particular model to be specified. Shah et al. (2014) used random forest imputation for imputing complex epidemiological data sets. They found that MI based on random forest techniques tends to be more efficient and produced narrower confidence intervals as compared to standard MI methods. However, they focused on the setting where few variables have missing values. One disadvantage of algorithms based on random forests is that they are computationally expensive to implement in high dimensions and do not account for the uncertainty of estimating parameters in the imputation models (Rubin (1987)).

Loh et al. (2016) implement CART and forests to overcome incomplete data problems when the auxiliary variables are numerous. The study shows that the CART and forests methods are more reliable than likelihood methods for MI but CART can be biased toward selecting variables that allow more splits (Loh and Shih (1997)), (Kim and Loh (2001)). The study by Burgette and Reiter (2010) suggests that inferences based on the CART imputation engine can be more reliable than default applications of MICE based on main-effects generalized linear models. However, despite of various merits, CART methods and other fully conditional specifications are subject to odd behaviours, e.g. CART can be biased toward selecting variables that allow more splits in high dimensions (Raghunathan et al. (2001)), (van Buuren and Oudshoorn (1999)). Categorical predictors with many levels can be a major hurdle for CART algorithms. For example, over two billion potential partitions are formed for a categorical predictor with 32 levels, which makes CART algorithms computationally inefficient for standard computers.

The joint modelling (JM) specification is an alternative to the FCS approach. JM involves specifying a multivariate distribution for the data and draws imputations from their conditional distributions by Markov Chain Monte Carlo (MCMC) methods. Joint distributions of the variables with missing values are also specified by parametric, non-parametric and semi parametric models. A non-parametric Bayesian joint modelling approach for MI for multivariate categorical data presented by Si and Reiter (2013) uses the Dirichlet process mixtures of multinomial distributions (DPMPM) (Dunson and Xing (2009)). This approach automatically models complex dependencies whereas other MI methods (log linear model or conditional logistic regressions) can fail to detect complex structures in high-dimensional categorical variables. Akande et al. (2017) compared the performance of various MI methods for categorical data. According to their study, the Bayesian mixture model approach dominates the approach based on chained equations (which uses generalized linear models) and is as reliable as imputations based on CART in MICE. They also found that the Bayesian joint modelling approach is substantially faster than the FCS methods for MI. However, in the presence of a large number of categorical and continuous variables, the sequential behaviour of CART can form suboptimal and unstable trees (Hastie et al. (2001)), (Marshall and Kitsantas (2012)), (Strobl et al. (2009)). Moreover, to implement a fully Bayesian, joint modelling approach as suggested by Akande et al. (2017), one has to either discard all continuous variables or to categorize them. Murray and Reiter (2016) extended the Bayesian, joint modelling approach for multivariate continuous and categorical variables. However, this approach involves knowledge of complicated models to create the dependence structure between the continuous and the categorical variables. Schafer (1997) uses a JM approach called general location models for a mixture of continuous and categorical variables. Despite of being superior to FCS and CART in many ways, He (2010) suggests that the JM approaches can lack the flexibility needed to represent complex data structures arising in many studies (van Buuren (2007)).

Various recursive partitioning (RP) techniques (Iacus and Porro (2007, 2008)), (Nonyane and Foulkes (2007)), (Burgette and Reiter (2010)), (Stekhoven and Bühlmann (2012)), (Doove et al. (2014)) were proposed to overcome the problem of ignoring interactions in chained equations but most of the proposed methods combine recursive partitioning with single imputation instead of multiple imputation.

An approach called multilevel singular value decomposition (SVD) is used by Husson et al. (2018) for mixed data. SVD uses the between and within groups variability to impute values. One major drawback of SVD is that it cannot be implemented with MI. Geneviève et al. (2018) addressed main effects and interaction challenges in mixed and incomplete data frames.

MI by multiple correspondence analysis (MIMCA) (Audigier et al. (2017b)) utilizes the dimensionality reduction property of multiple correspondence analysis to impute categorical data with a high number of categories. Estimates obtained by MIMCA are as reliable as methods using MI with log linear models or conditional logistic regressions. MIMCA is less time-consuming on data sets with high dimensions than the other multiple imputation methods. However, MIMCA is limited to only categorical variables. Imputation methods that treat the categorical data as continuous, for example, as multivariate normal, can work well for some problems but are known to fail in others, even in low dimensions (Ake (2005)), (Allison (2000)), (Bernaards et al. (2007)), (Finch (2010)), (Graham and Schafer (1999)), (Horton et al. (2003)), (Yucel et al. (2011)).

An iterative singular value decomposition (SVD) algorithm for MI can be a good choice for quantitative (Hastie et al. (2015)), qualitative (Audigier et al. (2017a)) and mixed data (Audigier et al. (2016)) because of better performance than their counterparts. However, these methods cannot be suitable for the complex data we address in this paper.

Recently, hybrid techniques for imputations have gained a lot of attention (Ankaiah et al. (2011)), (Tang et al. (2015)), (Liyong et al. (2016)), (Shukur and Lee (2015)). For example, Ankaiah and Ravi (2011) propose a hybrid two stage imputation method involving the K-means algorithm and a multi-layer perceptron (MLP) in stage 1 and stage 2, respectively. Also, Nishanth et al. (2012) proposed a hybrid clustering and model based method, where they combine the K-means with an artificial neural network (ANN). Nishanth and Ravi (2013) propose an online data imputation framework incorporating data mining techniques. Considering the local similarity of data, Li et al. (2013) borrowed the idea from clustering and applied it to the problem of missing data imputation. Azim et al. (2014) present a hybrid model that uses a multi-layer perceptron and a fuzzy c-means clustering working in sequence for data imputation. Liang et al. (2015) also proposed a novel missing value imputation method using the stacked auto-encoder and incremental clustering (SAIC). However, obtaining good clustering results may become challenging due to the expansion of the data volume with existing clustering algorithms. Multiple Imputation using grey theory and entropy based on clustering (MIGEC) is another hybrid missing data method proposed by Ting et al. (2014). The MIGEC method divides the complete data into clusters and selects the nearest cluster based on grey theory for each incomplete instance and imputes values using a weighted average based on the information entropy.

Various other MI approaches are suggested in nested imputation (Rubin (2003)), where a set of a variable is imputed based on the former set. Two-stage multiple imputation by Harel (2007), Harel and Schafer (2003), Reiter and Drechsler (2007), Reiter and Raghunathan (2007) are examples for nested imputations. These methods explicitly manage two multiple imputation procedures in a dependent structure (Rubin (2003)). Weirich et al. (2014) extended nested imputation methods in both continuous and categorical background variables for a large-scale assessment. However, we think these procedures are computationally more extensive, implemented in limited ways and require further research. Zhao and Long (2016) did some recent work for imputation methods in the presence of high-dimensional data. However, they focused on the setting where only one variable has missing values. Most recently, Nikfalazar et al. (2019) proposed a new hybrid imputation method that deals with the missing data issue in the Mobility in Cities Database (MCD). Their hybrid method combines features of decision trees and fuzzy clustering into an iterative algorithm for missing data imputation.

When dealing with large scale complex data with missing values in high-dimensional situations, we desire a hybrid multiple imputation approach (HMI) that (i) avoids odd behaviours of FCS techniques in high dimensions (ii) avoids difficulties of creating complicated models for the dependence between the continuous and the categorical variables as in JM approaches (iii) avoids the problem of a specification of clusters (iv) offers efficient computation. HMI is a flexible and practical technique, which combines properties of existing approaches to handle missing values in large scale complex surveys. We propose a HMI technique which applies FCS MI techniques to impute continuous variables based on information obtained by categorical variables that are already imputed by a JM MI approach.

4. Materials and methods

Before introducing the proposed hybrid architecture, a brief description of FCS and JM MI methods, Rubin's inference and various estimates used for comparing the performance of the imputation algorithms is given below.

4.1. Fully Conditional Specification (FCS): Chained Equations

The FCS method specifies an imputation model for each variable with missing values conditional on the other variables in the data set. Missing values are sequentially imputed in each iteration. Imputation starts from the first variable with missing values.

In the first step, initial values for missing values in all variables are specified, i.e. Y_1^0 , ... ,Y_1^0.

In the second step, at iteration t: for j = 1 to p, most recently imputed values, i.e. X, Y_1^t, ... ,Y_{j-1}^t,Y_{j+1}^{t-1} , ... ,Y_p^{t-1} of all other variables, X, Y_2^{t-1}, ... , Y_p^{t-1} for j=1 and Y_1^{t-1}, ... ,Y_p^{t-1} use a univariate method to impute all missing values in the jth variable Y_j^t. Here, X denotes a set of variables that have no missing values. Repeat the second step until the maximum number of iterations is reached. The above steps (including the first one) are repeated M times to get M imputations. The starting values for each chain are generated with a different seed for random numbers to generate different initial values.

4.2. Fully Bayesian joint modelling (JM) using Dirichlet process infinite mixtures of products of multinomials (DPMPM)

The fully Bayesian, joint modelling (JM) approach known as "Dirichlet process mixtures of products of multinomial distributions model" (DPMPM) (Dunson and Xing, (2009)) is described as:

1. Assume that each individual *i* belongs to exactly one of *K* < ∞ latent classes.
2. For *i = 1,…, n*, let $g_i \epsilon \{1, \dots, k\}$ indicate the class of individual *i*, and let π_k =*Pr* $(g_i = k)$. Assume further that $\pi = \{\pi_1, \dots, \pi_k\}$ is the same for all individuals.
3. Within any class, we suppose that each of the *j* variables independently follows a class-specific multinomial distribution, i.e. for any value

$$y_j \ \epsilon \ \{1, \dots, d_j\}, \text{let } \phi^{(j)}_{k_c j} = Pr(y_{ij} = y_j \,|g_i = k).$$

Note that d_j denotes the number of categories of the *j*-th variable.

Mathematically, the finite mixture model can be expressed as follows:

$$y_{ij}|\mathrm{g}_i, \phi \overset{ind}{\sim} \text{Multinomial } (\phi^{(j)}_{g_i 1}, \dots, \phi^{(j)}_{g_i d_j}) \text{ for all } i \text{ and } j \qquad (4.1)$$

$$\mathrm{g}_i|\, \pi \sim \text{Multinomial } (\pi_1, \dots, \pi_K) \text{ for all } i \qquad (4.2)$$

For prior distributions on ϕ and π , we have

$$\pi_k = V_k \left(\prod_{l<k} 1 - V_g \right) \text{ For } k=1,\dots,K$$

$$V_k \overset{iid}{\sim} Beta\ (1,\ \alpha)$$

$$\alpha \sim Gamma\ (a_\alpha,\ b_\alpha\ )$$

$$\phi_{kj} \sim Dirichlet \quad (\ a_{j1}\ , \dots , a_{jd_j})$$

We set a_{j1}=...= a_{jd_j} = 1 for all *j*, and (a_α = 0.25; b_α = 0.25). In order to get complete data sets, first the latent class indicator for each individual is drawn from the full conditional and then each missing y_{ij} is drawn from the class specific, independent multinomial distributions.

4.3. Rubin's inference:

For *m = 1,…,M,* let $q^{(m)}$ and $u^{(m)}$ be respectively the point estimates of *Q* (e.g. the estimated regression coefficient in an analysis model) and the variance estimates of $q^{(m)}$ of the interesting analysis model, e.g. a parametric regression model. Valid inferences for a scalar *Q* are obtained by combining the $q^{(m)}$ and $u^{(m)}$, using Rubin's (1987) rules as follows:

$$\overline{q}_M = \sum_{m=1}^{M} \frac{q^{(m)}}{M}, \tag{4.3}$$

$$b_M = \sum_{m=1}^{M} \frac{(q^{(m)} - \overline{q}_M)^2}{M-1}, \tag{4.4}$$

$$\overline{u}_M = \sum_{m=1}^{M} \frac{u^{(m)}}{M}\ , \tag{4.5}$$

$\overline{q}_M$ can be used to estimate *Q* and the variance of $\overline{q}_M$ can be estimated by

$$T_M = \left(1 + \frac{1}{M}\right) b_M + \overline{u}_M, \tag{4.6}$$

with degrees of freedom $v_M = (M-1)(1 + \frac{\overline{u}_M}{(\left(1+\frac{1}{M}\right)b_M)^2})$. (4.7)

5. Proposed hybrid architecture

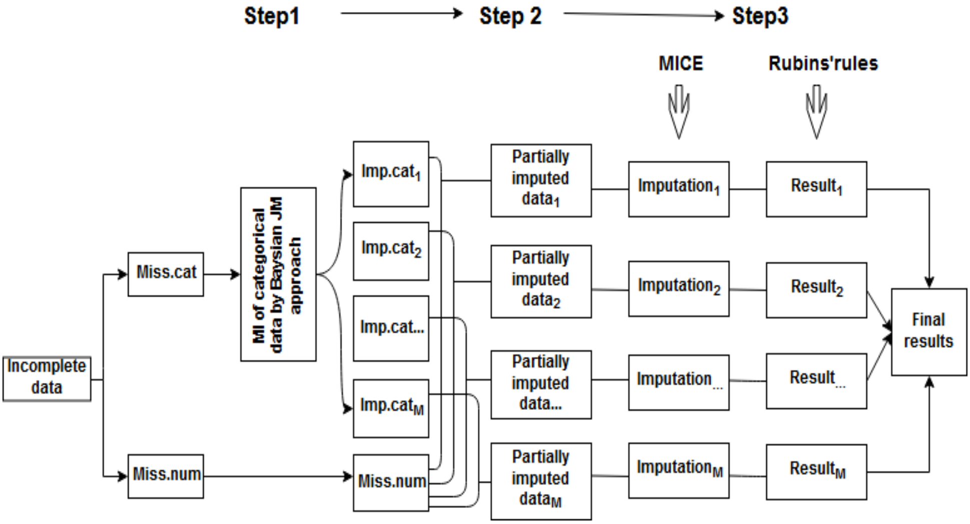


Figure 1. Schematic diagram illustrating the proposed hybrid architecture

A schematic diagram illustrating the proposed hybrid architecture is provided in Figure 1. The proposed missing data imputation approach is a 3-stage approach. **Step 1**: We begin by partitioning incomplete data into two different groups, i.e. categorical data → Miss.cat and incomplete continuous data → Miss.num, where Miss.cat and Miss.num may contain missing values. After partitioning, multiple complete versions → Imp.cat are created for Miss.cat by applying a fully Bayesian joint modelling approach to MI. In this step, Miss.num still contains missing values. **Step 2**: All variables in the data set Miss.num are added to each of the Imp.cat data sets, resulting in *M* partially imputed datasets where values in the continuous variables may be missing and values in the categorical variables have already been imputed in step 1. **Step 3:** Incomplete continuous variables in the *M* partially imputed datasets are imputed using MICE such that the draws from the posterior predictive distribution of the unobserved continuous data depend on the given categorical variables, which have been already imputed by the fully Bayesian joint modelling MI.

To implement the HMI approach, we combine a JM approach “DPMPM” with the FCS approach MICE. We select “DPMPM” due to its computational efficiency, its ability to automatically model complex dependencies and its successful implementation for the case of high-dimensional categorical variables in various fields, i.e. econometrics (Chib and Hamilton (2002), Hirano (2002)), social science (Kyung, Gill, and Casella (2010)), and finance (Rodrı´guez and Dunson (2011)). MICE is selected due to its open source character and popularity. R (R Core Team (2018)) software, version 3.0.1 is used to perform all calculations. The two R packages “mice” (van Buuren and Groothuis-Oudshoorn, (2011)), version 2.17

and “NPBayesImputeCat” (Quanli et al. (2018)), version 0.1 are used to implement the HMI approach. The “default” function of “mice” uses predictive mean matching (PMM) for continuous variables, logistic regression for factor variables with two levels and multinomial logit model for more than two categories. We also use the package 'mitools' (Thomas (2019)) to combine the results from MI. Default versions of chained equations using “mice” fail to impute missing values in the child data. The neural net function, called by “mice” for categorical variables with more than two categories, stops the default version because of exceeded "maximum allowable number of weights". The function “nnet” is used to prevent running code that will take a very long time to complete when there are factor variables with many levels. This gives an indication that complex dependence structures in the data make it complicated to identify them by the default application of MICE. Therefore, we did not implement the default version and compare two HMI approaches, i.e. “H.CART” and “H.DEF” with the MICE based method “Mice$_{\text{CART}}$” (classification and regression trees (CART)). “H.CART” and “H.DEF” combine a fully Bayesian joint modelling approach with the MICE algorithms “CART” and “Default”, respectively. To implement the hybrid approach, we examine a small prior specification for a_α and b_α (i.e. a_α= 0.25, b_α = 0.25) with a moderate number of mixture components (i.e. *k*=80).

6. Simulation studies

To investigate the performance of the HMI method via simulation, we generate a large number *(X=39)* of mixed type variables. First, we generate 31 binary *(X_b)* variables. A multivariate normal (MVN) distribution is used to generate correlated random covariates C_i comprising 1000 observations. The marginal distributions are: C_i ~ *N (0, 0.5),* where *i={1,…,31}.*The correlation structure is given as:

$$R = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}.$$

Where ρ = *0.5*. Random covariates (C_i) are transformed into binary values (X_b) using the following threshold:

$$X_{b_i} = \begin{cases} 0 & if \quad C_i \le 0\,, \\ 1 & if \quad C_i > 0, \end{cases}$$

where *i*={*1,…,31*}.

In order to generate two multilevel categorical covariates, i.e. (X_{m_1} and X_{m_2}), we first generate two random covariates from normal distributions (ND) given as: C_{32} ~ $N\ (\mu_1; \sqrt{2})$, C_{33} ~ $N\ (\mu_2; \sqrt{2})$, where μ_1 and μ_2 are described as:

$$\mu_1 = 0.1 + 0.1 \sum_{i=1}^{31} X_{b_i} + 0.1 X_{b_2} X_{b_3} + 0.1 X_{b_5} X_{b_8} + 0.1 X_{b_2} X_{b_{29}}.. \tag{6.1}$$

$$\mu_2 = 0.1 + 0.1 \sum_{i=1}^{31} X_{b_i} + 0.1 C_{32} + 0.1 X_{b_2} X_{b_3} + 0.1 X_{b_5} X_{b_8} + 1.1 X_{b_2} X_{b_{29}}. \tag{6.2}$$

Further, all observations in $C_{31}$and C_{32} are randomly split into various homogeneous groups and two multilevel categorical variables X_{m_1} and X_{m_2} are formed with four and six categories respectively.

To encode complex dependence relationships with higher order interactions, we generate another binary covariate $X_{b_{32}}$ from Bernoulli distributions with probabilities governed by the logistic regression with

$$logit\ Pr\ (X_{b_{32}}) = 0.001 - 0.01X_{b_1} - 0.09X_{b_2} - 0.09X_{b_3} - 0.09X_{b_4} + 0.05X_{b_5} + 0.08X_{b_6} - 0.02\,X_{b_7} + 0.08\,X_{b_8} + 0.01X_{b_9} + 0.01\,X_{b_{10}} - 0.02\,X_{b_{11}} + 0.01X_{b_{i12}} - X_{b_{13}} + 0.02X_{b_{14}} - 0.01X_{b_{15}} + 0.02\,X_{b_{16}} - 0.03X_{b_{17}} - 0.02X_{b_{18}} - 0.07X_{b_{19}} + 0.08X_{b_{20}} + 0.08X_{b_{21}} + 0.01X_{b_{22}} + 0.09X_{b_{23}} + 0.09X_{b_{24}} + 0.05X_{b_{25}} + 0.08X_{b_{26}} - 0.02X_{b_{27}} + 0.08X_{b_{28}} + 0.08X_{b_{29}} - 0.01X_{b_{30}} + 0.09\,X_{b_{31}} + 0.02\,C_{32} + 0.02C_{33} + 0.02\,X_{b_{12}}X_{b_{29}} - 0.02X_{b_{15}}X_{b_{18}}X_{b_{29}}\,.$$

(6.3

We then generate two continuous covariates, i.e. X_{n_1} and X_{n_2} from normal distributions (ND) as follows:

$$X_{n_1} \sim N\left(\mu_3; \sqrt{0.5}\right).$$

Where, $\mu_3 = -2 - 1.5X_{b_1} + 2.15X_{b_2} + 2.25\,X_{b_3} - 3.6\,X_{b_4} - 1.88X_{b_5} + 1.11\,X_{b_6} + 2X_{b_7} - 5X_{b_8} + X_{b_9} - 2X_{b_{10}} + 2X_{b_{11}} + 5X_{b_{12}} - 2X_{b_{13}} + 3X_{b_{14}} + 4X_{b_{15}} + X_{b_{16}} + X_{b_{17}} - X_{b_{18}} - X_{b_{19}} - X_{b_{20}} - X_{b_{21}} - X_{b_{22}} + 2X_{b_{23}} - X_{b_{24}} + X_{b_{25}} + X_{b_{26}} + X_{b_{27}} + X_{b_{28}} + X_{b_{29}} + X_{b_{30}} + X_{b_{31}} + 2C_{32} - C_{33} + X_{b_{32}} + 2X_{b_{11}}X_{b_{12}}X_{b_{13}} - 2\,X_{b_{15}}X_{b_{18}} + 2X_{b_{12}}\,X_{b_{29}}$.

(6.4)

And

$$X_{n_2} \sim N\left(\mu_4; \sqrt{0.5}\right). \tag{6.5}$$

Where, $\mu_4 = \mu_3 + X_{n_1}$.

(6.6)

Both continuous covariates are highly positively correlated, i.e. r = *0.9*.

We then define a covariate dependent continuous response with expectation

$$\mu_y = 1 + \sum_{i=1}^{32} X_{b_i} + \sum_{i=1}^{4} X_{n_i} + \sum_{i=2}^{4} X_{m_{1\_i}} + \sum_{i=2}^{6} X_{m_{2\_i}} + X_{b_9}X_{b_{15}} + X_{b_1}X_{b_{17}} + X_{b_{14}}X_{b_{20}} + \epsilon.$$

(6.7)

Additionally, a random component $\epsilon \sim N\,(0;\ 0.5)$ is added. The regression coefficients for categorical variables with multiple levels are expressed as dummy variables, e.g. $\sum_{i=2}^{4} X_{m_{1\_i}}$ and $\sum_{i=2}^{6} X_{m_{2\_i}}$ in the predictor (all coefficients are 1.0).

Equations 6.1–6.7 include higher-order interactions to represent complex dependence structures. Imputation approaches based on log-linear models or chained equations may fail to capture these structures. There is no particular importance of the specific values of the coefficients. Nonzero coefficients are specified for higher order interactions for generating complex dependencies. The analysis model of interest is the linear model. Observations in all covariates can be missing (at random) with probabilities based on a logistic probability distribution model. Probabilities for missing for a random covariate *X* are given as:

$$\pi_{X_i} = \frac{e^{(-2-X_j)}}{(1+e^{(-2-X_j)})}.$$

Here, *i={1,…,39}* and *j ≠ i.* Missingness in X_i is attributed solely to other observed variable X_j. This yields 10% of the observations to be MAR. Based on recommendations in the MI literature (White et al. (2011)), (van Buuren (2012)), we decided to include all of the variables from the generated data in the imputation model to ensure that the imputation model preserves the relationships between the variables of interest (Schafer (1997)), (Moons et al. (2006)). Based on *Z =1000* simulation runs, the parameters of interest are estimated using the aforementioned Rubin's method. According to Rubin (1987), the number of suitable imputations for useful statistical inferences can be determined by a fraction of missing data. A surprisingly high relative efficiency can be obtained with no more than five imputations. Fichman and Cummings (2003) suggest, that *M=10* imputations are more than suitable in almost any realistic application. Therefore, ten imputed datasets are generated for each of the proposed and the MICE MI methods. Two hundred iterations (for each imputation step) are run to insure convergence and to obtain results of the simulations in a reasonable time. To compare the performance of the imputation algorithms, two error-based measurements were chosen to evaluate the quality of MI: Root mean square error (RMSE) and empirical standard errors (ESE) (Akande et al. (2017)), (Armina et al. (2017)). Smaller values for RMSEs and ESEs indicate better performance (Oba et al. (2003)). RMSE and ESE are calculated using the following formulas:

$$\text{Root mean square error } (\mathrm{RMSE}_{\bar{q}_m}) = \sqrt{\frac{\sum_{z=1}^{Z}\left(\bar{q}_M^z - \beta\right)^2}{Z}}, \tag{6.8}$$

$$\text{Empirical standard errors } (\mathrm{ESE}_{\bar{q}_m}) = \sqrt{\frac{\sum_{z=1}^{Z}\left(\bar{q}_M^z - \bar{q}\right)^2}{Z}}, \tag{6.9}$$

where $\bar{q}_M^z$ denotes the estimated parameter pooled over *M* imputed data sets in simulation run number *z* and β denotes the original parameter. The arithmetic mean of $\bar{q}_M^z$ and ($\sqrt{T_M}$) across all *z = {1,…,Z}* simulations are denoted as $\bar{q}$ and $\overline{\sqrt{T}}$. The amount of bias can be calculated by a simple difference, i.e.

$$Bias = RMSE - ESE \tag{6.10}$$

The coverage rates of at least 95% are calculated as:

$$\text{Coverage rate}_{\bar{q}_m} = \frac{\sum_{z=1}^{Z} 1\,[\beta \in CI\,(\bar{q}_M^z, T_M^z)]}{Z}, \tag{6.11}$$

where 1 [$\beta \in CI\,(\bar{q}_M^z,\ T_M^z)$] is an indicator function. The indicator function is equal to one when the confidence interval based on $\bar{q}_M^z$ and T_M^z contains β and equal to zero otherwise.

Table 1 gives the performance of the MI methods. Means for CI coverage and RMSEs over all beta coefficients are presented in Table 2. Various researchers (White et al. (2011)), (van Buuren, 2012)) recommend graphical comparisons of the imputation methods. For that purpose, boxplots of standard errors ($\sqrt{T_M}$) and point estimates ($\overline{q}_M$) for the regression coefficients for the 1000 simulation runs are presented in Figures 2 and 3 respectively.

6.1. Results

Table 1. Performance of methods for MI

Estimates	Parameter	MICECART	H.DEF	H.CART
RMSEs (ESEs)	$X_{b_{23}}$	0.158(0.114)	0.148(**0.089**)	**0.122**(0.110)
	$X_{m_{1\_2}}$	0.158(0.155)	0.228(**0.122**)	0.173(0.158)
	$X_{m_{2\_3}}$	0.187(0.148)	0.167(**0.114**)	**0.164**(0.145)
	$X_{b_{32}}$	0.045(0.032)	0.071(**0**)	**0.032**(0.032)
	X_{n_2}	0.063(0.063)	0.071(**0.032**)	**0.055**(0.055)
	$X_{b_1}X_{b_{17}}$	**0.190**(0.182)	0.239(**0.130**)	0.195(0.190)
$\bar{q}(\sqrt{T})$	$X_{b_{23}}$	0.891(0.192)	1.119(**0.137**)	0.947(**0.137**)
	$X_{m_{1\_2}}$	1.038(0.266)	0.808(**0.193**)	0.928(0.272)
	$X_{m_{2\_3}}$	0.887(0.245)	1.122(**0.176**)	0.920(0.249)
	$X_{b_{32}}$	0.969(0.049)	1.065(**0.027**)	1.006(0.047)
	X_{n_2}	1.014(0.088)	0.935(**0.049**)	0.995(0.086)
	$X_{b_1}X_{b_{17}}$	0.951(0.319)	0.800(**0.255**)	0.958(**0.225**)
Bias	$X_{b_{23}}$	0.044	0.059	**0.012**
	$X_{m_{1\_2}}$	0.772	**0.615**	0.656
	$X_{m_{2\_3}}$	**0.039**	0.053	0.671
	$X_{b_{32}}$	0.013	0.071	**0**
	X_{n_2}	0.956	**0.886**	0.909
	$X_{b_1}X_{b_{17}}$	0.008	0.109	**0.005**
Coverage(%)	$X_{b_{23}}$	99	95	100
	$X_{m_{1\_2}}$	100	94	100
	$X_{m_{2\_3}}$	100	97	100
	$X_{b_{32}}$	97	29	99
	X_{n_2}	99	83	100
	$X_{b_1}X_{b_{17}}$	100	96	100

Root mean square errors and empirical standard errors (top), point estimates, standard errors and bias for different methods (middle) and estimated coverage probability (bottom) for MI methods under the Missing at Random (MAR) assumption. The middle panel lists the mean estimated standard errors and point estimates across the simulated data sets. All results are based on 10 imputations. Estimates are shown for only six regression coefficients, i.e. for variables $X_{b_{23}}$, $X_{m_{1\_2}}$, $X_{m_{2\_3}}$, $X_{b_{32}}$, X_{n_2} , $X_{b_1}X_{b_{17}}$. Bold figures indicate the smallest mean root mean square errors, mean empirical standard errors and amount of bias among the three imputation variants.

Table 2. Results over all beta coefficients

Estimates	MICE$_{\text{CART}}$	H.DEF	H.CART
CI coverage	98.66	91.91	99.89
RMSEs	0.184	0.170	**0.146**

Means for CI coverages and RMSEs are estimated over all regression coefficients for all MI methods. Bold values indicate the smallest mean for RMSEs over all regression coefficients among the three imputation variants.

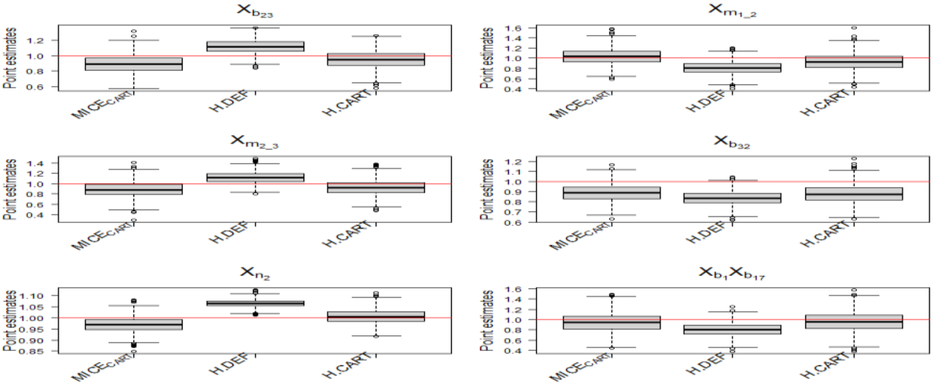


Figure 2. Simulated data: Boxplots for the point estimates ($\overline{q}_M$) across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations. Point estimates are shown for only six regression coefficients, i.e. for variables $X_{b_{23}}$, $X_{m_{1\_2}}$, $X_{m_{2\_3}}$, $X_{b_{32}}$, X_{n_2} , $X_{b_1}X_{b_{17}}$ .The horizontal red lines indicate the respective “true” values

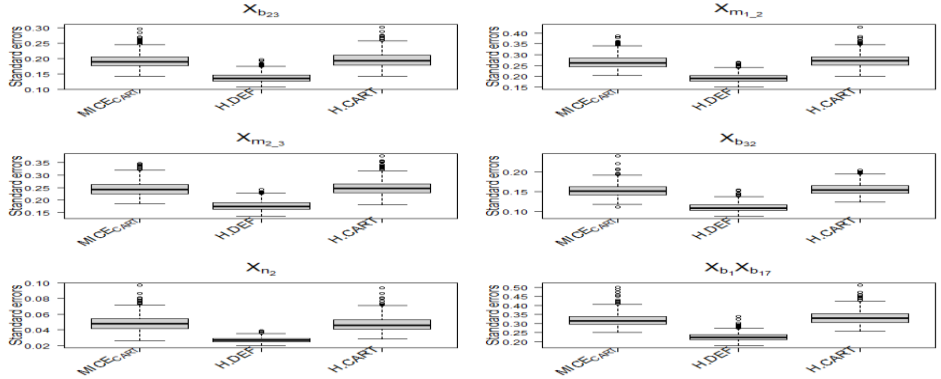


Figure 3. Simulated data: Boxplots for standard errors ($\sqrt{T_M}$) across 1000 simulations by imputation methods under Missing at Random (MAR) and ten imputations. Standard errors are shown for only six regression, i.e. $X_{b_{23}}$, $X_{m_{1\_2}}$, $X_{m_{2\_3}}$, $X_{b_{32}}$, X_{n_2} , $X_{b_1} X_{b_{17}}$ coefficients

The average point estimates based on H.CART are closer to the corresponding true values than those based on CART. H.CART tends to be less biased as compared to the CART method for all types of covariates and interaction terms, whereas H.DEF tends to be upward biased for binary and the multilevel covariate with four levels and slightly downward biased for the multilevel covariate with six levels, for the continuous covariates and the interaction terms as compared to the CART method (Figure 2). There seem to be similarities in the structure among all MI methods (i.e. all methods are downward biased) for binary covariate $X_{b_{32}}$, which was generated with higher order interactions. The H.DEF method tends to have smaller standard errors as compared to two relevant methods for all covariates, whereas the H.CART method tends to have similar standard errors as compared to CART for most of the cases (Figure 3). The estimated RMSEs, ESEs and averages of standard errors for the H.CART method are smaller for all types of covariates except the multilevel covariate with many categories. H.CART shows similar ESEs and averages of standard errors and slightly higher RMSEs for the multilevel covariate with more categories as compared to CART. The H.DEF method shows smaller ESEs and averages of standard errors for all types of covariates and slightly higher RMSEs for most of the covariates as compared to the other methods (Table 1). The H.DEF method led to more overall accuracy with smaller means for RMSEs over all beta coefficients as compared to CART (Table 2). A possible explanation for the efficiency gain with H.DEF is that it was able to make better use of the available information by accommodating nonlinearities among the predictors. For the most part, coverage rates for H.CART are in line with those from CART and produce almost identical results. In most cases, coverage probabilities for H.CART were 100%, which suggests that these confidence intervals may be too conservative. The simulated coverage rates of the 95% confidence intervals based on H.DEF are near to nominal 95% for most cases. Few of the incidences in H.DEF led to under-coverage. All but one of the incidences, i.e. $X_{b_{32}}$in which coverages dip below 30% occur. This severe under-coverage suggests that H.DEF (which uses the Bayesian approach for categorical and PMM as default for continuous covariates) might performing not well for continuous covariates but works well for categorical covariates. This might be one of the reasons that H.DEF gets biased results. Increasing M can lead to obtain coverage rates that are close to nominal in the case of under-coverages. Nevertheless, the H.DEF method led to coverage rates that are close to nominal over all beta coefficients as compared to CART (Table 2).

7. Imputation of MICS child data

The data for MICS is collected at both family and person level and it allows the study of relationships between health indicators and other characteristics. In this study, we use the child data set from the MICS Punjab 2014 household survey. The MICS Punjab data for children contains more than two hundred indicators on a variety of a child’s conditions. For example, indicators on a child’s mental development (e.g. a child is able to pick up small object with 2 fingers, etc.), a child’s nutrition intake in diet (e.g. a child drank or ate vitamin or mineral supplements, etc.) and vaccinations (e.g. ever had vaccination card, etc.). The MICS data for children contains a complex data structure for categorical variables with multiple levels and large amounts of missingness, which can be problematic for MICE. It can be tedious for MICE to specify imputation models and interaction terms in the presence of large databases with hundreds of variables and multicollinearity (Van Buuren and Oudshoorn 1999). It was not possible to have a proper comparison of the proposed and existing MI approaches in such case. Therefore, multiple categories for categorical variables were reduced by merging them, and a sub-sample of 52 variables, which contains information on child health, nutrition and development, is selected from MICS Punjab 2014 children data. Among these variables, 43 background variables are categorical with multiple categories and the remaining are continuous. Demographical variables like “district” and “area” are also included in the sub-sample. In this sub-sample, 5 variables have between 6 and 21% of missing values, 17 variables have 48% of missing values, 27 variables have between 50% and 86% of missing values, and 1 variable has more than 90% of missing values. Of all variables, only 3, i.e. “sex”, “wealth” and “area”, have complete records (see additional file). The variable “district” has 36 levels, hence keeping the analysis comparable and challenging at the same time. There are various reasons listed for item non response in the methodology of MICS i.e. nonresponse, don’t know and not reached, etc. Without distinguishing reasons for item non response, we assume that the items are MAR in the data under consideration. Similar to the simulation study, all of the variables from the sub-sample are included in the imputation model.

After imputations, parameters of interest for the child health are estimated using linear models for continuous response (height for age percentiles NCHS). The response variable, “height for age percentiles NCHS”, is obtained by using a table of Z-scores (percentile = the area from infinity to Z). Based on the evidence from demographical and behavioural risk factors associated to height, two continuous covariates i.e. “age”, “polio_vacc.” and two categorical variables, i.e.

“grains_in_diet” (Yes/ No) and “eggs_in_diet” (Yes/ No) are selected as potential determinants in the analysis model. Since there are no true values to compare for in the real data example, we calculated complete case (CC) estimates for comparison purposes (Table 5). The R package “VIM” (Templ et al. (2012)) is utilized for exploring data and the pattern of missing values. Figure 4 shows graphics of the incomplete predictors. Graphics for the remaining variables in the sub-sample are provided in an additional file. Similar to the simulation study ESEs, average point estimates and average standard across the 200 simulations are calculated for real data. Computational time and ESEs for MI methods are shown in Tables 3 and 4 respectively. Figures 5 and 6 display the average point estimates and average standard errors for the MI methods across the 200 simulations.

7.1. Results

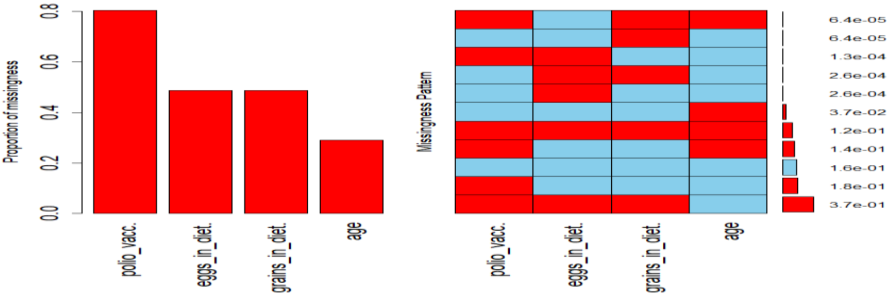


Figure 4. Real data: Aggregateplot in R, graphics of incomplete predictors. For purposes of displaying the graphical depiction, only four variables with proportions of missing values ranges between 18-28 were selected

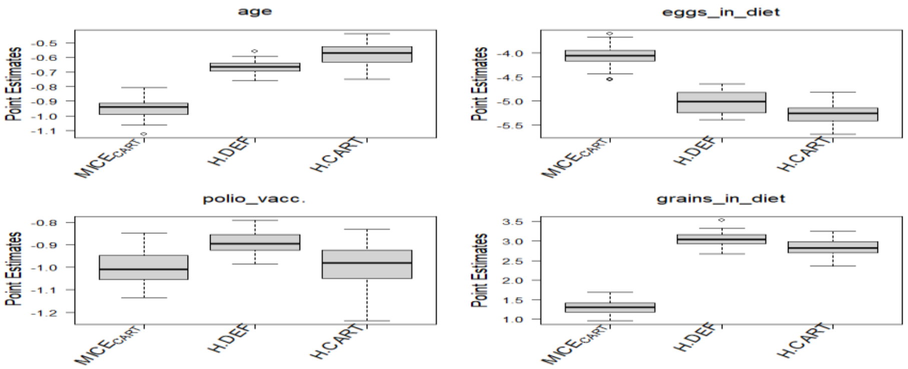


Figure 5. Real data: Boxplots for point estimates ($\overline{q}_M$) across 200 simulations by imputation methods under Missing at Random (MAR) and ten imputations

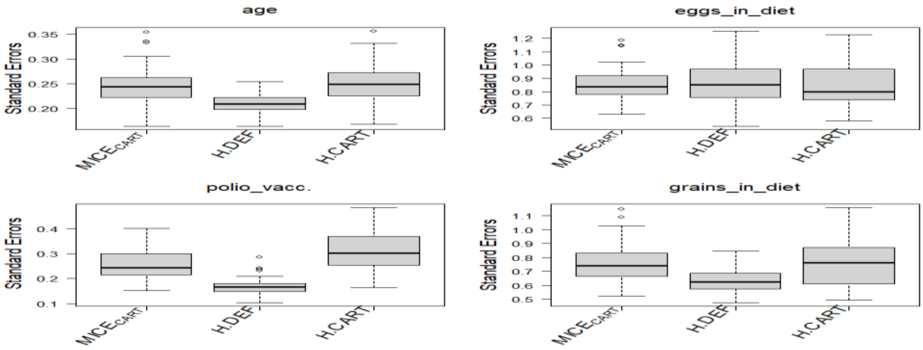


Figure 6. Real data: Boxplots for standard errors ($\sqrt{T_M}$) across 200 simulations by imputation methods under Missing at Random (MAR) and ten imputations.

Table 3. Real data: Time taken for various MI methods

Method	Default	CART	H..DEF	H.CART
Time	No run	3.25d	22.78h.	21.21h

Note: time = the time to complete 10 multiple imputation by variants of MI across 1000 simulations, h = hours, d = days, and Not Run = the program not able to complete multiple imputation on this subset. The maximum number of iterations is set to 200.

Table 4. Real data: ESEs for various MI methods

Variables	CART	H.DEF	H.CART
age	0.06	**0.04**	**0.06**
eggs_in_diet	0.21	0.22	**0.20**
polio_vacc.	0.07	**0.04**	0.09
grains_in_diet	0.17	**0.16**	0.21

Empirical standard errors by imputation methods under Missing at Random (MAR) and ten imputations. Cases where both HMI methods result in minimum between imputation variances (ESEs) as compared to CART are highlighted in bold.

Table 5. Real data: complete case (CC) estimates

Variables	est	se
age	3.542	0.899
eggs_in_diet	-9.866	1.305
polio_vacc.	-0.808	0.242
grains_in_diet	0211	1,342

The CC analysis uses only the complete cases (n = 4264), "est" and "se" denote the point estimates and standard errors of the coefficients of the linear model, respectively.

Figure 4 displays graphics of incomplete predictors. The bar plot on the left side shows the proportions of missing values in the predictors. The continuous predictor "polio_vacc." has the highest amount of missing values (i.e. about 80%) while the amount is rather small in the other three variables (i.e. less than 60% for two binary predictors and less than 40% for predictor "age"). An aggregation plot on the right side shows all existing combinations of missing (red) and imputed observed (blue) values. Additionally, the frequencies of different combinations are visualized by a small bar plot and by the number of their occurrences on the right side (Templ et al. (2012)). The aggregation plot reveals that missing values in the variable "polio_vacc." are also missing in the two binary variables. We note that the standard errors for all of the coefficients are smaller compared to the (absolute) point estimates under all MI methods (see Figures 5-6). This happens most likely due to sampling variability in the multiple imputation inferences. The empirical example with real data indicated that the CART and HMI variants yielded differing point estimates. We noticed that point estimates in CART are nearer to the estimates in complete case analysis for most of the cases with larger standard errors as compared to hybrid methods (see Table 5, Figures 5-6). Figure 6 displays smaller standard errors for H.DEF as compared to CART. ESEs for HMI variants are also smaller as compared to CART for most of the cases (see Table 5), suggesting better performance over CART. Given the results produced by the MI methods, a look at the computation times in Table 3 may be useful for a further comparison. Almost 4 days were taken by CART to run on standard computers, whereas, surprisingly, this time was reduced to almost one day when HMI methods were applied. We also applied the proposed methods to the full MICS data set with hundreds of variables and categories with multiple levels. We found that the proposed methods have a good capacity to perform for the MICS data where the MICE methods simply fail.

8. Conclusion and remarks

We acknowledge that results of MI can be biased even when complex multivariate data is MAR (White and Carlin, 2010). However, in this paper, we assumed that the missing data mechanism is MAR. We applied our hybrid strategy to handle missing data in large scale survey data with complex dependence structures among categorical variables and a high percentage of missing information. Identification of complex dependence structures among mixed type covariates will be difficult for JM and FCS MI methods in high dimensions. We obtain promising results by performing an illustrative analysis. The results obtained from the simulation studies and a real data example confirm the potential of our proposed approach to handle missing data under MAR. Superiority of H.DEF was its efficiency relative to the other imputation inference methods. The H.DEF method outperformed the other methods with respect to RMSEs, ESEs and standard errors but its point estimates were downwardly biased for a few regression coefficients, which led to under-coverage of the confidence intervals. H.CART gives estimates with less bias but over-coverage of confidence intervals. There was no noticeable difference in coverage and standard errors between H.CAT and CART. H.CART produces smaller RMSEs and ESEs for most parts and 3 times less computational cost as compared to MICE. A problem of the HMI approach is that it does not use the information available on the continuous variables for imputing the categorical variables. Further work is needed to use iterative procedures to develop strong relationships between the categorical and continuous variables. Currently, we are implementing solutions for this problem and we use the concept of categorizing continuous variables. We are working on the development of a new R package that will implement the proposed HMI approach with the hope that it will contribute in MI of large scale survey data.

Acknowledgments

The authors thank the editor and two referees for their suggestions, which greatly helped improve the article.

REFERENCES

ANDERSON, A. B., BASILEVSKY, A., HUM, D. P., (1983). Missing data: A review of the literature. In J. D. W. P. H. Rossi and A. B. Anderson (Eds.), Handbook of survey research, New York: Academic Press.

ARNOLD, B. C., PRESS, S. J., (1989). Compatible Conditional Distributions. Journal of the American Statistical Association, 84, pp. 152–156.

ALLISON, P. D., (2000). Multiple imputation for missing data: A cautionary tale. Sociological Methods and Research, 28, pp. 301–309.

AKE, C. F., (2005). Rounding after multiple imputation with non-binary categorical covariates (paper 112-30). In Proceedings of the Thirteenth Annual SAS Users Group International Conference, SAS Institute Inc., Cary, NC, pp. 1–11.

ANDRIDGE, R. R. (2009). Statistical methods for missing data in complex sample surveys. PhD thesis, The University of Michigan.

AKMATOV, M. K., (2011). Child abuse in 28 developing and transitional countries--results from the Multiple Indicator Cluster Surveys, Int J Epidemiol, 40(1), pp. 219–27.

ANKAIAH, N., RAVI, V., (2011). A novel soft computing hybrid for data imputation, Proceedings of the 7th international conference on data mining (DMIN), Las Vegas, USA.

AZIM, S., AGGARWAL, S. (2014). Hybrid model for data imputation: using fuzzy c means and multi layer perceptron. Advance Computing Conference (IACC), 2014 IEEE International. IEEE, pp. 1281–1285.

AUDIGIER, V., HUSSON, F., JOSSE, J., (2016). A principal component method to impute missing values for mixed data, Advances in Data Analysis and Classification, 10(1), pp. 5–26.

AKANDE, O., LI, F., REITER, J., (2017). An empirical comparison of multiple imputation methods for categorical data, Amer. Statist, 71, pp. 162–170.

ARMINA, R., ZAIN, A.M., ALI, N.A., SALLEHUDDIN, R., (2017). A review on missing value estimation using imputation algorithm, Journal of Physics: Conference Series, 892, pp. 012004.

AUDIGIER, V., WHITE, I. R., JOLANI, S., DEBRAY, T., QUARTAGNO, M., CARPENTER, J., ESCHE-RIGON, M., (2017a), Multiple imputation for multilevel data with continuous and binary variables, arXiv preprint, arXiv:1702.00971.

AUDIGIER, V., HUSSON, F., JOSSE, J., (2017b). MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. Statistics and Computing, 27, pp. 501–518.

BREIMAN, L., (2001). Random Forests. Machine Learning, 45(1), pp. 5–32.

BERNAARDS, C. A., BELIN, T. R., SCHAFER, J. L., (2007). Robustness of a multivariate normal approximation for imputation of binary incomplete data, Statistics in Medicine, 26, pp. 1368–1382.

BURGETTE, L. F., REITER, J. P., (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. American Journal of Epidemiology, Oxford University Press, 172(9), pp. 1070–6.

CHIB, S., HAMILTON, B. H., (2002). Semiparametric Bayes analysis of longitudinal data treatment models, Journal of Econometrics, 110, pp. 67–89.

CAPPA, C., KHAN, S.M., (2011). Understanding caregivers' attitudes towards physical punishment of children: evidence from 34 low- and middle-income countries, Child Abuse Negl, 35(12), pp. 1009–21.

DUNSON, D. B., XING, C., (2009). Nonparametric Bayes modeling of multivariate categorical data, Journal of the American Statistical Association, 104, pp. 1042-1051.

DENG, Y., CHANG, C., IDO, M.S., LONG, Q., (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. Scientific Reports, 6.

DOOVE, LISA, L., VAN BUUREN, S., ELISE, D., (2014). Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects, Computational Statistics and Data Analysis, Elsevier, 72, pp. 92–104.

EROSHEVA E. A., FIENBERG S. E., JUNKER B. W. (2002). Alternative statistical models and representations for large sparse multi-dimensional contingency tables, Annales de la Faculté des Sciences de Toulouse, 11, pp. 485–505.

FICHMAN, M., CUMMINGS, J. N., (2003). Multiple Imputation for Missing Data: Making the most of What you Know, Organizational Research Methods, 6(3), pp. 282–308.

FINCH, W. H., (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. Journal of Data Science, 8, pp. 361–378.

GELMAN, A., SPEED, T. P., (1993). Characterizing a joint probability distribution by conditionals, Journal of the Royal Statistical Society Series B: Statistical Methodology, 55, pp. 185–188.

GRAHAM, J. W., SCHAFER, J. L., (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. H. Hoyle (Ed.), Statistical strategies for small sample research, Thousand Oaks, CA: Sage, pp.1–29.

GENEVIÈVE, R., OLGA, K., JULIE, J., ÉRIC M., ROBERT, T., (2018). Main effects and interactions in mixed and incomplete data frames. arXiv preprint, arXiv:1806.09734.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J., (2001). The Elements of Statistical Learning; Data Mining, Inference, and Prediction, second ed. Springer Verlag, New York.

HIRANO, K., (2002). Semiparametric Bayesian inference in autoregressive panel data models. Econometrica, 70, pp. 781–799.

HAREL, O., SCHAFER, J. L., (2003). Multiple Imputation in two Stages. Proceedings of the Federal Committee on Statistical Methodology Research Conference, Washington D. C.

HORTON, N. J., LIPSITZ, S. P., PARZEN, M., (2003). A potential for bias when rounding in multiple imputation. The American Statistician, 57, pp. 229–232.

HAREL, O., (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. Statistical Methodology, 4, pp. 75–89.

HE, Y., (2010). Missing data analysis using multiple imputation: getting to the heart of the matter. Circ Cardiovasc Qual Outcomes, 3, pp. 98–105.

HASTIE, T., MAZUMDER, R., LEE, J. D., ZADEH,R., (2015). Matrix completion and low-rank svd via fast alternating least squares, J. Mach. Learn. Res., 16(1), pp. 3367–3402.

HOLDER, L., (2015). Multiple Imputation in Complex Survey Settings: A Comparison of Methods within the Health Behaviour in School-aged Children Study, Queen’s University

HUSSON, F., J. JOSSE, B. NARASIMHAN, G. ROBIN., (2018). Imputation of mixed data with multilevel singular value decomposition, arXiv e-prints, arXiv:1804.11087.

IACUS, S. M., PORRO, G., (2007). Missing data imputation, matching and other applications of random recursive partitioning. Comput. Statist. Data Anal, 52, pp. 773–789.

IACUS, S. M., PORRO, G., (2008). Invariant and metric free proximities for data matching: an R package. J. Stat. Softw, 25, pp. 1–22.

KIM, H., LOH, W.Y., (2001). Classification trees with unbiased multiway splits. Journal of the American Statistical Association, 96, pp. 589–604.

KYUNG, M., GILL, J., CASELLA, G., (2010). Estimation in Dirichlet random effects models. Annals of Statistics, 38, pp.979–1009.

WIRTH, K. E., TCHETGEN TCHETGEN, E. J., (2014). Accounting for selection bias in association studies with complex survey data. Epidemiology (Cambridge, Mass.), 25(3), pp. 444–453.

LOH, W., SHIH, Y., (1997). Split selection methods for classification trees. Statistica Sinica, 7, pp. 815–840.

LITTLE, R. J. A., RUBIN, D. B., (2002). Statistical analysis with missing data (2nd ed.). New York: Wiley.

LEE, K.J., GALATI, J. C., SIMPSON, J. A., CARLIN, J. B., (2012). Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study. Stat Med, 31(30), pp. 4164–74.

LI, D., GU, H., ZHANG, L.Y., (2013). A hybrid genetic algorithm-fuzzy c-means approach for incomplete data clustering based on nearest-neighbor intervals. J. Soft Computing, 17, pp. 1787–1796.

LIANG, Z., ZHIKUI, C., ZHENNAN, Y., YUEMING, HU., (2015). A Hybrid Method for Incomplete Data Imputation. 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, New York, pp. 1725–1730.

LIYONG, Z., WEI, L., XIAODONG, L., WITOLD, P., CHONGQUAN, Z., LU, W., (2016). A Global Clustering Approach Using Hybrid Optimization for Incomplete Data Based on Interval Reconstruction of Missing Value, International Journal of Intelligent Systems, 31(4), pp. 297–313.

LOH, W. Y., ELTINGE, J., CHO, M., LI, Y., (2016). Classification and Regression Tree Methods for Incomplete Data from Sample Surveys, arXiv preprint arXiv:1603.01631.

LEE, K. J., CARLIN, J. B., (2017). Multiple imputation in the presence of non-normal data. Stat Med, 36(4), pp. 606–17.

MARKER, D. A., JUDKINS, D. R., WINGLEE, M. (2002), Large-Scale Imputation for Complex Surveys. Survey Nonresponse, Wiley: New York, pp. 329–341.

MOONS, K. G. M., DONDERS, R. A. R. T., STIJNEN, T., HARRELL, F. E., (2006). Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol, 59(10), pp. 1092–101.

MORRIS, T. P., IAN, R. W., PATRICK, R., (2014). Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws. BMC Medical Research Methodology, BioMed Central, 14(1), 75.

MARSHALL, R. J., KITSANTAS, P., (2012). Stability and structure of cart and span search generated data partitions for the analysis of low birth weight. J. Data Sci, 10, pp. 61–73.

MURRAY, J. S., REITER, J. P., (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. Journal of the American Statistical Association, 111, pp. 1466–1479.

NONYANE, B. A. S., FOULKES, A. S., (2007). Multiple imputation and random forests (MIRF) for unobservable, high-dimensional data. Int J Biostat, 3, pp. 1–18.

NISHANTH, K. J., RAVI, V., ANKAIAH, N., BOSE, I., (2012). Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. Expert Sys Appl, 39(12), pp. 10583–10589.

NISHANTH, K. J., RAVI, V., (2013). A computational intelligence based online data imputation method: An application for banking. J. Inf. Process. Syst. 9, pp. 633–650.

NIKFALAZAR, S., YEH C. H., BEDINGFIELD, S., KHORSHIDI, H. A., (2019). A Hybrid Missing Data Imputation Method for Constructing City Mobility Indices. In: Islam R. et al. (eds.) Data Mining. AusDM 2018. Communications in Computer and Information Science, Vol. 996. Springer, Singapore.

OBA, S., SATO, M., TAKEMASA, I., MONDEN, M., MATSUBARA, K., ISHII, S., (2003). A Bayesian missing value estimation method for gene expression profile data. Bioinformatics, 19, pp. 2088–2096.

QUANLI, W., DANIEL, M.V., REITER, J. P., JIGCHEN, H., (2018). NPBayesImputeCat: Non-Parametric Bayesian Multiple Imputation for Categorical Data. R package version 0.1, https://CRAN.R-project.org/package=NPBayesImputeCat.

RUBIN, D. B., (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

RAGHUNATHAN, T. W., LEPKOWKSI, J. M., VAN HOEWYK, J., SOLENBEGER, P. A., (2001). Multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27, pp. 85–95.

RUBIN, D. B., (2003). Nested multiple imputation of NMES via partially incompatible MCMC. Statistica Neerlandica, 57(1), pp. 3–18.

REITER, J. P., DRECHSLER, J., (2007). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. IAB Discussion Paper, 20, pp. 1–18.

REITER, J. P., RAGHUNATHAN, T. E., (2007). The multiple adaptions of multiple imputation, Journal of the American Statistical Association, 102, pp. 1462–1471.

RODRI´GUEZ, A., DUNSON, D. B., (2011). Nonparametric Bayesian models through probit stick-breaking processes. Bayesian Analysis, 6, pp. 145–178.

R Core Team (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria,

https://www.Rproject.org/.

SCHAFER, J. L., (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

STROBL, C., MALLEY, J., ZEILEIS, A., (2009). An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol. Methods, 14, pp. 323–348.

SU, Y.S., GELMAN, A., HILL, J., YAJIMA, M., (2011). Multiplebimputation with diagnostics (mi) in R: Opening windows into the black box. Journal of Statistical Software, 45(2), pp. 1–31.

SEAMAN, S., BARTLETT, J., WHITE, I., (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Med Res Methodol, 12(1), pp. 1–13.

STEKHOVEN, D. J., BÜHLMANN, P., (2012). MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics, 28, pp.112–118.

SI, Y., REITER, J. P., (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. Journal of Educational and Behavioral Statistics, 38, pp. 499–521.

SHAH, A.D., JONATHAN, W. B., JAMES, C., OWEN, N., HARRY, H., (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using Mice: A Caliber Study. American Journal of Epidemiology, 179 (6). Oxford University Press, pp. 764–74.

SHUKUR, O. B., LEE, M. H., (2015). Imputation of missing values in daily wind speed data using hybrid AR-ANN method. Modern Applied Science.

TEMPL, M., ANDREAS, A., ALEXANDER, K., BERND, P., (2012). VIM: Visualization and Imputation of Missing Values, http://cran.r-project.org/web/packages/VIM/VIM.pdf.

TING, J., YU, B., YU, D., MA, S., (2014). Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering, Applied intelligence, 40(2), pp. 376–388.

TANG, J., ZHANG, G., WANG, Y., WANG, H., LIU, F., (2015). A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. Transportation Research Part C: Emerging Technologies, 51, pp. 29–40.

THOMAS, L., (2019). mitools: Tools for Multiple Imputation of Missing Data. R package version 2.4, https://CRAN.R-project.org/package=mitools.

VAN BUUREN, S., OUDSHOORN, C. G. M., (1999). Flexible multivariate imputation by MICE. Tech. rep., TNO Prevention and Health, Leiden.

VAN BUUREN, S., GROOTHUIS-OUDSHOON, K., (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), pp. 1–67.

VAN BUUREN, S., (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. Statistical Methods in Medical Research, Sage Publications Sage UK: London, England, 16(3), pp. 219–42.

VERMUNT, J. K., VAN GINKEL, J. R., VAN DER ARK, L. A., SIJTSMA, K., (2008). Multiple imputation of incomplete categorical data using latent class analysis. Sociological Methodology, 38, pp. 369–397.

VAN BUUREN, S., (2012). Flexible imputation of missing data. Boca Raton: CRC Press.

WHITE, I. R., ROYSTON, P., WOOD, A. M., (2011). Multiple imputation using chained equations: issues and guidance for practice. Stat Med, 30(4), pp. 377–99.

WHITE, I.R., CARLIN, J. B., (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med, 29(28), pp. 2920–31.

WEIRICH, S., HAAG, N., HECHT, M., BÖHME, K., SIEGLE, T., LÜDTKE, O., (2014). Nested multiple imputation in large-scale assessments. Large Scale Assess. Educ., 2, pp. 1–18.

XIE, X., MENG, X.-L., (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial? Statistica Sinica 27, pp. 1485–1594 (including discussion).

YUCEL, R.M., HE, Y., ZASLAVSKY, A. M., (2011). Gaussian-based routines to impute categorical variables in health surveys. Stat Med, 30(29), pp. 3447–60.

ZHU, J., M., EISELE, M., (2013). Multiple Imputation in a Complex Household Survey, The German Panel on Household Finances (PHF): Challenges and Solutions. PHF User Guide.

ZHAO, Y., LONG, Q., (2016). Multiple imputation in the presence of high-dimensional data. Statistical Methods in Medical Research, 25, pp. 2021–2035.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 59–70, DOI 10.21307/stattrans-2019-034
Submitted – 03.10.2017; Paper ready for publication – 06.02.2019

LINEAR CHOLESKY DECOMPOSITION OF COVARIANCE MATRICES IN MIXED MODELS WITH CORRELATED RANDOM EFFECTS

Anasu Rabe[1], D. K. Shangodoyin[2], K.Thaga[3]

ABSTRACT

Modelling the covariance matrix in linear mixed models provides an additional advantage in making inference about subject-specific effects, particularly in the analysis of repeated measurement data, where time-ordering of the responses induces significant correlation. Some difficulties encountered in these modelling procedures include high dimensionality and statistical interpretability of parameters, positive definiteness constraint and violation of model assumptions. One key assumption in linear mixed models is that random errors and random effects are independent, and its violation leads to biased and inefficient parameter estimates. To minimize these drawbacks, we developed a procedure that accounts for correlations induced by violation of this key assumption. In recent literature, variants of Cholesky decomposition were employed to circumvent the positive definiteness constraint, with parsimony achieved by joint modelling of mean and covariance parameters using covariates. In this article, we developed a linear Cholesky decomposition of the random effects covariance matrix, providing a framework for inference that accounts for correlations induced by covariate(s) shared by both fixed and random effects design matrices, a circumstance leading to lack of independence between random errors and random effects. The proposed decomposition is particularly useful in parameter estimation using the maximum likelihood and restricted/residual maximum likelihood procedures.

Key words: correlated random effects, covariance matrix, linear Cholesky decomposition, linear mixed models.

1. Introduction

Linear mixed models are a class of models (Laird and Ware, 1982) that provide parameter estimates (inference) for population (fixed effects) and subject-specific (random effects) characteristics via separate covariance structures.
Let $Y_i = (y_{i1}, \ldots, y_{in_i})^T$ be $n_i \times 1$ vector of responses measured on the i^{th} subject

[1] Department of Mathematics and Computer Science, Umaru Musa 'Yaradua University, PMB 2218, Dutsinma Road, Katsina, Katsina State, Nigeria. E-mail: anasu.rabe@umyu.edu.ng

[2] Department of Statistics, University of Botswana, Corner of Notwane and Mobuto Road, Pvt Bag 00706, Gaborone, Botswana. E-mail: dkshangodoyin@mopipi.ub.bw

[3] Department of Statistics, University of Botswana, Corner of Notwane and Mobuto Road, Pvt Bag 00706, Gaborone, Botswana. E-mail: kthaga@mopipi.ub.bw

$(i = 1,\dots,m)$ from a total of $n = \sum_{i=1}^{m} n_i$ measurements. A linear mixed model (Laird and Ware, 1982) for the i^{th} subject is represented by:

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i \tag{1.1}$$

where X_i is $n_i \times p$ design matrix for the $p \times 1$ vector of fixed-effects regression coefficients β, Z_i is $n_i \times q$ design matrix for the $q \times 1$ vector of random effects γ_i and ε_i is $n_i \times 1$ vector of error terms. For model (1.1), we assume that:

i. Error terms ε_i are independent within i^{th} subject and are normally distributed with zero mean and $n_i \times n_i$ covariance matrix Σ_i: $E(\varepsilon_i) = 0$ and $\varepsilon_i \sim N(0, \Sigma_i)$.

ii. The random effects γ_i are independent and normally distributed with mean zero and $q_i \times q_i$ covariance matrix Δ_i. X and Z share no covariate(s) so that γ_i and ε_i are independent: $E(\gamma_i) = 0$, $\gamma_i \sim N(0, \Delta_i)$ and $Cov(\gamma_i, \varepsilon_i) = 0$.

iii. The response variable y_i is normally distributed with mean $X_i\beta$ and covariance matrix V_i: $y_i \sim N(X_i\beta, V_i)$, where $V_i = Z_i\Delta_i Z_i^T + \Sigma_i$.

In practice, assumptions on random effects are difficult to satisfy. For example, non-normality of random effects has been proven in the literature by several authors, with Lange and Ryan (1989) providing some concrete examples. Assumption of independence between ε_i and γ_i may also not hold when research interest require that X and Z have common covariate(s). For example (Gelfand et al., 1995), in growth studies where individual profiles are centered about a population baseline curve. In such cases, individual models incorporate the baseline population covariate. Another example is the analysis of CD4 cell counts in HIV studies where all subjects are HIV-positive at baseline, but not all were diagnosed for the disease. Interest here is to develop a model that incorporates diagnosis as a baseline covariate and therefore should be incorporated in both X and Z. Also, in hierarchical mixed effects models (Pinheiro and Bates, 2000), the model structure is conditional on the random effects, making γ and ε_i inherently dependent.

Several approaches have been proposed in the literature to address these drawbacks, and procedures based on Cholesky decomposition of the random effects covariance matrix Δ_i provide additional advantage of guaranteeing the positive definiteness of the resulting factors, circumventing constraints of high dimensionality and statistical interpret-ability of the resulting parameters. We review some of these Cholesky-based procedures in the following section.

2. Variants of Cholesky decomposition

The standard Cholesky decomposition of a real, symmetric, positive definite matrix $\Sigma_{p\times p}$ is

$$\Sigma = U^T U \tag{2.1}$$

where U is an upper triangular with positive diagonal elements. The main advantage of this decomposition when used in parameter estimation procedures, such as maximum likelihood (ML) and residual/restricted maximum likelihood (REML), is that it provides an unconstrained parameterization of the parameters in Σ, circumventing the positive definiteness constraint. However, Pinheiro and Bates (1996) showed that the Cholesky factors U are not unique and the unconstrained $\frac{p(p+1)}{2}$ parameters lack a meaningful statistical interpretation with respect to the entries in Σ. To overcome these drawbacks, several classical and Bayesian approaches have been proposed in the literature.

Under the classical approach, Pourahmadi (1999) proposed the modified Cholesky decomposition (MCD) for modelling parameters of the precision matrix and developed a ML procedure (Pourahmadi, 2000) for normal generalized linear models:

$$\Sigma^{-1} = L^T D^{-1} L \tag{2.2}$$

where entries in the unit lower triangular Cholesky factor L are interpreted as negatives of autoregressive coefficients when a response variable y_t is regressed on its predecessors $y_{t-1},\dots,y_1$ and entries on the diagonal factor D as logarithms of their innovations.

However, despite the good performance of the proposed MCD, the procedure left a number of questions unanswered:

First, the proposed ML estimation procedure (and its restricted extensions) works well only when measurement times are identical across subjects, and hence may not be applicable to unbalanced data sets, particularly since it utilizes the sample covariance matrix S^2 as an initial value for Σ, which does not exist in unbalanced data settings (Pan and MacKenzie, 2006). In such cases, either the ML procedure is enhanced (Holan and Spinka, 2007) or some numerical optimization approach is adopted (Zimmermann et al., 1998). Second, the use of a saturated mean structure may be unnecessary (Pan and MacKenzie, 2003), except when the mean model is incorporated in searching the joint mean-covariance parameter space. Third, the regressogram, as proposed and utilized in model selection for the dependence and innovation variance, failed to capture the joint mean-covariance structure since the mean model was not included. However, Garcia et al. (2012) showed that as a data-driven graphical tool for model selection, they are powerful graphical tools in joint mean-covariance model selection for incomplete longitudinal data. Fourth, when subject-specific characterization is the focus of research interest, a linear mixed modelling (LMM) framework may be more flexible than a generalized linear modeling (GLM) framework.

These questions raised a number of issues and stimulated keen interest in modeling covariance structures under different frameworks and perspectives. Pan and Mackenzie (2003) observed that parameter estimates based on MCD are not optimal, and to address the first question, they proposed extending the procedure to unbalanced data with optimal parameter estimates achieved via joint search of the mean-covariance space. Zhang and Leng (2012) proposed a moving average

Cholesky decomposition (MACD) for Σ as the inverse of precision matrix:

$$\Sigma = L^{-1}DL^{-T} \tag{2.3}$$

where the entries in L^{-1} have a moving average (MA) interpretation. It has the same advantages and limitations as MCD and only differs in its MA interpretation. Li and Pourahmadi (2013) utilized MCD in developing a procedure that circumvent the effect of violating normality assumption on random effects in linear mixed models. However, their procedure was based on the assumption that design matrices X and Z share no common covariates and may result into inefficient parameters when such assumptions are violated. More recently, Lee et al. (2017) proposed an autoregressive moving average Cholesky decomposition (ARMACD) by combining the modified Cholesky decomposition and moving average Cholesky decomposition to address high-dimensionality and positive definiteness constraints:

$$\Lambda_i\Sigma_i\Lambda_i^T = L_iD_iL_i^i \tag{2.4}$$

where Λ_i is unit lower triangular matrix with generalized autoregressive parameters (GARPs) $-\phi_{i,tj}$ at its $(t,j)^{th}$ position, L_i is unit lower triangular with generalized moving average parameters (GMAPs) $\iota_{i,tj}$ at its $(t,j)^{th}$ position $(j<t)$ and $D_i = diag(\sigma_{i1}^2,\ldots,\sigma_{in_i}^2)$ is diagonal with innovation variances (IV) σ_{ij}^2. This decomposition subsumes a wide variety of covariance structures which are more flexible and with better forecasting performance than separate higher order AR or MA models. The combination of MCD and MACD creates a unified framework with models that allow nonstationarity and heteroscedasticity in parameter estimates.

Under a Bayesian framework, Daniels and Zhao (2003) proposed modelling the random effects covariance matrix Δ_i for the i^{th} subject $(i = 1,\ldots,m)$ using the modified Cholesky decomposition

$$\Lambda_i\Delta_i\Lambda_i^T = D_i \tag{2.5}$$

where $D_i = diag(\sigma_{i1}^2,\ldots,\sigma_{iq}^2)$ is diagonal with innovation variances (IV) σ_i^2 and Λ_i is unit lower triangular with GARPs $-\phi_{i,tj}$ as its $(t,j)^{th}$ entry. This variant also provides the advantages of overcoming the positive definite constraint and statistical interpretation of parameters. By adopting a Bayesian approach, they gained additional flexibility in obtaining the sampling distribution of the random effects using a simple Gibbs sampler, which sample from the posterior distribution of parameters. To incorporate heterogeneity in Δ_i, the random effects were allowed to depend only on subject-specific covariates and parsimony was achieved by regressing the parameters using these covariates. However, when random effects differ on different linear combinations of these covariates, separate covariance structures need to be fit for each combination and misspecification of a structure can lead to inefficient parameter estimates. Another drawback is their assumption of statistical independence of the repeated measurement given the random effects. This assumption restricted the application of their approach in longitudinal studies. Chen and Dunson (2003) proposed an alternative Cholesky decomposition for selecting the random effects components. Their approach factored the random effects covariance matrix Δ into

$$\Delta = D\Lambda\Lambda^T D \tag{2.6}$$

where D is a diagonal matrix with elements proportional to standard deviation of the random effects and Λ is a unit lower triangular matrix with off-diagonal elements describing correlations among the random effects. With separate factors for variance and correlation, their approach is computationally more tractable and provides some flexibility in selecting the random effects components. However, their approach is based on the assumption that components of the random errors and random effects are mutually independent. Gaskins and Daniels (2013) extended the Cholesky-based joint mean-covariance modelling to longitudinal data from several groups of subjects. They proposed a data-driven nonparametric method that simultaneously estimates the covariance matrix from each group by developing nonparametric priors using the matrix stick-breaking process. More recently, Han and Lee (2016) proposed a variant ARMACD decomposition:

$$\Lambda_i\Sigma_i\Lambda_i = C_i\Delta_b C_i^T + L_i D_i L_i^T \tag{2.7}$$

where Λ_i, L_i and D_i were as described above, Δ_b is the random effects covariance matrix and $C_i = (c_{i1},\ldots,c_{in_i})^T$ is the random effects design matrix. This approach has all the advantages of ARMACD as proposed by Lee et al. (2017), but model parameters are obtained conditional on random effects in the linear mixed model. By allowing X and Z to have common covariate(s), Gelfand et al. (1995) proposed a modelling procedure in which parameters are hierarchically centred to account for between-level correlations, and to ensure model identification. However, the random effects covariance matrix of their proposed procedure is positive semi-definite, leading to poor convergence properties in some parameter estimates.

To account for heteroscedasticity in the random errors ε_i via modelling the variance function, Pinheiro and Bates (2000) proposed a variant Cholesky decomposition of the random errors covariance matrix Σ_i as

$$\Sigma_i = D_i C_i D_i \tag{2.8}$$

where D_i is diagonal, describing the variance of the random errors and C_i is triangular with all diagonal elements positive, describing the correlation of the random errors. The variance function model was proposed, conditional on the random effects γ_i, as a function of the conditional population mean response $\mu_{ij} = E(y_i|\gamma_i)$. Now, with these conditional dependencies, the assumption of independence between ε_i and γ_i no longer holds and $Cov(\varepsilon_i,\gamma_i) \neq 0$. To circumvent the consequences of this violation, they allow common covariates between X and Z, approximating μ_{ij} by the best linear unbiased predictor (BLUP)

$$\hat{\mu}_{ij} = x_{ij}^T\beta + z_{ij}^T\gamma_i \tag{2.9}$$

where x_{ij} and z_{ij} denote the j^{th} rows of X_i and Z_i, respectively.
In this article, we propose a more efficient approach to address the violation of this key assumption. We improve efficiency in inference and gain more insight into model bahaviour by modelling the correlation structure between ε_i and γ_i when $Cov(\varepsilon_i, \gamma_i) \neq 0$, and to achieve this, we propose a linear Cholesky decomposition of the random effects covariance matrix Δ_i . Our approach is based on linear transformation of inner products of functions of Cholesky factors when subjected to left-right or lu decomposition. We next discuss this transformation.

3. Linear Cholesky decomposition

The proposed linear decomposition is based on upper triangular Cholesky factor U. Let $\Delta = U^T U$ be $q \times q$ standard Cholesky decomposition with U upper triangular. We subject U to lu factorization, obtaining:

$$\Delta = [lu(U)]^T \, lu(U) = (\Phi + \Psi)^T \, (\Phi + \Psi) \tag{3.1}$$

$$\Delta^{-1} = \left[lu\left(U^{-1}\right)\right]^T \left[lu\left(U^{-1}\right)\right] = \left(\rho + \Psi^{-1}\right)^T \left(\rho + \Psi^{-1}\right) \tag{3.2}$$

where $\Phi(-\phi_{i,kj})$ and $\rho(\theta_{i,kj})$ are upper triangular with zeros on the diagonal, parameters $-\phi_{i,kj}$ and $\theta_{i,kj}$ in $(k,j)^{th}$ positions, respectively, and innovation $\Psi = diag(\psi_{11}, \ldots, \psi_{qq})$ with $\Psi^{-1} = diag(\psi_{11}^{-1}, \ldots, \psi_{qq}^{-1})$ as its inverse.
Expanding (3.1) and (3.2), we obtain

$$\Delta = (\Phi + \Psi)^T (\Phi + \Psi) = \Phi^T \Phi + \Phi^T \Psi + \Psi^T \Phi + \Psi^T \Psi \tag{3.3}$$

$$\Delta^{-1} = \left(\rho + \Psi^{-1}\right)^T \left(\rho + \Psi^{-1}\right) = \rho^T \rho + \rho^T \Psi^{-1} + \Psi^{-T} \rho + \Psi^{-T} \Psi^{-1} \tag{3.4}$$

Definition 1 *Let Δ be represented by (3.3). Linear Cholesky decomposition is defined by*

$$\Delta = \Phi^T \Phi + \rho^T \rho + \Psi^T \Psi \tag{3.5}$$

$$= \Delta_{AR} + \Delta_{MA} + \Delta_{IV} \tag{3.6}$$

with Cholesky factors Φ, ρ and Ψ, respectively, describing the correlation structure in γ_i, the correlation structure between ε_i and γ_i, and the innovation of γ_i.

Also, for the precision matrix we have:

Definition 2 *Let Δ^{-1} be represented by (3.4). Linear Cholesky decomposition is defined by*

$$\Delta^{-1} = \rho^T \rho + \Phi^T \Phi + [\Psi]^{-T} \Psi^{-1} \tag{3.7}$$

$$= \Delta_{MA} + \Delta_{AR} + \Delta_{IV}^{-1} \tag{3.8}$$

The following theorem provides the basis for linear Cholesky decomposition:

Theorem 1 *Linear Cholesky decomposition of real, symmetric positive definite Δ is*

$$\begin{aligned} \Delta(\Theta) &= \Phi^T\Phi + \rho^T\rho + \Psi^T\Psi \\ &= \Delta_{AR} + \Delta_{MA} + \Delta_{IV} \end{aligned} \tag{3.9}$$

where $\Theta = (-\phi_{i,j}, \theta_{i,j}, \psi_{ii})$ with $-\phi_{ij} = corr(\gamma_i, \gamma_j)$ for $i \neq j$, $\theta_{ij} = corr(\varepsilon_{ik}, \gamma_{jk})$ and $\psi_{ii} = log\,(diag\,[\Delta_{ii}])$.

Proof 1 *Let $\Sigma_{p\times p} = U^TU$ be the standard Cholesky decomposition, then lu decomposition of upper triangular U results into ul (upper-lower) factors (see Stewart (1998), page 183):*

$$[lu(U)]^T\, lu(U) = (U^*\Psi)^T\,(U^*\Psi) \tag{3.10}$$

$$= \Psi^T U^{*T} U^* \Psi \tag{3.11}$$

with U^* unit upper triangular and Ψ lower triangular (diagonal) matrices.
Let 'b' be $p\times 1$ suitably chosen vector such that columns of U^* can be sequentially extracted via repeated multiplication, forming a Krylov sequence $b, U^*b, U^{*2}b, \ldots, U^{*(p-1)}b$. Define as Krylov matrix

$$\mathscr{K} = \left[b, U^*b, U^{*2}b, \ldots, U^{*(p-1)}b\right] = \langle b, U^*b\rangle \tag{3.12}$$

where $\langle .,.\rangle$ is an inner product, with the columns forming an ordered basis whose linear combinations span the Krylov subspaces $\mathscr{K}_1, \ldots, \mathscr{K}_{p-1} = \mathsf{K}^{(p-1)} \subseteq \mathrm{F}^p$, where F^p is $p - dimensional$ vector field. lu decomposition described by (3.11) is nonlinear in the factors. For a linear decomposition, we have from (3.3)

$$\Delta = \Phi^T\Phi + \underbrace{\Phi^T\Psi + \Psi^T\Phi} + \Psi^T\Psi$$

where Φ is strictly upper triangular with dependence parameters and Ψ is diagonal with variance parameters. The under-braced equation is a function of inner products of the respective Cholesky factors. There are several, well-established matrix linear transformations (such as Lyapunov stability transformation, see Carlson and Datta (1979)) that can be used in obtaining meaningful interpretation of this function of inner products.
If we let each product describe the rate of change in value of the correlation parameters θ between ε_i and γ_i (through shared covariate(s)) as

$$\partial f\,[\Phi(\phi)] = \left[\Phi^T(\theta)\Psi(\theta)\right]\partial\theta \tag{3.13}$$

$$\partial f\,[\Psi(\psi)] = \left[\Psi^T(\theta)\Phi(\theta)\right]\partial\theta \tag{3.14}$$

then, we obtain the correlation between ε_i and γ_i (through shared covariates, with respect to parameters in Φ) using a direct differentiation result by De Hoog et al. (2011), as

$$\begin{aligned}\left[\Phi^T(\theta)\Psi(\theta)+\Psi^T(\theta)\Phi(\theta)\right]\partial\theta &= \partial f\left[\Psi(\theta)\Phi(\theta)\Psi^T(\theta)\right]\\ &= \Phi^T(\theta)\Psi(\theta)\Phi(\theta)\\ &= \langle\Phi,\Psi\Phi\rangle \qquad (3.15)\end{aligned}$$

The columns of $\Phi(\theta)=\left[0,U^*b,U^{*2}b,\ldots,U^{*(n-1)}b\right]$ form linear combinations that span the same Krylov subspaces, but $\Phi(\theta)$ has the first column as zero vector, with zeros on its diagonal while $U^*\left(u^*_{ij}\right)$ is a unit triangular with the first column as a unit vector, having the first entry 1. However, congruence of $\langle b,U^*b\rangle$ to $\langle\Phi(\theta),\Psi(\theta)\Phi(\theta)\rangle$ implies congruence of $U^*(u^*_{ij})$ to $\Phi(\theta)$ and can be exploited in establishing an equivalence relation between them by changing the basis of U^*: Define a new basis $Z=(z_1,\ldots,z_p)\in\mathsf{K}^{(p-1)}$ for the column space of U^* and let $Z=SU^*$ where S is invertible, then

$$f(Z)=Z^T\Psi Z=\langle Z,\Psi Z\rangle=\langle SU^*,\Psi SU^*\rangle$$

Now, $\langle SU^*,\Psi SU^*\rangle$ is congruent to $\langle\Phi(\theta),\Psi(\theta)\Phi(\theta)\rangle$ if there exists a non-singular matrix Q such that

$$\begin{aligned}Q\Phi(\theta) &= SU^*Q \;\Rightarrow\; \Phi(\theta) = Q^{-1}SU^*Q\\ Q\Psi(\theta)\Phi(\theta) &= \Psi(\theta)SU^*Q \;\Rightarrow\; \Psi(\theta)\Phi(\theta) = Q^T\Psi(\theta)SU^*Q \qquad (3.16)\end{aligned}$$

Then, we have

$$\begin{aligned}\Phi^T(\theta)\Psi(\theta)\Phi(\theta) &= \left(Q^{-1}SU^*Q\right)^T\left(Q^T\Psi(\theta)SU^*Q\right)\\ &= Q^TU^{*T}S^TQ^{-T}Q^T\Psi(\theta)SU^*Q\\ &= (U^*Q)^T S^T\Psi(\theta)S(U^*Q)\\ &= (U^*Q)^T I(U^*Q)\\ &= (U^*Q)^T(U^*Q) \qquad (3.17)\end{aligned}$$

with invertible S diagonalizing $\Psi(\theta)$ to identity: $S^T\Psi(\theta)S\longrightarrow I$. The matrix $Q=[q_1,\ldots,q_p]$ is obtained using the Lanczos algorithm. The main advantage of the lu factorization is that columns of U can be reconstructed using the non-zero rows of U^* as coefficients and we exploit this feature in estimating MA parameters by reducing an upper Hessenberg matrix H to tridiagonal via column operations:

$$\begin{aligned} U^*Q &= [U^*q_1,\ldots,U^*q_p] \\ &= [q_1,\ldots,q_p]\begin{pmatrix} h_{11} & h_{12} & h_{13} & \ldots & \ldots & h_{1p} \\ h_{21} & h_{22} & h_{23} & \ldots & \ldots & h_{2p} \\ 0 & h_{32} & h_{33} & \ldots & \ldots & h_{3p} \\ \ldots & 0 & h_{43} & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & h_{(p-1)p} \\ 0 & 0 & \ldots & \ldots & h_{(p-1)p} & h_{pp} \end{pmatrix} \\ &= QH = \rho(\theta) \end{aligned} \tag{3.18}$$

where $\rho(\theta)$ is tridiagonal. With U^* being symmetric,

$$H = Q^{-1}U^*Q$$

is also symmetric and tridiagonal. Note that Q need not be orthogonal. Now, using the above relations, we have

$$\begin{aligned} \Phi^T(\theta)\Psi(\theta)\Phi(\theta) &= (U^*Q)^T\,(U^*Q) \\ &= \rho^T(\theta)\rho(\theta) \end{aligned} \tag{3.19}$$

To ensure that our approach also provides the basic advantage offered by Cholesky decompositions, we show that Δ_i is positive definite:

Theorem 2 *Let Δ_i be represented by the linear Cholesky decomposition (3.5). Then, Δ_i is positive definite.*

Proof 2 *By definition (3.5), $\Delta_i = \Phi^T\Phi + \rho^T\rho + \Psi^T\Psi$. Then, for any conformable nonzero vector x, we have*

$$\begin{aligned} x^T\Delta x &= x^T\left(\Phi^T\Phi + \rho^T\rho + \Psi^T\Psi\right)x \\ &= x^T\Phi^T\Phi x + x^T\rho^T\rho x + x^T\Psi^T\Psi x \\ &= (\Phi x)^T\,\Phi x + (\rho x)^T\,\rho x + (\Psi x)^T\,\Psi x \\ &= Y_1^T Y_1 + Y_2^T Y_2 + Y_3^T Y_3 \\ &= \Sigma_i y_{1i}^2 + \Sigma_j y_{2j}^2 + \Sigma_k y_{3k}^2 > 0 \end{aligned} \tag{3.20}$$

where $Y_1 = \Phi x, Y_2 = \rho x, and\ Y_3 = \Psi x$. Therefore, $x^T\Delta x > 0$ and Δ is positive definite.

4. Conclusions

We propose a linear Cholesky decomposition for estimating correlation parameters between random errors ε_i and random effects γ_i in linear mixed models when the independence assumption between the two does not hold. Our approach can be regarded as an extension of the Pinheiro and Bates (2000) result, in which their Cholesky decomposition of Σ_i has two factors, while our decomposition of Δ_i has three factors. Application of this decomposition to parameter estimation using the maximum likelihood and restricted/residual maximum likelihood procedures is the topic of our ongoing research.

Acknowledgements

We would like to thank the editorial office and two referees for their constructive feedback, which considerably improved the original version of the paper. The first author is grateful to Umaru Musa Yar'adua University Katsina for funding his research.

REFERENCES

CARLSON, D. H., DATTA, B. N., (1979). The Lyapunov matrix equation $SA + A^*S = S^*B^*BS$. Linear Algebra and Its Applications, 28, pp. 43–52.

CHEN, Z., DUNSON, D. B., (2003). Random effects selection in linear mixed models. Biometrics, 59, pp. 762–769.

DANIELS, M. J., ZHAO, Y. D., (2003). Modeling the random effects covariance matrix in longitudinal data. Statistics in Medicine, 22(10), pp. 1631–1647.

DE HOOG, F. R., ANDERSSEN, R. S., LUKAS, M. A., (2011). Differentiation of matrix functions using triangular factorization. Mathematics of Computation, 80(275), pp. 1585–1600.

GARCIA, T. P., KOHLI, P., POURAHMADI, M., (2012). Regressogram and mean-covariance models for incomplete longitudinal data, Journal of the American Statistical Association, 66(2), pp. 85–91.

GASKINS, J. T., DANIELS, M. J., (2013). A nonparametric prior for simultaneous covariance estimation, Biometrika, 100(1), pp. 125–138.

GELFAND, A. E., SAHU, S. K., CARLIN, B. P., (1995). Efficient parameterization for normal linear mixed models, Biometrika, 82(3), pp. 479–488.

HAN, E-J., LEE, K., (2016). Dynamic linear mixed models with ARMA covariance matrix, Communications for Statistical Applications and Methods, 23(6), pp. 575–585.

HOLAN, S., SPINKA, C., (2007). Maximum likelihood estimation for joint mean-covariance models for unbalanced repeated measures data, Statistics and Probability letters, 77, pp. 319–328.

LAIRD, N. M., WARE, J. H., (1982). Random-effects models for longitudinal data, Biometrics, 38(4), pp. 963–974.

LANGE, N., RYAN, L., (1989). Assessing normality of random effects models, Annals of Statistics, 17, pp. 624–642.

LEE, K., BAEK C., DANIELS, M. J., (2017). ARMA Cholesky factor models for the covariance matrix of linear models, Computational Statistics and Data Analysis, 115, pp. 267–280.

LI, E., POURAHMADI, M., (2013). An alternative REML estimation of covariance matrices in linear mixed models, Statistics and Probability Letters, 83, pp. 1071–1077.

PAN, J-X., MACKENZIE, G., (2003). On modeling mean-covariance structures in longitudinal studies, Biometrika, 90(1), pp. 239–244.

PAN, J-X., MACKENZIE, G., (2006). Regression models for covariance structures in longitudinal studies, Statistical Modeling, 6, pp. 43–57.

PINHEIRO, J. C., BATES, D. M., (1996). Unconstrained parameterizations for variance-covariance matrices, Statistics and Computing, 6, pp. 289–296.

PINHEIRO, J. C., BATES, D. M., (2000). Mixed-effects models in S and S-Plus, Springer-Verlag, New York, pp. 205–207.

POURAHMADI, M., (1999). Joint mean-covariance models with application to longitudinal data: Unconstrained parameterization, Biometrika, 86(3), pp. 677–690.

POURAHMADI, M., (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix, Biometrika, 87(2), pp. 425–435.

STEWART, G. W., (1998). Matrix algorithms volume I: Basic decompositions, SIAM Publications, Philadelphia, U.S.A., p. 183.

ZHANG, W., LENG, C., (2012). A moving average Cholesky factor model in covariance modelling for longitudinal data, Biometrika, 99(1), pp. 141–150.

ZIMMERMANN, D. L., VINCENT, N-A., HAMMOU, E., (1998). Computational aspects of likelihood-based estimation of first order antedependence models, Journal of Statistical Computation and Simulation, 60(1), pp. 67–84.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 71–87, DOI 10.21307/stattrans-2019-035
Submitted – 14.05.2019; Paper ready for publication – 01.10.2019

MODELLING LANGUAGE EXTINCTION USING SUSCEPTIBLE-INFECTIOUS-REMOVED (SIR) MODEL

N. A. Ikoba[1], E. T. Jolayemi[2]

ABSTRACT

The study presents a stochastic epidemic model applied to the model of indigenous language extinction. The Susceptible-Infectious-Removed (SIR) categorization of an endemic disease has been reformulated to capture the dynamics of indigenous language decline, based on the assumption of non-homogeneous mixing. The time in which an indigenous language is expected to be extinct was derived using a modified SIR model with the population segmented into several sub-communities of small sizes representing family units. The data obtained from the 2016 indigenous language survey conducted in several parts of Nigeria and from the 2013 Nigeria Demographic Health Survey (NDHS) were used to estimate the key parameters of the model for Nigeria's several indigenous languages. The parameters of interest included the basic reproduction number, the threshold of endemicity, and the time in which a language is expected to be extinct, starting from the endemic level. On the basis of the time in which a language is expected to be extinct, several of the surveyed languages appeared to be in a precarious condition, while others seemed virile, thanks to a high language transfer quotient within families.

Key words: language extinction, stochastic epidemic model; non-homogeneous mixing, quasi-stationary distribution, time in which a language is expected to be extinct.

1. Introduction

The world today is littered with thousands of languages and several hundreds have been documented to have become extinct (Crystal, 2000). Of the known 7,102 living languages, 22% of them have been categorized as 'in trouble', 13% dying, while there is a loss rate of about 6 languages per year (Lewis *et al.*, 2015).

It is claimed that more than 400 Nigerian languages are endangered and there is a declining level of transfer of indigenous language ability to the younger generation (Ohiri-Aniche, 2014). It was therefore projected that many Nigerian languages will become extinct by 2084 (Ohiri-Aniche, 2014).

A lot of languages globally are tethering on the brink of extinction (Nuwer, 2014). Over the past century alone, it is estimated that around 400 languages -

[1] Department of Statistics, University of Ilorin, Ilorin, Nigeria. E-mail: ikoba.na@unilorin.edu.ng.
[2] Department of Statistics, University of Ilorin, Ilorin, Nigeria.

about one every three months - have gone extinct, and it is estimated that 50% of the world's remaining 6,500 languages will be gone by the end of the 21st century (Nuwer, 2014).

Languages usually reach the point of crisis after being displaced by a socially, politically and economically dominant one. Sometimes, especially in immigrant communities, parents will decide not to teach their children their heritage language, perceiving it as a potential hindrance to their success in life (Nuwer, 2014).

The problem of declining use of indigenous languages at the expense of global languages like English and French has become much pronounced in previously colonized countries in the developing world.

While there is a common view that many indigenous languages are dying, the depth of the problem is yet to be sufficiently captured, as there has been minimal use of scientific approaches to reflect the gravity of the problem especially among indigenous languages of sub-Saharan Africa. Thus, the desire to provide an analytical model that sufficiently captures the language decline in a society over an extended period of time is the main motivation for this research.

A research that motivated the pursuit of language modelling using epidemic models is the work of Daley and Kendall (1964) which modelled the spread of rumours using the basic SIR epidemic model.

The main focus in the adaptation of stochastic epidemic models to study the decline of indigenous languages is to derive conditions under which the indigenous language can be propagated without the threat of extinction. In other words, it is of interest to estimate thresholds above which such languages will continue to survive with minimal fluctuations around the threshold.

The SIR model was originally used to model the dynamics of an epidemic within a population consisting of susceptible individuals (S), infectious individuals (I) and those who have recovered or are removed (R) and could no longer contribute to the spread of the epidemic. There are numerous versions of the SIR epidemic model to capture different disease dynamics for both closed populations and populations that incorporate demographic turnover.

Under the context of indigenous language decline in a community, using the principle of the SIR epidemic model, the susceptible group is viewed as all children born into the community; the infectious group corresponds to children who later acquire indigenous language ability; while the removed group contains those who exit the community either through death or migration.

There are several useful similarities between the study of epidemics and language decline, with seemingly opposite objectives. While under the epidemic context, the goal of modelling is to ensure the termination of the spread of the disease in the population, in the context of language modelling, the goal is to ensure continuous propagation of the language to enhance its survival in the population. It is noted that this is a novel application of stochastic epidemic models to capture indigenous language dynamics. A search of the literature reveals that there are seemingly no papers applying epidemic models in the area of indigenous language extinction. A few of the relevant researches germane to this study are now presented.

Trapman and Bootsma (2009) established a relation between the spread of infectious diseases and the dynamics of the M/G/1 queueing system with processor sharing. They showed that the number of infectious individuals in a standard SIR epidemic model at the moment of first detection of the epidemic was geometrically distributed. They derived the distribution of the number of infectious individuals at the moment of first detection in a broad class of epidemics in large populations.

An inherent relationship exists between infectious diseases and human populations, as reflected by Dobson and Carper (1996). The relationship can be extended to that of languages and human populations.

Allen and Burgin (2000) compared the dynamics of some deterministic and stochastic discrete-time epidemic models with fixed and varying population. In some cases, the time to extinction was very long, and in such cases, if the probability distribution is conditioned on non-extinction, then for values of the basic reproduction number $R_0 > 1$, there exists a quasi-stationary distribution whose mean agrees with the deterministic endemic equilibrium. The expected duration of the epidemic was also investigated numerically.

Ball and Lyne (2000) analyzed the spread of an SIR epidemic in a closed, finite population using a household model. Both local and global infection was possible within the population. We are interested in such household models that could be modified to fit the goals of language modelling.

Nasell (2002) studied several stochastic models with demography for various endemic infections. Approximations of the quasi-stationary distributions and expected time to extinction were derived for the SI, SIS, SIR and SIRS models. The approximations were valid for sufficiently large population sizes, and in comparison with corresponding deterministic models, the stochastic models provided realistic parameter estimates.

Nasell (2005) examined quasi-stationarity and time to extinction for the classic endemic model, with focus restricted to the transition region in the parameter space where the quasi-stationary distribution is non-normal. An approximation of the marginal distribution of the infected individuals in quasi-stationarity was proposed. Simulation results showed that the analytical approximations performed reasonably.

Verdasca et. al (2005) studied the effect of spatial correlations on the spread of infectious diseases using a stochastic SIR model on complex networks. Heavy stochastic fluctuations tend to limit the utility of deterministic models under such circumstances.

Lindholm and Britton (2007) studied an SIR model with the population subdivided into k sub-communities of equal sizes n, assumed large. Lindholm and Britton (2007) model provides a useful platform for the adaptation of epidemic modelling to language extinction. The general SIR model with homogeneous mixing is called SIR-HM and the SIR model with heterogeneous mixing in which the population is divided into sub-communities is called SIR-SC (Lindholm and Britton, 2007).

Burr and Chowell (2008) analyzed SEIR-type models under the assumption of non-homogeneous mixing. Their goal was to evaluate possible retrospective signatures of non-homogeneous mixing behaviour. From simulated outbreaks, it was concluded that such signatures can detect non-SEIR-type behaviour in some of the social structures considered.

Schwartz et. al (2009) investigated random extinction processes in a class of epidemic models, examining the rate of disease extinction as a function of disease spread. It was shown that the effective entropic barrier for extinction in a SIS model displays scaling with the distance to the bifurcation point, with an unusual critical component. Analytical results were compared with numerical simulations and were found to be good.

Billings et. al (2013) considered the effect of randomly distributed intervention as disease control on large finite populations, and showed how intervention control modulates the expected time to extinction, which in turn was a function of population size and rate of infection spread.

In section two, the relevant SIR model is conceptualized to capture the language decline dynamics. The model analysis is also presented, and estimates of the quasi-stationary distribution, threshold of endemicity and expected time to extinction are obtained. The model parameter estimation from survey and historical data is also presented in section 3. The results from an indigenous language survey carried out in some parts of Nigeria in order to determine the extinction status of some Nigerian languages are presented and discussed in section 3. Finally, our conclusions are presented in section 4.

2. SIR model with demography for language extinction

The main interest motivating the adaptation of the relevant SIR model to study indigenous language extinction is the very close relationship between epidemics and languages with regards to extinction. When demographic turnover and heterogeneities in the population are incorporated into SIR models for endemic diseases with infectious periods that are of the same order as the lifetime distribution of individuals, it is seen that such scenario will sufficiently capture the indigenous language decline dynamics in a population.

2.1. Description of the SIR model with demography and heterogeneous mixing

One of the fundamental assumptions of the standard SIR epidemic model is that the population is a closed one, that is, there is no demographic turnover (births, deaths, migrations) in the population. For the SIR model with demography, this assumption is relaxed in that there can be births of susceptible individuals into the population and also deaths or migration out of the population. This relaxed assumption therefore introduces population dynamics as a component of the evolution of the epidemic.

The SIR model with demography provides a useful adaptation of endemic diseases modelling to capture the decline of languages with some modifications. Andersson and Britton (2000) provided an excellent description of the SIR model with demography, stressing its appropriateness in modelling endemic diseases, which have a long infectious period in relation to the lifetime of the individual. They noted that when modelling the spread of an endemic disease in a very large population, demographic turnover cannot be ignored. A disease is called endemic if it is able to persist in a population for a long time, without the need of introducing new infectious individuals from some external population (Lindholm, 2007). When the susceptible population in a large community is augmented fast enough through births and/ or migrations, the epidemic tends to persist for a very long time even without the introduction of new infectious individuals into the population.

For the language scenario, the assumption is imposed that the infectious period and the residual lifetime of the individual coincide. In fact, the overall lifetime of an individual is the sum of the language acquisition (latency) period and the infectious period. The system is viewed at specified epochs.

The focus in the adaptation of stochastic epidemic models to study the decline of indigenous languages is to derive conditions under which the indigenous language can be propagated without the threat of extinction. In other words, interest is centred in achieving those thresholds above which such languages will continue to survive with minimal fluctuations around the threshold.

According to Lindholm and Britton (2007), for diseases in which the infectious period is of the same order with the lifetime distribution, one can assume that when an individual recovers from the infection, this individual will likely be removed due to death within a relatively short time period. This is a fundamental backbone of the language adaptation of stochastic epidemic models, as it fairly approximates the scenario in the language setting, where it is assumed that the infectious period and the residual lifetime of the individual are identically distributed.

The relevant SIR model with demography is now conceptualized in the indigenous language extinction approach.

The population in the community contains k families labelled 1,2,...,k. Let n_i be the size of the i^{th} family, then the size of the entire population is

$$N = \sum_{i=1}^{k} n_i$$

Individuals are born into the population at a constant rate μk and each of them is assumed to have an exponentially distributed lifetime with intensity θ.

Initially, it is assumed that there are zero susceptible and k indigenous language-speaking individuals in the population, denoting the initial language-speaking population in the community before the advent of external influences like colonization. A given indigenous language speaker stays ‘infectious’ for a time period that is exponentially distributed with intensity ν (unless he dies of other causes before the end of that period). During that time, the infective in the j^{th} family makes contact with children in his family at rate β/n_j. With probability p, a contacted child imbibes the language.

All random variables and counting processes (which are Poisson processes due to the exponential lifetimes assumption) are assumed to be mutually independent. Some of the relevant variables are next defined:

X(t)= Number of susceptibles at time t; Y(t)= Number of language speakers at time t.

$\{X(t), Y(t), t \geq 0\}$ is a Markov process with transition rates:

$$\left\{\begin{array}{llc} From & to & at\ rate \\ (s_j, i_j) & (s_j+1, i_j) & \mu n_j \\ (s_j, i_j) & (s_j-1, i_j) & \theta s_j \\ (s_j, i_j) & (s_j-1, i_j+1) & (\beta/n(1+\varepsilon(k-1)))s_j(i_j+\varepsilon\sum_{u\neq j} i_u) \\ (s_j, i_j) & (s_j, i_j-1) & (\theta+\nu)i_j \end{array}\right. \tag{2.1}$$

We are mainly interested in the susceptible (S) and language-speaking (I) population, as the recovered/ removed population is assumed to have no further influence on the indigenous language dynamics.

Non-homogeneous mixing models such as those for outbreaks in social networks are often believed to provide better predictions of the benefits of the various mitigation strategies such as isolation or vaccination (Burr and Chowell, 2008). The homogeneous mixing assumption is often an unrealistic one, but due to the tractability of their analysis, such models provide useful insights on the disease dynamics. Social structure is a type of non-homogeneous mixing, such as most individuals having preferential contact with work or family members compared to the general population.

The application of the SIR model with demography and non-homogeneous mixing to the language scenario imply that an infectious individual can mainly infect his offspring. There is a sort of family-based transmission of language ability from parent to offspring. The dynamics of the spread of the language becomes more intricate with the assumption of non-homogeneous mixing. If the population is divided into families, some families may not have any indigenous language speaker (language-free), while most others will contain a mixture of language speakers and non-speakers. As the population evolves in time, the tendency is to have an increasing number of non-language speakers, mainly the young ones born into the population but unable to imbibe their indigenous language.

It is of interest to study situations where k is large in relation to the $n_j's$. This appropriately captures the scenario in communities where there are smaller family sizes in relation to the number of families in the community. It is also assumed that the contact rates within families are the same.

Define the parameter

$\varepsilon =$ proportion of an individual's contacts that are with other families.

$\varepsilon = 0$ implies that the families are isolated and there is no transmission between families. $\varepsilon = 1$ implies a single large community in which there is interaction among individuals in the population irrespective of their family.

The overall infectious pressure in the entire population is kept constant regardless of the value of ε.

The standard SIR model assumes homogeneous mixing in the population, therefore results from the analysis of the standard SIR model are not applicable when the assumption of homogeneous mixing is relaxed. An appropriate analysis of the SIR model with the population divided into sub-communities has been given by Lindholm and Britton (2007). For the case with sub-communities, an infected individual in the j^{th} family makes contacts with any given individual within its own family at rate β'/n_j and at rate $\varepsilon\beta'/n$ with a given individual in any of the k-1 surrounding sub-communities, where n is taken as the mean family size in the population.

Drawing from the principles of Lindholm and Britton (2007), the probability that a contact is within the j^{th} family is

$$\Pr(within\ contact) = \frac{1}{1+\varepsilon(k-1)}$$

and the basic reproduction number, R_0 is given as

$$R_0 = \frac{1}{(\theta+\nu)}\left(\frac{n\beta'}{n} + \frac{(k-1)n\beta'}{n}\right) = \frac{\beta'}{(\theta+\nu)}\big(1+\varepsilon(k-1)\big) \tag{2.2}$$

Define $\alpha = (\theta+\nu)/\theta$ as the ratio of mean lifetime to mean duration of language ability and $\beta = \beta'(1+\varepsilon(k-1))$, then the basic reproduction number can be expressed as

$$R_0 = \frac{\beta}{\theta\alpha} \tag{2.3}$$

which is of the same form as the SIR model with homogeneous mixing.

It is noted that the basic reproduction number, R_0 is an increasing function of ε. In fact, it can be seen from equation (2.2) that $R_0(1) = kR_0(0)$ and all other values of $R_0(\varepsilon)$ are bounded in the interval $(R_0(0), R_0(1))$. Hence, as the value of ε increases, the basic reproduction number also increases, with the additional property that $R_0(1)$ is an integer multiple of $R_0(0)$.

2.2. The quasi-stationary distribution

The quasi-stationary distribution is important when modelling endemic diseases, since interest is on the behaviour of the epidemic until it goes extinct. Similar reasoning also show that the quasi-stationary distribution in the language context is also of relevance, as a language that has progressed to the stage of quasi-stationarity has become endangered and is most likely to become extinct in the absence of any external revitalization measures. The quasi-stationary distribution of the population is defined as the conditional distribution that the process has not been absorbed after a long period of time. The endemic level can be thought of as the mean of this distribution, which the process fluctuates around (Lindholm and Britton, 2007).

Let $Q = \{q_{x,y}\}$ denote the quasi-stationary distribution.

$$q_{x,y} = \lim_{t\to\infty} Pr\{X(t) = x, Y(t) = y | Y(t) > 0\}$$

If in addition, the infectious period is exponentially distributed, then by reason of the memoryless property of the exponential distribution,

$$Pr(T_Q > t+s | T_Q > t, \big(X(0), Y(0)\big) \sim\ Q) = Pr(T_Q > s | \big(X(0), Y(0)\big) \sim\ Q)$$

where T_Q is the time to extinction of the language given that the process is in the quasi-stationary state.

A relevant proposition is next presented (Lindholm and Britton, 2007):

Proposition 1: The time to extinction given that the process is started in the quasi-stationary distribution T_Q, is exponentially distributed with mean

$$E(T_Q) = \tau = \frac{1}{\theta\,\alpha\,q_{.,1}} \tag{2.4}$$

where

$$q_{.,1} = \sum_x q_{x,1} \tag{2.5}$$

is the marginal probability that 1 indigenous language speaker is left in the population.

Lindholm and Britton (2007) derived an approximation of $q_{.,1}$, and thus T_Q using diffusion approximation.

Drawing from the central limit theorem, and with a population of size n, then approximations for the mean number of indigenous language speakers, μ_y and the standard deviation, σ_y are given as (Lindholm and Britton, 2007)

$\mu_y = n\frac{R_0-1}{\alpha R_0}$ and $\sigma_Y = \frac{\sqrt{n}}{R_0}\sqrt{}(R_0 - 1 + R_0^2/\alpha)$, and when $\alpha >> R_0$

$\sigma_Y \approx \sqrt{n(R_0-1)}/R_0$ and

$$q_{.,1} \approx \frac{R_0}{\sqrt{2\pi n(R_0-1)}}\exp\left(\frac{-n\,(R_0-1)}{2\alpha^2}\right) \tag{2.6}$$

The quasi-stationary distribution obtained by Lindholm and Britton (2007), as presented in equation (2.6) would not be appropriate in the case where the lifetime of the individual and the infectious period are of the same order (α small). This is the nature of the indigenous language dynamics, hence we propose an alternative approximation for the quasi-stationary distribution as

$$q_{.,1} \approx \frac{R_0}{\sqrt{2\pi n(R_0-1+R_0^2/\alpha)}}\exp\left(\frac{-n\,(R_0-1)^2}{2\alpha^2(R_0-1+R_0^2/\alpha)}\right) \tag{2.7}$$

which is better suited for the language characterization due to the fact that the residual lifetime and the period language ability are of the same order.
It is noted that equation (2.7) reduces to equation (2.6) when $\alpha >> R_0$.

2.3. Expected time to extinction of the language

As established in Andersson and Britton (2000), the expected time to extinction for the homogeneous mixing case of the SIR model, using the normal approximation (equation 2.6) yields

$$\tau \approx \frac{\sqrt{2\pi n(R_0-1)}}{\theta\alpha R_0}\exp\left(\frac{n\,(R_0-1)}{2\alpha^2}\right) \tag{2.8}$$

When the life length is long in relation to the length of the infectious period, the above approximation produces too wide estimates when the sub-community sizes are only moderately large (Lindholm and Britton, 2007). A better approximation, well suited for the language scenario with smaller family sizes, could be obtained using equation (2.7) instead:

$$\tau \approx \frac{\sqrt{2\pi n(R_0 - 1)}}{\theta\alpha R_0} \exp\left(\frac{n\,(R_0 - 1)^2}{2\alpha^2\left(R_0 - 1 + \frac{R_0^2}{\alpha}\right)} \right) \tag{2.9}$$

Under the scenario of long life length in relation to the length of the infectious period, Nasell (2005) proposed that the quasi-stationary distribution of the number of infected individuals could be approximated with a geometric distribution with $p = 1/\mu_Y$. If $Y \sim Geo(p)$, then $E(Y) = 1/p = \mu_Y$

When the quasi-stationary distribution of Y is approximated as such, then

$$\tau_n \approx \frac{n(R_0 - 1)}{\theta\alpha^2 R_0} \tag{2.10}$$

The expected time to extinction when all the k families are started at the endemic level, where there is a proportion ε of contacts between families, $\tau(\varepsilon)$ provides a credible challenge in obtaining the estimates of the mean time to extinction τ that is dependent on ε. In actual fact, $\tau(\varepsilon) = \tau(\varepsilon, n, k, \theta, \alpha, R_0)$.

When $\varepsilon = 0$, all k families are isolated and independent, and starting at the endemic level of infection, the expected time until one of the k families recovers is τ_n/k, due to independence and that the expected duration of an epidemic within a family is exponentially distributed with mean τ_n. Due to the assumption that the language-free states are absorbing states of the process, when $\varepsilon = 0$, the expected time until one of the k-1 remaining families recovers is $\tau_n/(k-1)$. Repeating the argument yields

$$\tau(0) = \tau_n \sum_{i=1}^{k} \frac{1}{i} \tag{2.11}$$

As the number of families, k increases, the sum $\sum_{i=1}^{k} \frac{1}{i}$ approaches a finite value, its limiting value. Hence, the time to extinction will increase with greater number of families even with zero interaction between families.

On the other hand, when $\varepsilon = 1$, all k families behave as one large community of size $\sum_{j=1}^{k} n_j$, and we can use the SIR-HM approximation of τ_n with n replaced by $\sum_{j=1}^{k} n_j$ (Lindholm and Britton, 2007).

$$\tau(1) = \tau_n$$

If the $n_j{}'s$, the size of the families are small, Lindholm and Britton (2007) suggested that the geometric approximation of τ_n may be used to approximate $\tau(0)$ and $\tau(1)$, to yield

$$\frac{\tau(0)}{\tau(1)} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{i}$$

which is less than 1 for $k > 1$, and this implies that $\tau(0) < \tau(1)$, hence $\tau(\varepsilon)$ is an increasing function.

For large sizes of the families, Lindholm and Britton (2007) also suggested that the truncated normal approximation of τ_n should be used. Furthermore, for population sizes such that $\mu_Y/\sigma_Y > 3$ and with $\alpha >> R_0$, τ_n is approximated by equation (2.8), and $\frac{\tau(0)}{\tau(1)}$ is also smaller than 1 for sufficiently large n.

As established by Andersson and Britton (2000), the stationary points of the process are the language-free state (1,0) and

$$(\hat{x},\hat{y}) = \left(\frac{\nu+\theta}{\beta}, \frac{\theta}{\nu+\theta}\left(1-\frac{\nu+\theta}{\beta}\right)\right) = \left(\frac{1}{R_0}, \frac{R_0-1}{\alpha R_0}\right) \tag{2.12}$$

It should be noted that the language-free state (1,0) is stable for $R_0 < 1$ and unstable for $R_0 > 1$. The point $(\hat{x},\hat{y})$ is stable for $R_0 > 1$ (and is otherwise negative).

If $R_0 < 1$, the language is predicted to die out fairly quickly in the community, and if $R_0 > 1$ then it will rise towards a positive level called the endemic level.

The basic reproduction number, R_0 works as a threshold determining the dynamics of the indigenous language, and is dimensionless. If $R_0 \leq 1$, the language will go extinct rather quickly. Otherwise the language has a positive probability to persist in the population over a long period of time.

Using principles from the Central Limit Theorem, let us assume that $R_0 > 1$ and suppose that the process is started close to the endemic level $(n\hat{x}, n\hat{y})$. The process is positively recurrent and it will become absorbed into the set of language-free states $\{(i,0), i \geq 0\}$ in finite time. Prior to absorption, small fluctuations may be observed around the endemic level and the nature of these fluctuations can be examined. Define for $t \geq 0$

$$\left(\tilde{X}_{t_i}, \tilde{Y}_{t_i}\right) = \sqrt{n}(\bar{X}_{t_i} - x(t_i), \bar{Y}_{t_i} - y(t_i))$$

Andersson and Britton (2000) established that $\left(\tilde{X}_{t_i}, \tilde{Y}_{t_i}\right)$ converges weakly on compact time intervals to a Gaussian process $(\tilde{X}, \tilde{Y})$ with mean vector $\underline{0}$ and covariance matrix Σ.

The time to extinction of the indigenous language, T_Q is defined as (Andersson and Britton, 2000)

$$T_Q = \inf\{i \geq 0 : Y(t_i) = 0\}$$

T_Q is finite for any fixed population size if $R_0 > 1$.

According to Andersson and Britton (2000), it is a classical (and very difficult) problem to obtain the estimates of T_Q. Not even the expected value $E(T_Q)$ is easily estimated. One approach is to let the population size to become very large and regard language extinction as the result of a large deviation from a high endemic level. In this regard, asymptotic approximations of $E(T_Q)$ are available. There is also an heuristic derivation of an approximate expression for $E(T_Q)$ (Nasell, 1999), noting that the coefficient of variation of the number of infectious individuals in endemicity is given by

$$\frac{\sqrt{n\hat{\Sigma}_{22}}}{n\hat{y}} = \frac{\sqrt{n\left(R_0 - 1 + \frac{R_0^2}{\alpha}\right)}}{R_0} \frac{\alpha R_0}{n(R_0 - 1)} \approx \frac{\alpha}{\sqrt{n(R_0 - 1)}} \tag{2.13}$$

The last approximation is appropriate when $\alpha >> R_0^2$.

For a real-life disease in a community with heterogeneous mixing, extinction may be caused by a normal fluctuation from a not so high endemic level or by a large deviation from a high level. In other words, in the absence of a catastrophe that may wipe out the disease and the population, the extinction of the disease follows a gradual process.

Simulation results from Andersson and Britton (2000) also show that, for realistic parameter values, the Nasell (1999) formula gives a much better approximation to the observed time to extinction than does the formula derived by van Herwaarden and Grasman (1995).

The distribution of T_Q depends on the parameters $R_0, \alpha,$ and θ as well as the size of the population. The extinction times will increase as the basic reproduction number increases. In fact, a population that sufficiently reproduces itself will have a high value of R_0, which in turn, will produce higher extinction times.

2.4. Threshold of endemicity of the language

At endemicity, the system fluctuates minimally around the endemic level. For the process $\{X(t), Y(t), t \geq 0\}$, we are interested in the proportion of indigenous language speakers at time t. We introduce the random variable

$$Z(t) = \frac{Y(t)}{X(t) + Y(t)} = \frac{Y(t)}{N(t)}$$

At time t=0, it is assumed that there are no susceptible individuals and m language-speaking individuals, that is, the initial population had indigenous language ability. This implies that $Z(0) = 1$. The value of the random variable $Z(t)$ at the endemic level, ($\hat{z}(t)$) is the threshold beyond which the language becomes endangered, called the *threshold of endemicity*. Once this threshold is reached, there is a high probability that the language will become extinct in the population in finite time, in the absence of any external intervention.

The threshold of endemicity is given as

$$\hat{z} = \frac{\hat{y}}{\hat{x} + \hat{y}} = \frac{R_0 - 1}{\alpha + R_0 - 1} \tag{2.14}$$

Once a language gets to the endemic level, then the possibility of extinction from that point onwards can be very high. At the endemic level, there are small fluctuations before extinction. The endemic level is a somewhat stationary position of the process.

For language survivability, there should only be a small proportion of families in the language-free state.

2.5. Parameter estimation

In order to estimate the parameters of the SIR model of language extinction for a specified indigenous language, input parameters like the birth rate, life expectancy, language transfer rate, mean language-speaking period, mean family sizes, etc, have to be extracted from historical data or via a survey of the language. These parameters are the inputs needed to obtain the estimates derived in the previous section.

Using these input parameters, the basic reproduction number is computed, the expected time to extinction is also obtained, as well as all the other relevant metrics to ascertain the virility of the indigenous language, as described in the previous section.

3. Results and discussions

An indigenous language survey was conducted in some cities in Nigeria in 2016 and language use data was extracted for some tribes over two generations. The questionnaire was constructed such that a respondent could provide demographic information and details on the language ability of his siblings and children, if any. In addition, data on the fertility profile of Nigerian women and average life expectancy of Nigerians, obtained from the 2013 Nigeria Demographic Health Survey (NDHS) (NPC, 2014) were used as inputs to estimate the model parameters for the surveyed languages.

The surveyed languages were Yoruba, Igbo, Bini, Urhobo, Esan and Isoko. Apart from the major languages of Yoruba and Igbo, the other languages emanate mainly from Edo and Delta states in the Niger Delta region of Nigeria.

The questionnaire contained 19 brief questions with the goal of eliciting information on the basic demographic characteristics of the respondent, indigenous language use ability of the respondent's sibling as well as the respondent's children, if any. The questionnaire also contained questions relating to the possible reasons for lack of intergenerational transfer of language ability from parent to children. A total of 607 respondents provided language use data on themselves, their parents, siblings and their children, if any. Hence, the data contained information on over 5,000 persons across three generations.

Data from the indigenous language survey questionnaire were processed via the Statistical Package for the Social Sciences (SPSS) and R computing software to yield the relevant estimates of the conceptualized model parameters.

Table 1 is a summary of the estimates of the model input parameters for the group of parents and non-parents (first and second generation, respectively) surveyed in the various languages. Table 2 captures the estimates of the quasi-stationary distribution, time to extinction, and threshold of endemicity, for the two generations. The estimate of the mean lifetime was 54 years (NPC, 2014), and this was assumed to be the same for all the languages.

Table 1. Estimates of the model input parameters (mean family size n, birth rate μ, indigenous language transfer rate β, basic reproduction number R_0, mean duration of the language transmission period $1/\nu$, and the ratio mean lifetime to the mean period of language ability α) for the first (1st) and second (2nd) generations respectively

Language	n		μ		β		$R_0(0.001)$		R_0		$1/\nu$	α
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd		
Igbo	8	5	5.97	2.82	5.61	1.80	4.22	1.35	3.83	1.23	25	3.16
Yoruba	8	5	5.84	3.04	5.39	1.44	4.00	1.80	3.64	1.64	26	3.08
Urhobo	9	5	6.27	3.04	4.43	1.44	3.33	1.08	3.03	0.98	25	3.16
Esan	9	6	6.20	3.78	5.57	2.61	4.08	1.91	3.71	1.74	27	3.00
Bini	9	5	7.09	3.00	6.46	1.50	4.68	1.09	4.25	1.00	28	2.93
Isoko	10	5	7.21	2.25	5.90	1.06	4.49	0.81	4.08	0.73	24	3.25
Others	10	5	7.11	3.17	6.29	1.83	4.61	1.34	4.19	1.22	27	3.00

Table 2. Estimates of the marginal quasi-stationary distribution ($q_{.,1}$), time to extinction (τ) and threshold of endemicity ($\hat{z}$) for the languages surveyed for the first (1st) and second (2nd) generations respectively

Language	$q_{.,1}$		τ_n		$\tau(0)$		τ		$(\hat{z})$	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
Igbo	0.1286	0.2558	32.0	5.06	165.82	26.25	132.84	66.8	0.47	0.07
Yoruba	0.1275	0.2213	33.1	11.13	171.67	57.74	137.65	79.3	0.46	0.17
Urhobo	0.1245	0.3280	32.61	-	169.15	-	137.26	52.1	0.39	-
Esan	0.1104	0.1930	39.44	15.31	204.61	79.42	163.02	93.2	0.47	0.20
Bini	0.1022	0.3053	43.33	0	224.78	0	180.34	60.4	0.53	0
Isoko	0.1040	-	38.59	-	200.20	-	159.84	-	0.49	-
Others	0.0941	0.2524	45.68	5.41	236.96	28.06	191.34	71.3	0.52	0.07

The mean family sizes n, ranged from 5 to 10. In order to ensure that all the time variables were in the same unit of time, the corresponding annual birth rate based on the generational birth rates were computed and used for the subsequent calculations. The mean infectious period $1/\nu$ was taken as the difference between the mean lifetime and the mean age at marriage. It varied across the languages surveyed.

Since the number of households in Enumeration Areas (EAs), the basic sampling unit in Nigerian population censuses, ranged between 80 and 100, we chose a value of k=100, which is a reasonable value for the number of households in a typical community.

For the choice of minimal interaction between families, we chose a value of $\varepsilon = 0.001$, which is smaller than Lindholm and Britton (2007) choice of 1/365=0.0027 and yields $\Pr(within\ contact) = 0.91$, in comparison with Lindholm and Britton (2007) choice which yields $\Pr(within\ contact) = 0.79$.

The choice of the value of ε, the proportion of an individual's language interaction with other families was motivated by the desire to keep the probability of inter-family interaction below 0.1. That is, we desired that the probability of within family interaction should be around 0.9, or that 90% of the language interactions among children should be within their family, enforcing the assumption of intergenerational transfer of language ability from parent to children.

The basic reproduction number R_0 when $\varepsilon = 0.001$ was slightly higher than the corresponding value when $\varepsilon = 0$ for all the tribes. The values of R_0 were between 3.03 and 4.68 for the first generation data and between 0.73 and 1.80 for the second generation data.

The values of α, the ratio of life's length to the length of the infectious period, were between 2.93 and 3.25 among the tribes. We observed minimal variation in the values of α, pointing to the fact that similar conditions affect individuals in the population irrespective of their language.

The marginal distribution of 1 infectious individual in the family at quasi-stationarity, $q_{.,1}$, had 0.13 as its largest values in the first generation (corresponding to the Igbo language) and 0.33 for the Urhobo language in the second generation.

Due to the generational decline in the values of R_0, the expected time to extinction for the families, τ_n also exhibited declines between the two generations. Similarly, the same scenario is replicated for $\tau(0)$ and τ.

The threshold of endemicity $\hat{z}$ also showed sometimes sharp decline between both generations across the surveyed languages. However, as should be expected for languages on the lower fringes of the extinction scale, the declines were steeper. The Urhobo and Bini languages had the lowest threshold of endemicity.Higher values of $\hat{z}$ imply that the process will remain in endemicity for a longer time before it is absorbed, while lower values imply that the process will only spend little time until absorption.

Once the process gets to endemicity/ quasi-stationarity, then in finite time, in the absence of any external intervention, the process will be absorbed even with $R_0 > 1$.

For the language to remain virile, it should be far away from the quasi-stationary level and programmes and policies put in place to ensure steady growth of the language-speaking population.

The total fertility rate, which is the expected number of children an average Nigerian woman would have born at the end of the childbearing cycle (15–49 years) was about 5.5 (NPC, 2014). This value was not too far off from our survey data for all the tribes.

4. Conclusion

The SIR model of language extinction provides a useful approach to ascertaining the status of any language globally, with the use of historical or survey data. This conceptualization therefore provides a tool that can be deployed to document the status of the world's numerous indigenous languages.

On the basis of the expected time to extinction, conditioned on quasi-stationarity, several of the surveyed Nigerian languages were seen to be in precarious conditions, while a few others are seemingly virile based on a high language transfer quotient within families. While it is difficult to correctly predict the time to extinction due to the limited nature of the survey, the rapid decline in indigenous language ability among younger Nigerians point to the possibility of extinction as the older generation exit the population.

The threshold of endemicity, which reflects the proportion of the language-speaking population in stationarity, also point to decline and threat of extinction in relation to the younger generation across the surveyed Nigerian languages.

The goal of indigenous language stakeholders is to enhance the level of use of the indigenous language in both private and public domains. In essence, language practitioners will desire the value of ε to be very close to 1. That will signal the virility of any indigenous language

In the presence of historical data, the model could provide a good perspective of the indigenous language dynamics vis a vis demographic, economic and social changes. In such situations, the progression of any indigenous language could be tracked effectively, so that when the proportion of speakers fall below the allowed threshold, necessary short-term and long-term interventions could be made by governments and other stakeholders.

Areas of possible extension of the model framework are given below:

i. The model could be modified to accommodate greater heterogeneities (age, sex, location, etc), as there could be certain segments of the population with varying indigenous language transmission rates and residual lifetimes;

ii. a simulation scheme could be conceptualized and developed to sufficiently capture the dynamics of any language decline, using initial parameter estimates obtained from survey or historical data;

iii. the value of the global infection rate can be modelled as a function of k, the number of families in the population, as the present model produces estimates of R_0 that tend to be far higher as k increases;

iv. better approximations for the expected time to extinction could be designed when
there are smaller family sizes; and

v. models that adequately accommodate the transient behaviour of language decline over times could be developed to study the indigenous languages.

REFERENCES

ALLEN, L., BURGIN, A. M., (2000). Comparison of Deterministic and Stochastic SIS and SIR Models in Discrete Time. Mathematical Biosciences, 163, pp. 1–33.

ANDERSSON, H., BRITTON, T., (2000). Stochastic Epidemic Models and Their Statistical Analysis, Lecture Notes, Department of Mathematics, Stockholm University, Sweden.

BALL, F. G., LYNE, O. D., (2000). Stochastic Multitype SIR Epidemics among a Population Partitioned into Households, Adv. in App. Prob., 33, pp. 99–123.

BILLINGS, L., MIER-Y-TERAN-ROMERO, L., LINDLEY, B., SCHWARTZ, I., (2013). Intervention-Based Stochastic Disease Eradication, PLoS ONE 8(8): e70211.

BURR, T. L., CHOWELL, G., (2008). Signatures of Non-Homogeneous Mixing in Disease Outbreaks. Mathematical and Computer Modelling, 48, pp. 122–140.

CRYSTAL, D., (2000). Language Death, New York: Cambridge University Press.

DALEY, D. J., KENDALL, D. G., (1964). Epidemics and Rumours. Nature, 204 (4963), 1118.

DOBSON, A. P., CARPER, E. R., (1996). Infectious Diseases and Human Population History, BioScience, 46(2), pp. 115–126.

HAGENAARS, T. J., DONNELY, C. A., FERGUSON, N. M., (2004). Spatial Heterogeneity and the Persistence of Infectious Diseases. J. Theo. Bio., 203, pp. 33–50.

LEWIS, M. PAUL, GARY F., SIMONS, CHARLES, D. FENNIG, (eds.)., (2015). Ethnologue:
Languages of the World, Eighteenth edition. Dallas, Texas: SIL International. www.ethnologue.com.

LINDHOLM, M., (2007). Stochastic epidemic models for endemic diseases: the effect of population heterogeneities. Research Report 2007:10, Licentiate thesis, Department of Mathematics, Stockholm University, Sweden.

LINDHOLM, M., BRITTON, T., (2007). Endemic persistence or disease extinction: the effect of separation into families, Theo. Pop. Bio., 72, pp. 253–263.

NASELL, I., (1999). On the time to extinction in recurrent epidemics. Journal of the Royal Statistical Society, B 61, pp. 309–330.

NASELL, I., (2002). Stochastic Models of Some Endemic Infections. Mathematical Biosciences, 179, pp. 1–19

NASELL, I., (2005). A new look at the critical community size for childhood infections. Theoretical Population Biology, 67, pp. 203–216.

National Population Commission, (2014). Nigeria Demographic and Health Survey 2013.

Published by National Population Commission (Nigeria) and ICF International, www.npc.gov.ng.

NUWER, R., (2014, May 6). Why We Must Save Dying Languages. Retrieved from www.bbc.co.uk.

OHIRI-ANICHE, C., (2014). More than 400 Nigerian Indigenous Languages Endangered,
Vanguard Newspaper, Nigeria, www.vanguardngr.com, 26-04-14.

SCHWARTZ, I. B., BILLINGS, L., DYKMAN, M., LANDSMAN, A., (2009). Predicting Extinction Rates in Stochastic Epidemic Models. Journal of Statistical Mechanics:
Theory and Experiment, www.stacks.iop.org/JSTAT/2009/P01005 .

VERDASCA, J. A., TELO DA GAMA, M. M., NUNES, A., BERNADINO, N. R., PACHECO, J. M., GOMES, M. C., (2005). Recurrent Epidemics in Small World Networks, J. Theor. Biol., 233(4), pp. 553–561.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 89–111, DOI 10.21307/stattrans-2019-036
Submitted – 28.08.2018; Paper ready for publication – 01.10.2019

ESTIMATING POPULATION COEFFICIENT OF VARIATION USING A SINGLE AUXILIARY VARIABLE IN SIMPLE RANDOM SAMPLING

Rajesh Singh[1], **Madhulika Mishra**[2]

ABSTRACT

This paper proposes an improved estimation method for the population coefficient of variation, which uses information on a single auxiliary variable. The authors derived the expressions for the mean squared error of the proposed estimators up to the first order of approximation. It was demonstrated that the estimators proposed by the authors are more efficient than the existing ones. The results of the study were validated by both empirical and simulation studies.

Key words: coefficient of variation, simple random sampling, auxiliary variable, mean square error.

1. Introduction

It is a prominent fact in the theory of sample surveys that suitable use of auxiliary information increases the efficiency of the estimators used for estimating the unknown population parameters. Some important works illustrating use of auxiliary information at estimation stage are Singh et al. (2005), Singh et al. (2007), Khoshnevisan et al. (2007), Singh et al. (2009), Singh and Kumar (2011), Malik and Singh (2013) and Singh et al. (2018). Over a vast period of time a substantial amount of work has been done by several authors for the estimation of population mean, population variance but little attention has been given to the estimation of the population coefficient of variation. Das and Tripathi (1992–93) first proposed the estimator for the coefficient of variation when samples were selected using simple random sampling without replacement (SRSWOR) scheme. Other works include Patel and Shah (2009) and Ahmed, S.E. (2002). Breunig (2001) suggested an almost unbiased estimator of the coefficient of variation. Sisodia and Dwivedi (1981) suggested a modified ratio estimator using the coefficient of variation of auxiliary variable.

[1] Department of Statistics, Banaras Hindu University, Varanasi-221005, India. ORCID ID: http://orcid.org/0000-0002-9274-8141.

[2] Corresponding Author: Department of Statistics, Banaras Hindu University, Varanasi-221005, India. E-mail: madhulika1707@gmail.com. ORCID ID: http://orcid.org/0000-0002-6408-0746.

Rajyaguru and Gupta (2005) also worked on the problem of estimation of the coefficient of variation under simple random sampling and stratified random sampling.

The coefficient of variation is extensively used in biology, agriculture and environmental sciences.

A brief summary of the paper is as follows.

Section 1 is introductory in nature, comprises the works that have been already done in the sampling literature. In Section 2 we considered five estimators for comparison purposes and their properties. In Section 3, we proposed two log type estimators for the coefficient of variation, one general type estimator and one wider type. In Section 4, an empirical study was carried out in support of our results. In Section 5, we carried out a simulation study to validate our theoretical results and have presented them with the help of bar graphs. In Section 6 we finally concluded our results.

Let us consider a finite population P = (P_1, P_2……… P_N) of size 'N' consisting of distinct and identifiable units. Let the study and auxiliary variables be denoted by Y and X, and let Y_i and X_i be their values corresponding to ith unit in the population (i = 1, 2……… N). We define:

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ as the population mean for the study variable

$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ as the population mean for the auxiliary variable

$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$ as the population mean square for the study variable

$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$ as the population mean square for the auxiliary variable

$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)$as the population covariance between the study and auxiliary variable, X and Y.

Let us suppose that a sample of size 'n' has been drawn from this population of size 'N' units using SRSWOR technique. For this sample let y_i and x_i denote values of the i^{th} sample unit corresponding to study variable Y and auxiliary variable X respectively.

For the sample observations, we define:

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ as the sample mean for the study variable Y

$\bar{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$ are the sample mean for the auxiliary variable X

$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ as the sample mean square for the study variable

$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ as the sample mean square for the auxiliary variable

$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)$as the sample covariance term.

Now, let us define

$$\in_0 = \frac{\bar{y}}{\bar{Y}} - 1, \in_1 = \frac{\bar{x}}{\bar{X}} - 1, \in_2 = \frac{s_y^2}{S_y^2} - 1 \text{ and } \in_3 = \frac{s_x^2}{S_x^2} - 1$$

such that

$$\mathrm{E}(\in_0) = \mathrm{E}(\in_1) = \mathrm{E}(\in_2) = \mathrm{E}(\in_3) = 0$$

$$\mathrm{E}(\in_0^2) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)C_y^2 \ , \ \mathrm{E}(\in_1^2) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)C_x^2 , \ \mathrm{E}(\in_2^2) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)(\lambda_{40} - 1) \ ,$$

$$\mathrm{E}(\in_3^2) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)(\lambda_{04} - 1)$$

$$\mathrm{E}(\in_0 \in_1) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)\rho C_y C_x \ , \ \mathrm{E}(\in_0 \in_2) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)C_y \lambda_{30} \ \mathrm{E}(\in_0 \in_3) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)C_y \lambda_{12} ,$$

$$\mathrm{E}(\in_1 \in_2) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)C_x \lambda_{21} \ , \ \mathrm{E}(\in_1 \in_3) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)C_x \lambda_{03} \ ,$$

$$\mathrm{E}(\in_2 \in_3) = \left(\frac{1-\mathrm{f}}{\mathrm{n}}\right)(\lambda_{22} - 1)$$

Here, $f = \frac{\mathrm{n}}{\mathrm{N}}$: Sampling fraction, $C_y = \frac{\mathrm{S_y}}{\bar{\mathrm{Y}}}$ and $C_x = \frac{\mathrm{S_x}}{\bar{\mathrm{X}}}$ are the population coefficient of variation for the study variable Y and auxiliary variable X, respectively. Also ρ_{xy} denotes the correlation coefficient between X and Y.

In general,

$$\mu_{rs} = \frac{1}{\mathrm{N-1}}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^r\left(x_i - \bar{X}\right)^s \text{ and } \lambda_{rs} = \frac{\mu_{\mathrm{rs}}}{\mu_{20}^{\mathrm{r}/2}\,\mu_{02}^{\mathrm{s}/2}} \text{ respectively.}$$

2. Existing estimators

- The usual unbiased estimator to estimate the population coefficient of variation using information on a single auxiliary variable is defined below:

$$t_0 = \hat{C}_y = \frac{\mathrm{s_y}}{\bar{\mathrm{y}}} = \frac{\mathrm{S_y}\left(1+\in_2\right)^{1/2}}{\bar{Y}\left(1+\in_0\right)}$$

$$\approx\left(1\text{-}\in_0 + \in_0^2 + \frac{\in_2}{2} - \frac{\in_0\in_2}{2} - \frac{\in_2^2}{8}\right)C_y \quad (2.1)$$

Its mean squared error (MSE) is given by:

$$MSE\left(t_0\right) \approx \mathrm{C_y^2}\left(\frac{1-f}{n}\right)\left[C_y^2 + \frac{\lambda_{40}-1}{4} - C_y\lambda_{30}\right] \quad (2.2)$$

- Solanki et al. (2015) introduced a difference type estimator for the population coefficient of variation C_y as:

$$C_d = \hat{\mathrm{C}}_{\mathrm{y}} + \alpha_2\left(C_x - \hat{C}_x\right) \quad (2.3)$$

MSE of C_d is given by:

$$MSE\left(C_d\right) \approx \mathrm{C_y^2}\left(\frac{1-f}{n}\right)\left[\left\{C_y^2 + \frac{\lambda_{40}-1}{4} - C_y\lambda_{30}\right\} - C_y^2\,\frac{\left\{\rho C_y C_x - \frac{C_y\lambda_{12}}{2} - \frac{C_x\lambda_{21}}{2} + \frac{\lambda_{22}-1}{4}\right\}}{\left\{C_x^2 + \frac{\lambda_{04}-1}{4} - C_x\lambda_{03}\right\}}\right]$$

(2.4)

- Solanki et al. (2015) defined another class of estimator for the population coefficient of variation C_y as:

$$C_d^* = \alpha_1 \hat{C}_y + \alpha_2 \left(\hat{C}_x - C_x \right) \tag{2.5}$$

MSE of C_d^* is given by:

$$MSE\left(C_d^*\right) \approx \alpha_1^2 C_y^2 A + \alpha_2^2 C_x^2 B + C_y^2 + 2\alpha_1 \alpha_2 C_y C_x C - 2\alpha_1 C_y^2 D - 2\alpha_2 C_y C_x E \tag{2.6}$$

Here,

$$\left.\begin{aligned}
A &= 1 + \left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left\{3C_y^2 - 2C_y \lambda_{30}\right\} \\
B &= \left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left\{C_x^2 + \frac{\lambda_{04} - 1}{4} - C_x \lambda_{03}\right\} \\
\mathrm{C} &= \left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left\{C_x^2 - \frac{\lambda_{04} - 1}{8} - \frac{C_x \lambda_{03}}{2} + \rho C_y C_x - \frac{C_y \lambda_{12}}{2} - \frac{C_x \lambda_{21}}{2} + \frac{\lambda_{22} - 1}{4}\right\} \\
D &= 1 + \left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left\{C_y^2 - \frac{\lambda_{40} - 1}{8} - \frac{C_y \lambda_{30}}{2}\right\} \\
\mathrm{E} &= \left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left\{\mathrm{C}_{\mathrm{x}}^2 - \frac{\lambda_{04} - 1}{8} - \frac{\mathrm{C}_{\mathrm{x}} \lambda_{03}}{2}\right\}.
\end{aligned}\right\} \tag{2.7}$$

On differentiating equation (2.6) with respect to α_1 and α_2, we obtain their optimum values as:

$$\alpha_{1opt} = \frac{BD - CE}{AB - C^2} \tag{2.8}$$

$$\alpha_{2opt} = \frac{\mathrm{C}_{\mathrm{y}}}{\mathrm{C}_{\mathrm{x}}}\left(\frac{AE - CD}{AB - C^2}\right) \tag{2.9}$$

On substituting these optimum values of α_1 and α_2, in equation (2.6), we obtain the Minimum MSE for the estimator C_d^* as:

$$\min MSE\left(C_d^*\right) \approx \alpha_{1opt}^2 C_y^2 A + \alpha_{2opt}^2 C_x^2 B + C_y^2 + 2\alpha_{1opt} \alpha_{2opt} C_y C_x C - 2\alpha_{1opt} C_y^2 D - 2\alpha_{2opt} C_y C_x E \tag{2.10}$$

- Adichwal et al. (2016) proposed a two-parameter ratio-product-ratio estimator for the population coefficient of variation as:

$$t_{r1} = \alpha\left[\frac{(1\text{-}\beta)\bar{x}+\beta\bar{X}}{\beta\bar{x}+(1-\beta)\bar{X}}\right]\hat{C}_y + (1-\alpha)\left[\frac{\beta\bar{x}+(1-\beta)\bar{X}}{(1-\beta)\bar{x}+\beta\bar{X}}\right]\hat{C}_y \tag{2.11}$$

$$t_{r2} = \gamma\left[\frac{(1\text{-}\delta)s_x^2+\delta S_x^2}{\delta s_x^2+(1-\delta)S_x^2}\right]\hat{C}_y + (1-\gamma)\left[\frac{\delta\, s_x^2+(1-\delta)S_x^2}{(1-\delta)s_x^2+\delta S_x^2}\right]\hat{C}_y \tag{2.12}$$

MSE of the estimators t_{r1} and t_{r2} are respectively given by:

$$MSE(t_{r1}) \approx \mathrm{MSE}(\mathrm{C_y}) - \frac{1}{4}\left(\frac{1-f}{n}\right)\{2\rho_{xy}C_y - \lambda_{21}\}^2 C_y^2 \tag{2.13}$$

$$MSE(t_{r2}) \approx \mathrm{MSE}(\mathrm{C_y}) - \frac{1}{4}\left(\frac{1-f}{n}\right)\frac{\{(\lambda_{22}-1)-2\lambda_{12}C_y\}^2}{(\lambda_{04}-1)} C_y^2 \tag{2.14}$$

3. Proposed estimators

We have proposed some estimators for the coefficient of variation based on information on a single auxiliary variable.

Motivated by Mishra and Singh (2017), we propose improved log type estimators for estimating the population coefficient of variation given by:

estimators t1 and t2 as:

a) $$t_1 = \hat{\mathrm{C}}_{\mathrm{y}} + \alpha\log\left(\frac{\hat{C}_x}{C_x}\right) \tag{3.1}$$

b) $$t_2 = \hat{\mathrm{C}}_{\mathrm{y}}(1+w_1) + w_2\log\left(\frac{\hat{C}_x}{C_x}\right) \tag{3.2}$$

Expressing the estimator t_1 and in terms of $\in$'s and then taking expectations up to the first order of approximation, we get MSE of the estimator t_1 as:

$$MSE(t_1) \approx \mathrm{C_y^2}\left(\frac{1-f}{n}\right)\left[\mathrm{C_y^2} + \frac{(\lambda_{40}-1)}{4} - C_y\lambda_{30}\right] + \alpha^2\left(\frac{1-f}{n}\right)\left[\mathrm{C_x^2} + \frac{(\lambda_{04}-1)}{4} - C_x\lambda_{03}\right] + 2C_y\alpha\left(\frac{1-f}{n}\right)\left[\rho C_y C_x + \frac{(\lambda_{22}-1)}{4} - \frac{(C_y\lambda_{12})}{2} - \frac{(C_x\lambda_{21})}{2}\right] \tag{3.3}$$

$$MSE(t_1) = \mathrm{C}_{\mathrm{y}}^2 A_1 + \alpha^2 A_2 + 2C_y \alpha A_3 \tag{3.4}$$

Here,

$$\left.\begin{aligned}
\mathrm{A}_1 &= \left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left\{\mathrm{C}_{\mathrm{y}}^2 + \frac{(\lambda_{40}-1)}{4} - \mathrm{C}_{\mathrm{y}}\lambda_{30}\right\},\\
\mathrm{A}_2 &= \left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left\{\mathrm{C}_{\mathrm{x}}^2 + \frac{(\lambda_{04}-1)}{4} - \mathrm{C}_{\mathrm{x}}\lambda_{03}\right\},\\
\mathrm{A}_3 &= \left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left\{\rho\mathrm{C}_{\mathrm{y}}\mathrm{C}_{\mathrm{x}} + \frac{(\lambda_{22}-1)}{4} - \frac{\mathrm{C}_{\mathrm{x}}\lambda_{21}}{2} - \frac{\mathrm{C}_{\mathrm{y}}\lambda_{12}}{2}\right\}.
\end{aligned}\right\} \tag{3.5}$$

To obtain the optimum value of α , we partially differentiate the expression (3.4) with respect to α and we obtain the optimum value as:

$$\alpha_{opt} = \text{-}\,\mathrm{C}_{\mathrm{y}}\,\frac{A_3}{A_2} \tag{3.6}$$

Putting this optimum value of α in equation (3.4), we get the minimum value for $MSE(t_1)$ as:

$$\min MSE(t_1) \approx \mathrm{C}_{\mathrm{y}}^2\left(A_1 - \frac{A_3^2}{A_2}\right) \tag{3.7}$$

Expressing the estimators t_2 in terms of $\in$'s and then taking expectations up to the first order of approximation we get MSE of the estimator t_2 as:

$$MSE(t_2) \approx \mathrm{C}_{\mathrm{y}}^2\left(\frac{1-f}{n}\right)\left[C_y^2 + \frac{\lambda_{40}-1}{4} - C_y\lambda_{30}\right] + C_y^2 w_1^2\left[1 + 3\left(\frac{1-f}{n}\right)C_y^2 - 2\left(\frac{1-f}{n}\right)C_y\lambda_{30}\right] +$$

$$w_2^2\left(\frac{1-f}{n}\right)\left[C_x^2 + \frac{\lambda_{04}-1}{4} - C_x\lambda_{03}\right] + 2C_y^2 w_1\left(\frac{1-f}{n}\right)\left[2C_y^2 + \frac{\lambda_{40}-1}{8} - \frac{3}{2}C_y\lambda_{30}\right] +$$

$$2C_y w_1 w_2\left(\frac{1-f}{n}\right)\left[\frac{C_x^2}{2} + \rho C_y C_x + \frac{\lambda_{22}-1}{4} - \frac{\lambda_{04}-1}{4} - \frac{C_y\lambda_{12}}{2} - \frac{C_x\lambda_{21}}{2}\right] +$$

$$2C_y w_2\left(\frac{1-f}{n}\right)\left[\rho C_y C_x + \frac{\lambda_{22}-1}{4} - \frac{C_y\lambda_{12}}{2} - \frac{C_x\lambda_{21}}{2}\right] \tag{3.8}$$

$$MSE(t_2) = \mathrm{C}_{\mathrm{y}}^2 B_1 + C_y^2 w_1^2 B_2 + w_2^2 B_3 + 2C_y^2 w_1 B_4 + 2C_y w_1 w_2 B_5 + 2C_y w_2 B_6 \tag{3.9}$$

Here,

$$\left.\begin{aligned}
B_1 &= \left(\frac{1-f}{n}\right)\left[C_y^2 + \frac{\lambda_{40}-1}{4} - C_y\lambda_{30}\right] \\
B_2 &= 1 + 3\left(\frac{1-f}{n}\right)C_y^2 - 2\left(\frac{1-f}{n}\right)C_y\lambda_{30} \\
B_3 &= \left(\frac{1-f}{n}\right)\left[C_x^2 + \frac{\lambda_{04}-1}{4} - C_x\lambda_{03}\right] \\
B_4 &= \left(\frac{1-f}{n}\right)\left[2C_y^2 + \frac{\lambda_{40}-1}{8} - \frac{3}{2}C_y\lambda_{30}\right] \\
B_5 &= \left(\frac{1-f}{n}\right)\left[\frac{C_x^2}{2} + \rho C_y C_x + \frac{\lambda_{22}-1}{4} - \frac{\lambda_{04}-1}{4} - \frac{C_y\lambda_{12}}{2} - \frac{C_x\lambda_{21}}{2}\right] \\
B_6 &= \left(\frac{1-f}{n}\right)\left[\rho C_y C_x + \frac{\lambda_{22}-1}{4} - \frac{C_y\lambda_{12}}{2} - \frac{C_x\lambda_{21}}{2}\right]
\end{aligned}\right\} \tag{3.10}$$

To obtain the optimum value of w_1 and w_2, we differentiate the expression (2.21) with respect to w_1 and w_2 and obtain the optimum values as:

$$w_{1opt} = \left(\frac{B_5 B_6 - B_3 B_4}{B_2 B_3 - B_5^2}\right) \tag{3.11}$$

$$w_{2opt} = C_y\left(\frac{B_6 B_2 - B_4 B_5}{B_5^2 - B_2 B_3}\right) \tag{3.12}$$

Putting these optimum values of w_1 and w_2 in equation (2.21), we get the minimum value for $MSE(t_2)$ as:

$$MSE(t_2)=\mathrm{C}_{\mathrm{y}}^2 B_1+C_y^2 w_{1opt}^2 B_2+w_{2opt}^2 B_3+2C_y^2 w_{1opt} B_4+2C_y w_{1opt} w_{2opt} B_5+2C_y w_{2opt} B_6 \quad (3.13)$$

c) Following Srivastava and Jhajj (1981), we propose a general class of estimators to estimate the population coefficient of variation C_y of the study variable Y using known mean and known variance of auxiliary variable X as:

$$t_3=\hat{C}_y H(u,v) \quad (3.14)$$

where $u=\frac{\bar{x}}{\bar{X}}$, $v=\frac{\mathrm{s}_{\mathrm{x}}^2}{\mathrm{S}_{\mathrm{x}}^2}$ and $H(u,v)$ is a function of u and v such that the point (u,v) assumes the value in a closed convex subset R_2 of two-dimensional real space containing the point $(1,1)$;

The function $H(u,v)$ is continuous and bounded in R_2;
$H(1,1)=1;$

The first and the second order partial derivatives of $H(u,v)$ exist and are continuous and bounded in R_2.

Expanding $H(u,v)$ about the point $(1,1)$ in a second order Taylor's series we obtain

$$t_3=\hat{\mathrm{C}}_{\mathrm{y}} H(u,v)=\hat{\mathrm{C}}_{\mathrm{y}} H[1+(u-1),1+(v-1)] \quad (3.15)$$

$$t_3=\hat{C}_y\left[H(1,1)+(u-1)\left.\frac{\partial H}{\partial u}\right|_{(1,1)}+(v-1)\left.\frac{\partial H}{\partial v}\right|_{(1,1)}+(u-1)^2\frac{1}{2}\left.\frac{\partial^2 H}{\partial u^2}\right|_{(1,1)}+(v-1)^2\left.\frac{\partial^2 H}{\partial v^2}\right|_{(1,1)}\right.$$
$$\left.+(u-1)(v-1)\frac{1}{2}\left.\frac{\partial^2 H}{\partial u\partial v}\right|_{(1,1)}\right] \quad (3.16)$$

$$t_3=\hat{\mathrm{C}}_{\mathrm{y}}\left[1+\in_1 H_1+\in_3 H_2+\in_1^2 H_3+\in_3^2 H_4+\in_1\in_3 H_5\right] \quad (3.17)$$

Here,

$$H_1 = \left.\frac{\partial H}{\partial u}\right|_{(1,1)}, \quad H_2 = \left.\frac{\partial H}{\partial v}\right|_{(1,1)}, \quad H_3 = \frac{1}{2}\left.\frac{\partial^2 H}{\partial u^2}\right|_{(1,1)}, \quad H_4 = \left.\frac{\partial^2 H}{\partial v^2}\right|_{(1,1)}, \quad H_5 = \frac{1}{2}\left.\frac{\partial^2 H}{\partial u \partial v}\right|_{(1,1)}$$

Substituting the value of $\hat{C}_y$ in the above expression (2.28), we get

$$t_3 = C_y\left(1 - \in_0 + \in_0^2 + \frac{\in_2}{2} - \frac{\in_0 \in_2}{2} - \frac{\in_2^2}{8}\right)\left(1 + \in_1 H_1 + \in_3 H_2 + \in_1^2 H_3 + \in_3^2 H_4 + \in_1 \in_3 H_5\right)$$

(3.18)

Mean square error of the estimator t3 is given by

$$MSE(t_3) = \mathrm{E}\left[\mathrm{t}_3 - C_y\right]^2 = \mathrm{C}_\mathrm{y}^2\mathrm{E}\left[-\in_0 + \in_1 H_1 + \frac{\in_2}{2} + \in_3 H_2 + \mathrm{O}(\in)\right]^2$$

(3.19)

Simplifying the expression (2.30), we get

$$MSE(t_3) = \mathrm{C}_\mathrm{y}^2\left(1 - \frac{f}{n}\right)\left[C_y^2 + C_x^2 H_1^2 - 2H_1\rho C_y C_x + \frac{(\lambda_{40} - 1)}{4} + (\lambda_{04} - 1)H_2^2 + (\lambda_{22} - 1)H_2 + \right.$$
$$\left. 2\left\{-\frac{C_y \lambda_{30}}{2} - C_y \lambda_{12} H_2 + \frac{C_x \lambda_{21}}{2} H_1 + C_x \lambda_{03} H_1 H_2\right\}\right]$$

(3.20)

In order to obtain the minimum MSE for the estimator t_3, we partially differentiate the expression (2.31) with respect to H_1 and H_2 to get the optimum values as

$$H_{1opt} = \frac{1}{2}\left[\frac{(\lambda_{04} - 1)\{2\rho C_y - \lambda_{21}\} - \lambda_{03}\{2C_y \lambda_{12} - (\lambda_{22} - 1)\}}{C_x\{(\lambda_{04} - 1) - \lambda_{03}^2\}}\right] \tag{3.21}$$

$$H_{2opt} = \frac{1}{2}\left[\frac{\lambda_{03}\{2\rho C_y - \lambda_{21}\} - \{2C_y \lambda_{12} - (\lambda_{22} - 1)\}}{\lambda_{03}^2 - (\lambda_{04} - 1)}\right] \tag{3.22}$$

Substituting these optimum values of H_1 and H_2 in equation (2.31), we obtain the expression for the minimum MSE of t_3

$$MSE(t_3)=\mathrm{C}_{\mathrm{y}}^2\left(1-\frac{f}{n}\right)\left[C_y^2+C_x^2H_{1opt}^2-2H_{1opt}\rho C_yC_x+\frac{(\lambda_{40}-1)}{4}+(\lambda_{04}-1)H_{2opt}^2+(\lambda_{22}-1)H_{2opt}+\right.$$
$$\left.2\left\{-\frac{C_y\lambda_{30}}{2}-C_y\lambda_{12}H_{2opt}+\frac{C_x\lambda_{21}}{2}H_{1opt}+C_x\lambda_{03}H_{1opt}H_{2opt}\right\}\right] \tag{3.23}$$

d) Again, following Srivastava and Jhajj (1981), we propose a wider class of estimators to estimate the population coefficient of variation C_y as:

$$t_4=H^*\left(\hat{C}_y,u,v\right) \tag{3.24}$$

where $u=\frac{\bar{x}}{\bar{X}}$, $v=\frac{\mathrm{s}_{\mathrm{x}}^2}{\mathrm{S}_{\mathrm{x}}^2}$ and $H^*\left(\hat{C}_y,u,v\right)$ is a function of $\hat{C}_y$, u and v such that the point $\left(\hat{C}_y,u,v\right)$assumes the value in a closed convex subset R_3 of three-dimensional real space containing the point $\left(C_y,1,1\right)$;

The function $H^*\left(\hat{C}_y,u,v\right)$ is continuous and bounded in R_3;

$H^*\left(C_y,1,1\right)=C_y$;

The first and the second order partial derivatives of $H^*\left(\hat{C}_y,u,v\right)$ exist and are continuous and bounded in R_3.

Expanding $H^*\left(\hat{C}_y,u,v\right)$ about the point $\left(C_y,1,1\right)$ in a second order Taylor's series, we have

$$t_4=H^*\left(\hat{C}_y,u,v\right)=\mathrm{H}^*\left[\mathrm{C}_{\mathrm{y}}+\left(\hat{C}_y-C_y\right),1+(u-1),1+(v-1)\right]$$
$$=H^*\left(C_y,1,1\right)+\left(\hat{C}_y-C_y\right)\frac{\partial H^*}{\partial\hat{C}_y}\bigg|_{\left(C_{y,1,1}\right)}+(u-1)\frac{\partial H^*}{\partial u}\bigg|_{\left(C_{y,1,1}\right)}+(v-1)\frac{\partial H^*}{\partial v}\bigg|_{\left(C_{y,1,1}\right)}+$$
$$\left(\hat{C}_y-C_y\right)^2\frac{1}{2}\frac{\partial^2H^*}{\partial\hat{C}_y^2}\bigg|_{\left(C_{y,1,1}\right)}+(u-1)^2\frac{1}{2}\frac{\partial^2H^*}{\partial u^2}\bigg|_{\left(C_{y,1,1}\right)}+(v-1)^2\frac{1}{2}\frac{\partial^2H^*}{\partial v^2}\bigg|_{\left(C_{y,1,1}\right)}+$$

$$\left(\hat{C}_y - C_y\right)(u-1)\frac{1}{2}\frac{\partial^2 H^*}{\partial \hat{C}_y \partial u}\bigg|_{\left(C_y,1,1\right)} + (u-1)(v-1)\frac{1}{2}\frac{\partial^2 H^*}{\partial u \partial v}\bigg|_{\left(C_y,1,1\right)} + (v-1)\left(\hat{C}_y - C_y\right)\frac{1}{2}\frac{\partial^2 H^*}{\partial v \partial \hat{C}_y}\bigg|_{\left(C_y,1,1\right)} \tag{3.25}$$

$$t_4 = \mathrm{C_y} + \left(\hat{C}_y - C_y\right) + \in_1 H_1^* + \in_3 H_2^* + \left(\hat{C}_y - C_y\right)^2 H_3^* + \in_1^2 H_4^* + \in_3^2 H_5^* + \left(\hat{C}_y - C_y\right)\in_1 H_6^* + \\ \in_1 \in_3 H_7^* + \left(\hat{C}_y - C_y\right)\in_3 H_8^* \tag{3.26}$$

Here,

$$H^*\left(C_y,1,1\right) = \mathrm{C_y}\,,\ \frac{\partial H^*}{\partial \hat{C}_y}\bigg|_{\left(C_y,1,1\right)} = 1\,,\ H_1^* = \frac{\partial H^*}{\partial u}\bigg|_{\left(C_y,1,1\right)}\,,\ H_2^* = \frac{\partial H^*}{\partial v}\bigg|_{\left(C_y,1,1\right)}\,,$$

$$H_3^* = \frac{1}{2}\frac{\partial^2 H^*}{\partial \hat{C}_y^2}\bigg|_{\left(C_y,1,1\right)}$$

$$H_4^* = \frac{1}{2}\frac{\partial^2 H^*}{\partial u^2}\bigg|_{\left(C_y,1,1\right)}\,,\ H_5^* = \frac{1}{2}\frac{\partial^2 H^*}{\partial v^2}\bigg|_{\left(C_y,1,1\right)}\,,\ H_6^* = \frac{1}{2}\frac{\partial^2 H^*}{\partial \hat{C}_y \partial u}\bigg|_{\left(C_y,1,1\right)}\,,$$

$$H_7^* = \frac{1}{2}\frac{\partial^2 H^*}{\partial u \partial v}\bigg|_{\left(C_y,1,1\right)}\,,\ H_8^* = \frac{1}{2}\frac{\partial^2 H^*}{\partial v \partial \hat{C}_y}\bigg|_{\left(C_y,1,1\right)}$$

Now, substituting the value of $\hat{C}_y$ in equation (2.37), we have

$$t_4 = \mathrm{C_y}\left(1 - \in_0 + \in_0^2 + \frac{\in_2}{2} - \frac{\in_0 \in_2}{2} - \frac{\in_2^2}{8}\right) + \in_1 H_1^* + \in_3 H_2^* + \left\{C_y\left(1 - \in_0 + \in_0^2 + \frac{\in_2}{2} - \frac{\in_0 \in_2}{2} - \frac{\in_2^2}{8}\right) - C_y\right\}^2 H_3^*$$

$$\in_1^2 H_4^* + \in_3^2 H_5^* + \left\{C_y\left(1 - \in_0 + \in_0^2 + \frac{\in_2}{2} - \frac{\in_0 \in_2}{2} - \frac{\in_2^2}{8}\right) - C_y\right\}\in_1 H_6^* + \in_1 \in_3 H_7^* + \\ \left\{C_y\left(1 - \in_0 + \in_0^2 + \frac{\in_2}{2} - \frac{\in_0 \in_2}{2} - \frac{\in_2^2}{8}\right) - C_y\right\}\in_3 H_8^* \tag{3.27}$$

$$MSE\left(t_4\right) = \mathrm{E}\left[\mathrm{t}_4 - C_y\right]^2 = \mathrm{E}\left[C_y\left(-\in_0 + \frac{\in_2}{2}\right) + \in_1 H_1^* + \in_3 H_2^* + \mathrm{O}(\in)\right]^2 \tag{3.28}$$

After simplifying the expression (2.39), we get:

$$MSE(t_4)=\left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left[C_y^2\left\{C_y^2+\frac{(\lambda_{40}-1)}{4}-C_y\lambda_{30}\right\}+C_x^2H_1^{*2}+(\lambda_{04}-1)H_2^{*2}+2C_y\left(\frac{C_x\lambda_{21}}{2}-\rho C_yC_x\right)H_1^*+\right.$$
$$\left.2C_x\lambda_{03}H_1^*H_2^*+2C_y\left(\frac{(\lambda_{22}-1)}{2}-C_y\lambda_{12}\right)H_2^*\right] \qquad (3.29)$$

In order to obtain the minimum MSE for the estimator t_4 we partially differentiate the expression (2.40) with respect to H_1^* and H_2^* and obtain optimum values as:

$$\left.\begin{array}{l} H_{1opt}^*=\dfrac{\mathrm{C_y}}{2\mathrm{C_x}}\left[\dfrac{(\lambda_{04}-1)\{2\rho C_y-\lambda_{21}\}-\lambda_{03}\{2C_y\lambda_{12}-(\lambda_{22}-1)\}}{\lambda_{04}-\lambda_{03}^2-1}\right] \\[2ex] H_{2opt}^*=\dfrac{\mathrm{C_y}}{2}\left[\dfrac{\lambda_{03}(2\rho C_y-\lambda_{21})-\{2C_y\lambda_{12}-(\lambda_{22}-1)\}}{\lambda_{03}^2-(\lambda_{04}-1)}\right] \end{array}\right\} \qquad (3.30)$$

Substituting these optimum values of H_1^* and H_2^* in equation (2.40), we obtain the expression for the minimum MSE of t_4

$$MSE(t_4)=\left(\frac{1\text{-}\mathrm{f}}{\mathrm{n}}\right)\left[C_y^2\left\{C_y^2+\frac{(\lambda_{40}-1)}{4}-C_y\lambda_{30}\right\}+C_x^2{H_{1opt}^*}^2+(\lambda_{04}-1){H_{2opt}^*}^2+2C_y\left(\frac{C_x\lambda_{21}}{2}-\rho C_yC_x\right)H_{1opt}^*+\right.$$
$$\left.2C_x\lambda_{03}H_{1opt}^*H_{2opt}^*+2C_y\left(\frac{(\lambda_{22}-1)}{2}-C_y\lambda_{12}\right)H_{2opt}^*\right] \qquad (3.31)$$

4. Empirical study

In this section, we have carried out an empirical study to explicate the performance of our proposed estimator. We used the following data sets:

Population I: [Source : Murthy (1967), p. 399].

X: Area under wheat in 1963,

Y: Area under wheat in 1964,

N=34, n=15,

$\overline{X}$=208.88, $\overline{Y}$=199.44,

$C_X = 0.72$, C_Y=0.75, $\rho_{XY} = 0.98$,

$\lambda_{21} = 1.0045$, $\lambda_{12} = 0.9406$, $\lambda_{40} = 3.6161$, $\lambda_{04} = 2.8266$, $\lambda_{30} = 1.1128$, $\lambda_{03} = 0.9206$, $\lambda_{22} = 3.0133$

Population II: [Source : Sarjinder Singh (2003), p. 1116].

X: Number of fish caught in year 1993,

Y: Number of fish caught in year 1995,

N=69, n=40,

$\overline{X}$=4591.07, $\overline{Y}$=4514.89,

$C_X = 1.38$, C_Y=1.35,

$\lambda_{21} = 2.19$, $\lambda_{12} = 2.30$, $\lambda_{40} = 7.66$, $\lambda_{04} = 9.84$, $\lambda_{30} = 1.11$, $\lambda_{03} = 2.52$, $\lambda_{22} = 8.19$

In order to determine the Percent Relative Efficiency (PRE) of the estimators we have used the following formula

$$PRE(t, t_o) = \frac{Var(t_0)}{MSE(t)} \times 100$$

where $t = C_d, C_d^*, t_{r1}, t_{r2}, t_1, t_2, t_3 . t_4$.

Table 1. MSE and PRE of the estimators

ESTIMATOR	POPULATION-1		POPULATION-2	
	MSE	PRE	MSE	PRE
t_0	0.008016	100.00	0.0380	100.00
C_d	0.00123	651.7051	0.0298	127.4607
C_d^*	0.00122	654.4814	0.0285	133.2609
t_{r1}	0.006868	116.54	0.037313	102.04
t_{r2}	0.006963	114.95	0.037563	101.36

Table 1. MSE and PRE of the estimators (cont.(

ESTIMATOR	POPULATION-1		POPULATION-2	
	MSE	PRE	MSE	PRE
t_1	0.00123	651.7356	0.0299	127.4598
t_2	0.001038	771.9898	0.0283	134.6127
t_3	0.001203	666.4304	0.0297	128.345
t_4	0.001203	666.4304	0.0297	128.345

We can summarize the results from Table 1 as:

All the proposed estimators t_1, t_2, t_3 and t_4 are more efficient than the usual unbiased estimator t_0.The estimator t_1 turns out to be nearly as efficient as the difference type estimator C_d while all the remaining estimators, t_2, t_3 and t_4 are more efficient than the estimators C_d, C_d^*, t_{r1} and t_{r2}. Among all the estimators, t_2 is the most efficient because of the smallest value of MSE and highest value of PRE.

5. Simulation studies

This section describes the procedure that we adopted for the simulation study. We have used R programming for calculating MSE of the existing and proposed estimators. We followed the procedure adopted by Reddy et al. (2010) and have generated bivariate population with a specified correlation coefficient between the study and auxiliary variable. The algorithm is as follows:

1. Generate two independent random variables X from $N(\mu, \sigma^2)$ and Z from $N(\mu_1, \sigma_1^2)$ using Box-Muller method (Jhonson, 1987).

2. Set $Y = \rho X + \sqrt{1-\rho^2} Z$ where $0 < \rho = 0.75, 0.85, 0.95 < 1$.

3. Consider the population with the parameters $\mu = 2.5$, $\sigma^2 = 2$, $\mu_1 = 5$ $\sigma_1^2 = 3$ and repeat the steps 1-2 2000 times.

4. From the population of size N=2000, draw 1500 simple random samples (y_i, x_i) $(i = 1, 2, ..., n)$ without replacement of size $n = 30, 50, 70$.

5. For each of the sample, compute MSE of the estimators t_o , Cd , Cd^*, t_1, t_2 , t_{r1} and t_{r2}.

6. Compute the average MSE of the estimator by the following formula:

$$MSE(i)=\frac{1}{1500}\sum_{j=1}^{1500} mse_{\,j}(i) \quad \text{where } i=t_o\,,Cd\,,Cd^*,t_1\,,t_2\,,t_{r1}\; and\, t_{r2}\,.$$

Table 2. Table showing MSE and PRE of the existing and proposed estimators for different values of ρ and n

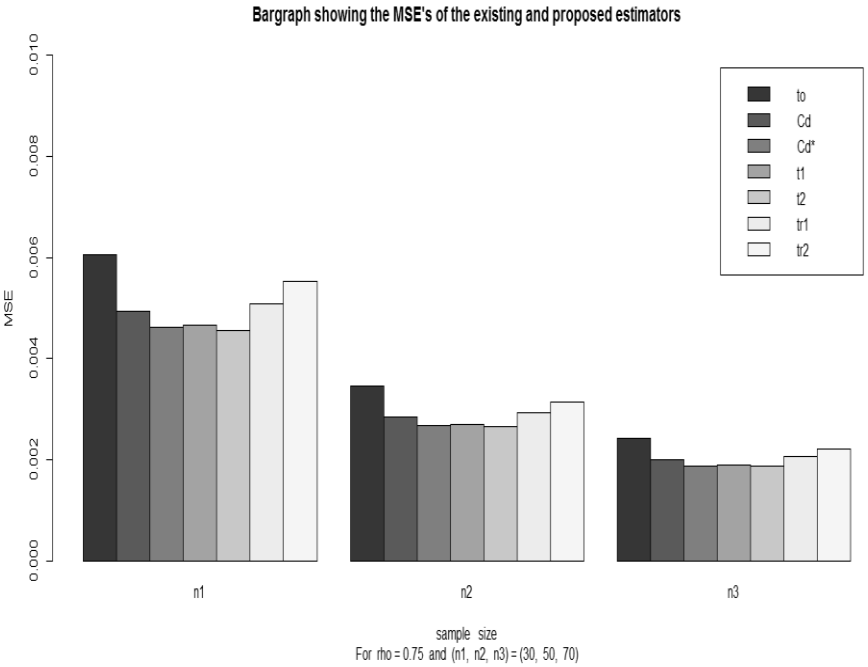
ρ	n	Estimator	MSE	PRE
		t_o	0.006053626	100.0000
		Cd	0.004924006	122.9410
		Cd^*	0.004617663	131.0970
		t_{r1}	0.005080027	119.1651
		t_{r2}	0.005532107	109.4270
		t_1	0.004668748	129.6626
		t_2	0.004552470	132.9744
	50	t_o	0.003450581	100.0000
		Cd	0.002835622	121.6869
		Cd^*	0.002671694	129.1533
		t_{r1}	0.002688403	117.5860
		t_{r2}	0.002650186	109.5933
		t_1	0.002934516	128.3506
		t_2	0.003148534	130.2015
	70	t_o	0.002412659	100.0000
		Cd	0.001990824	121.1889
		Cd^*	0.001879170	128.3896
		t_{r1}	0.002062564	116.9737
		t_{r2}	0.002200267	109.6530
		t_1	0.001887289	127.8373
		t_2	0.001868644	129.1128

0.85	30	t_o	0.006358341	100.0000
		Cd	0.004912595	129.4294
		Cd^*	0.003809327	166.9151
		t_{r1}	0.004876890	130.3769
		t_{r2}	0.005133219	123.8665
		t_1	0.003831045	165.9688
		t_2	0.003739557	170.0293
	50	t_o	0.003621428	100.0000
		Cd	0.002828737	128.0228
		Cd^*	0.002203557	164.3447
		t_{r1}	0.002825058	128.1895
		t_{r2}	0.002910006	124.4474
		t_1	0.002210627	163.8190
		t_2	0.002180249	166.1016
	70	t_o	0.002527309	100.0000
		Cd	0.001982556	127.4561
		Cd^*	0.001547597	163.3054
		t_{r1}	0.001984483	127.3535
		t_{r2}	0.002027634	124.6433
		t_1	0.001551035	162.9434
		t_2	0.001536162	164.5210

95	30	t_o	0.008647395	100.0000
		Cd	0.005851489	147.7811
		Cd^*	0.002426053	356.4389
		t_{r1}	0.005461214	158.3420
		t_{r2}	0.005113439	169.1094
		t_1	0.002430095	355.8459
		t_2	0.002364604	365.7015
	50	t_o	0.004896276	100.0000
		Cd	0.003355658	145.9111
		Cd^*	0.001397620	350.3295
		t_{r1}	0.003172583	154.3309
		t_{r2}	0.002841595	172.3073
		t_1	0.001398209	350.1820
		t_2	0.001376286	355.7601
	70	t_o	0.0034018746	100.0000
		Cd	0.0023435450	145.1593
		Cd^*	0.009782472	347.7520
		t_{r1}	0.0022234723	152.9983
		t_{r2}	0.0019631488	173.2866
		t_1	0.0009784789	347.6697
		t_2	0.0009677328	351.5304

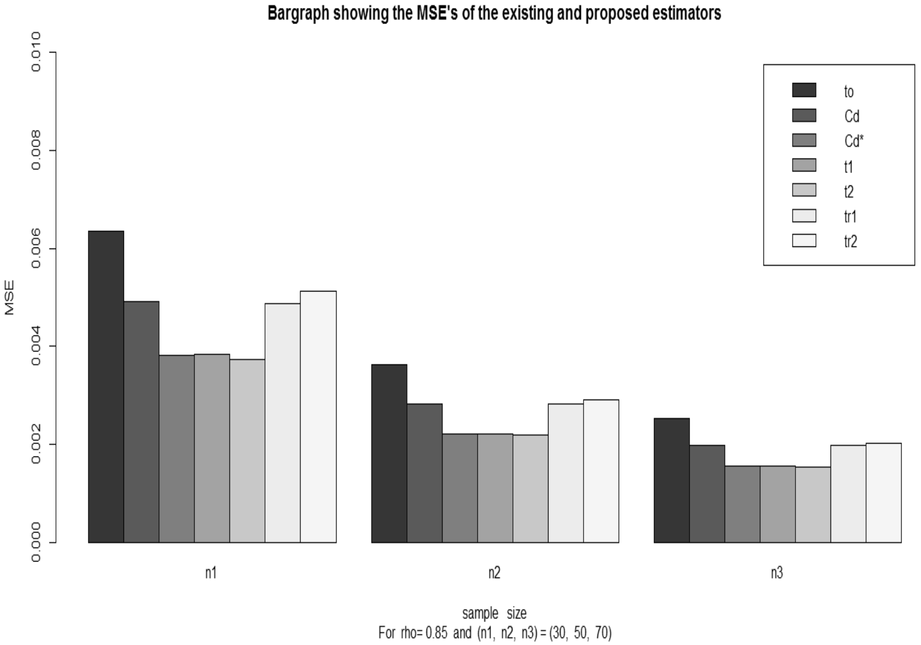
From the table, we can observe that for a particular value of ρ the value of MSE of the estimators decreases as the sample size increases. Also, we can see that in each of the cases among the proposed estimators t_1 and t_2 ,t_2 is more efficient amongst all the existing estimators t_o , Cd , Cd^*, t_{r1},t_{r2} and the proposed estimator t_1 while the estimator t_1 turns out to be more efficient than the existing estimators t_o , Cd , t_{r1}, t_{r2} and nearly as efficient as the estimator Cd^*. Hence, it turns out that the proposed estimator performs better than the existing estimators, therefore it is desirable to use the estimator in practice.

We have also shown the results through a bar diagram as below:



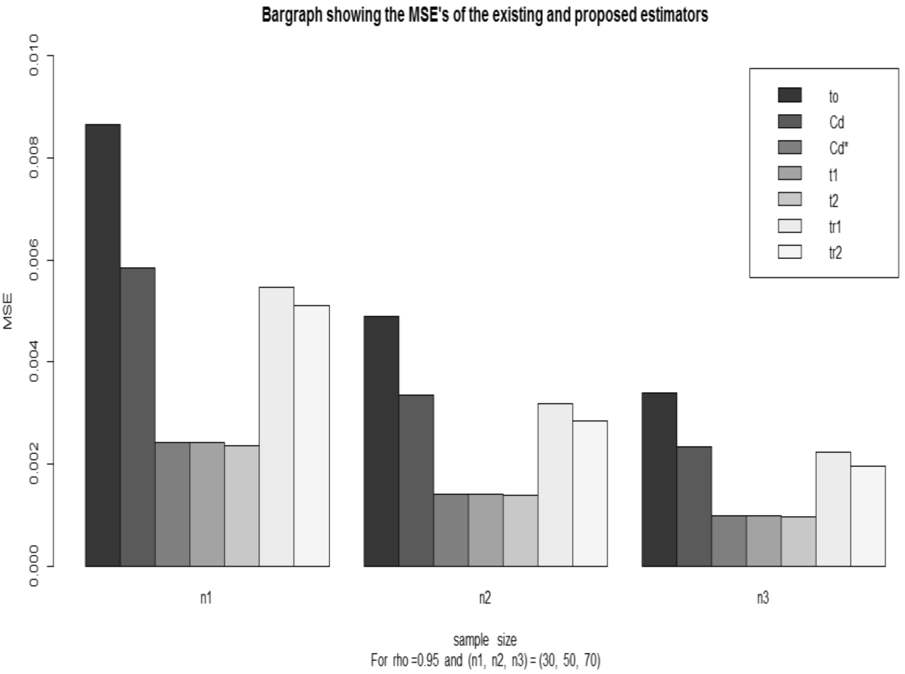
Bar graph showing MSEs of the existing and proposed estimators for $\rho = 0.75$ and (n1, n2, n3)= (30, 50, 70)

Explanation: It can be seen from the bar graph that for $\rho = 0.75$, MSE of all the estimators decreases as the value of the sample size (n) increases. And for a particular value of n, estimator t_2 has the least MSE among all the other estimators.



Bar graph showing MSEs of the existing and proposed estimators $\rho = 0.85$ and (n1, n2, n3) = (30, 50, 70)

Explanation: It can be seen from the bar graph that for $\rho = 0.85$, MSE of all the estimators decreases as the value of the sample size (n) increases. And for a particular value of n, estimator t_2 has the least MSE among all the other estimators.



Bar graph showing MSE of the existing and proposed estimators $\rho = 0.85$ and (n1, n2, n3)= (30, 50, 70)

Explanation: It can be seen from the bar graph that for $\rho = 0.95$, MSE of all the estimators decreases as the value of the sample size (n) increases. And for a particular value of n, estimator t_2 has the least MSE among all the other estimators.

Combined Explanation: From the above three bar graphs it can be summarized that for every value of $\rho = (0.75, 0.85, 0.95)$, the increase in the sample size causes a decrease in the mean square error of all the estimators. It is also evident that for a particular value of n, t_2 has the minimum MSE as compared to the other estimators.

6. Conclusion

In this paper we have proposed estimators for the population coefficient of variation and compared them with some existing estimators and saw from the empirical and simulation studies that the proposed estimator t_2 performs better than all the existing estimators t_o, Cd, C_d^*, t_{r1}, t_{r2} and the proposed estimator t_1. As regards t_1, it performs better than the estimators t_o, Cd, t_{r1}, t_{r2} but is no more better than the estimator C_d^*. For a better understanding of our results we have also considered a graphical approach and considered bar graphs to depict our results.

Acknowledgement

The authors are grateful and obliged to the Editor-in-Chief Prof. Włodzimierz Okrasa and the anonymous referees who devoted a part of their valuable time to provide us with their fruitful recommendations, which in turn helped us to improve the manuscript.

REFERENCES

AHMED, S. E., (2002). Simultaneous estimation of Co-efficient of Variation. Journal of Statistical Planning and Inference, 104, pp. 31–51.

ARCHANA, V., RAO, A., (2014). Some Improved Estimators of Co-efficient of Variation from Bi-variate normal distribution: A Monte Carlo Comparison. Pakistan Journal of Statistics and Operation Research, 10(1).

BREUNIG, R., (2001). An almost unbiased estimator of the co-efficient of variation. Economics Letters, 70(1), pp. 15–19.

DAS, A. K., TRIPATHI, T. P., (1981). A class of estimators for co-efficient of variation using knowledge on coefficient of variation of an auxiliary character, In annual conference of Ind. Soc. Agricultural Statistics, Held at New Delhi, India.

DAS, A. K., TRIPATHI, T. P., (1992). Use of auxiliary information in estimating the coefficient of variation, Alig. J. of. Statist, 12, pp. 51–58.

FINITE POPULATION-II, Model Assisted Statistics and application, 1(1), pp. 57–66.

KHOSHNEVISAN, M., SINGH, R., CHAUHAN, P., SAWAN, N., SMARANDACHE, F. (2007). A general family of estimators for estimating population means using known value of some population parameter(s), Far East Journal of Statistics, 22(2), pp. 181–191.

MALIK, S., SINGH, R., (2013). An improved estimator using two auxiliary attributes, Appli. Math. Compt., 219, pp. 10983–10986.

MISHRA, P., SINGH, R., (2017). A new log-product type estimator using auxiliary information, Jour. Sci. Res., 61(1&2), pp. 179–183.

MURTHY, M. N., (1967). Sampling theory and methods, Sampling theory and methods.

PATEL, P. A., RINA, S., (2009). A Monte Carlo comparison of some suggested estimators of co-efficient of variation in finite population, Journal of Statistics sciences, 1(2), pp. 137–147.

RAJYAGURU, A., GUPTA, P., (2005). On the estimation of the co-efficient of variation from

SINGH, H. P., TAILOR, R., (2005). Estimation of finite population mean with known coefficient of variation of an auxiliary character, Statistica, 65(3), pp. 301–313.

SINGH, H. P., SINGH, R., (2002). A class of chain ratio- type estimators for the coefficient of variation of finite population in two phase sampling, Aligarh Journal of Statistics, Vol. 22, pp. 1–9.

SINGH, S., (2003). Advanced Sampling Theory With Applications: How Michael Selected Amy (Vol. 2), Springer Science and Business Media.

SINGH, H. P., SINGH, R., ESPEJO, M. R., PINEDA, M. D., NADRAJAH, S., (2005). On the efficiency of a dual to ratio-cum-product estimator in sample surveys. Mathematical proceedings of the Royal Irish Academy, 105A (2), pp. 51–56.

SINGH, R., CHAUHAN, P., SAWAN, N., (2007). A family of estimators for estimating population means using known correlation coefficient in two-phase sampling. Statistics in Transition, 8(1), pp. 89–96.

SINGH, R., KUMAR, M., CHAUDHARY, M. K., KADILAR, C., (2009). Improved Exponential Estimator in Stratified Random Sampling, Pak. J. Stat. Oper. Res., 5(2), pp 67–82.

SINGH, R., KUMAR, M., (2011). A note on transformations on auxiliary variable in survey sampling. Mod. Assis. Stat. Appl., 6:1, pp. 17–19.

SINGH, R., MISHRA, P., BOUZA, C. N., (2018). Estimation of population mean using information on auxiliary attribute: A review, RG, DOI: 10.13140/RG.2.2.20477.87524.

SISODIA, B. V. S., DWIVEDI, V. K., (1981). Modified ratio estimator using coefficient of variation of auxiliary variable, Journal-Indian Society of Agricultural Statistics.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 113–134, DOI 10.21307/stattrans-2019-037
Submitted – 04.03.2019; Paper ready for publication – 25.10.2019

ESTIMATION OF INCOME CHARACTERISTICS FOR REGIONS IN POLAND USING SPATIO-TEMPORAL SMALL AREA MODELS

Alina Jędrzejczak[1,2], **Jan Kubacki**[2]

ABSTRACT

The paper presents the comparison of estimation results for spatial and spatiotemporal small area models. The study was carried out for income-related variables drawn from the Polish Household Budget Survey and explanatory variables from the Polish Local Data Bank for the years 2003-2013. The properties of EBLUPs (Empirical Best Linear Unbiased Predictors) based on spatiotemporal models, which utilize spatial correlation between neighbouring areas as well as historical data, were compared and contrasted with EBLUPs based on spatial models obtained separately for each year and with EBLUPs based on the Rao-Yu model. The computations were performed using sae, sae2 and spdep packages for R-project environment. In the case of sae package, the eblupFH, eblupSFH and the eblupSTFH functions were used for point estimation along with the mseFH, mseSFH and the pbmseSTFH functions for the MSE estimation, whereas the eblupRY function was applied for the purposes of sae2 package. The precision of direct estimators was guaranteed by the adoption of the Balanced Repeated Replication method. The results of the analysis demonstrate that a visible reduction of the estimation error was achieved for the implemented spatiotemporal small-area models, especially when significant spatial and time autocorrelations were observed. These results are even more valuable than those achieved by the means of the Rao-Yu model. In the computations, three author-defined functions were adopted, which not only enabled the author to perform the extract of random effects for spatial, spatiotemporal and Rao-Yu models, but also made it possible to obtain their decomposition with respect to spatial and temporal parts, thus creating a novel solution. The comparison was carried out using choropleth maps for spatial effects and distributions of temporal random effects for the considered years.

Key words: small area estimation, spatio-temporal models, Rao-Yu model, EBLUP estimation.

[1] Institute of Statistics and Demography, Faculty of Economics and Sociology, University of Łódź. E-mail: alina.jedrzejczak@uni.lodz.pl. ORCID ID: https://orcid.org/0000-0002-5478-9284.

[2] Centre of Mathematical Statistics, Statistical Office in Łódź. E-mail: j.kubacki@stat.gov.pl. ORCID ID: https://orcid.org/0000-0001-8281-0514.

1. Introduction

Statistical surveys are often designed to provide data that allow reliable estimation for the whole country and larger administrative units such as regions (in Poland – voivodships). However, for more specific variables the overall sample size is seldom large enough to yield direct estimates of adequate precision for all the domains of interest. In such cases, large estimation errors can make the inference unreliable and useless for decision-makers. The estimation errors can be reduced, however, by means of the model-based approach. Moreover, when an evident correlation exists between survey and administrative data, also the bias of the estimates can be reduced.

Small area estimation (SAE) offers a wide range of methods that can be applied when a sample size is insufficient to obtain high precision by means of conventional direct estimates. The techniques based on small area models - empirical best linear unbiased prediction (EBLUP) as well as empirical and hierarchical Bayes (EB and HB), seem to have a distinct advantage over other methods. One of these techniques is the spatio-temporal EBLUP technique, introduced by Marhuenda, Molina and Morales (2013). It is based on the assumption that the spatial relationships between domains can be modelled by a sum of two components: the simultaneous autoregressive process SAR (see: Pratesi and Salvati (2008), p. 114) and an additional time-related process described by AR(1) scheme (see: Rao, Yu (1992, 1994)). Both these assumptions are involved in the covariance structure of the spatio-temporal model. Related Spatial EBLUP (Spatial Empirical Best Linear Unbiased Prediction), which was introduced by Cressie (1991) and is explained in detail in Saei and Chambers (2003), Pratesi and Salvati (2004, 2005, 2008), Singh et al. (2005), Petrucci and Salvati (2006), can be considered as a reference point for spatio-temporal models. Recently, the Spatial EBLUP technique was used in ‘sae’ package (Molina, Marhuenda (2013)) for R-project environment published in CRAN resources. Moreover, some spatial econometric models were discussed in Griffith, D.A., Paelinck, J.H.P. (2011) and Kubacki, Jędrzejczak (2016), where MCMC (Markov Chain Monte Carlo) applications for spatial models are presented.

In the paper we compare several approaches to the spatial and spatio-temporal modelling implemented for small area estimation. In our opinion, spatio-temporal estimation can be sometimes useful with respect to the traditional EBLUP approach. This is because of better efficiency of such models, which not only incorporate ordinary spatial relationships (using proximity matrix), but also assume time-related dependencies. It can be useful when visible time-related relationships are observed in a data set. These models, using sample and auxiliary information from other domains as well as other time periods, can yield substantial quality improvements as compared to ordinary small area models, where only explanatory variables from administrative sources or other statistical surveys are used. It is also related to imposing certain constraints that can positively affect the quality of obtained estimates. The models with time-related covariance structure can additionally be helpful in the analysis of the dynamics of the observed phenomena, which can be supplementary related to the econometric models, including the panel models (Jędrzejczak, Kubacki (2016)).

2. Estimation for small areas using spatio-temporal Fay-Herriot model

In the paper the primary target results were those related to the Spatio-Temporal Fay-Ferriot model (STFH), being the extension of the classical Fay-Herriot small area model. The methodology for such models was described in detail by Marhuenda, Molina and Morales (2013). Under the spatio-temporal small area model, the area parameter for domain *d* at current time instant *T* is estimated as borrowing strength from other time instants and from the *D* domains. Let θ_{dt} represent the variable of interest determined for area *d* and time *t* where $d = 1, \dots, D$, and $t = 1, \dots, T$. If the direct estimator of this quantity is denoted by $\hat{\theta}_{dt}^{DIR}$, and the sampling errors are expressed as e_{dt}, which are assumed to be independent and normally distributed with known variances ψ_{dt}, the spatio-temporal model can be written as below

$$\hat{\theta}_{dt}^{DIR} = \theta_{dt} + e_{dt} \tag{1}$$

The relationship above is valid for all considered *d* and *t*. The equation (1) can also be expressed by means of the model which incorporates spatio-temporal relationships of the form

$$\theta_{dt} = \mathbf{x}_{dt}^{T}\beta + u_{1d} + u_{2dt} \tag{2}$$

Here, $\mathbf{x}_{dt}$ are the vectors of *p* auxiliary variables representing regression structure of θ_{dt}, with regression coefficients *β*. The area-time random effects can be expressed by $(u_{2d1}, \dots, u_{2dT})^T$ and are assumed identically and independently distributed for each area. Moreover, they follow the AR(1) process with autocorrelation parameter ρ_2, which can be described as

$$u_{2dt} = \rho_2 u_{2d,t-1} + \epsilon_{2dt}, \text{ where } |\rho_2| < 1 \text{ and } \epsilon_{2dt} \overset{iid}{\sim} N(0, \sigma_2^2) \tag{3}$$

The area-related random effects, expressed by $(u_{11}, \dots, u_{1D})^T$, are subject to the SAR process with variance parameter σ_1^2, spatial autocorrelation ρ_1 and proximity matrix **W** = ($w_{d,l}$), which can be obtained from an original proximity matrix $\mathbf{W}^0$, whose diagonal elements are equal to zero and the remaining entries are equal to 1 (when the two areas corresponding to the row and the column indices are considered as neighbour and zero otherwise). Then, **W** is obtained by row-standardization of $\mathbf{W}^0$, obtained by dividing each entry of $\mathbf{W}^0$ by the sum of elements in the same row. The area level random effects can be described as

$$u_{id} = \rho_1 \sum_{l \neq d} w_{d,l} u_{1l} + \epsilon_{1d} \text{ where } |\rho_1| < 1 \text{ and } \epsilon_{1d} \overset{iid}{\sim} N(0, \sigma_1^2) \tag{4}$$

Using the stacking notations for vectors and matrices one can present the following relationships for the considered model:

$$\mathbf{y} = \underset{1 \le d \le D}{col}\left(\underset{1 \le t \le T}{col}(\hat{\theta}_{dt}^{dir})\right), \quad \mathbf{X} = \underset{1 \le d \le D}{col}\left(\underset{1 \le t \le T}{col}(x_{dt}^{T})\right)$$

$$\mathbf{e} = \underset{1 \le d \le D}{col}\left(\underset{1 \le t \le T}{col}(e_{dt})\right), \mathbf{u}_1 = \underset{1 \le d \le D}{col}(u_{1d}), \mathbf{u}_2 = \underset{1 \le d \le D}{col}\left(\underset{1 \le t \le T}{col}(u_{2dt})\right)$$

Also, one can define additionally $\mathbf{Z}_1 = \mathbf{I}_D \otimes \mathbf{1}_T$, where $\mathbf{I}_D$, is the D x D identity matrix, $\mathbf{1}_T$ is the vector of 1's and has length T, and ⊗ is the Kronecker product, $\mathbf{Z}_2 = \mathbf{I}_n$, where *n=DT* is the total number of observations, $\mathbf{u} = (\mathbf{u}_1^T, \mathbf{u}_2^T)^T$ and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$.

Using the notation presented above we can describe the spatio-temporal model in terms of the general linear mixed model, which has the following form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

Let $\boldsymbol{\delta} = (\sigma_1^2, \rho_1, \sigma_2^2, \rho_2)$ denote the vector of covariance structure parameters involved in the model above. We can use the following relationships for the vector **e** related to direct estimation error: $\mathbf{e} \sim N(\mathbf{0}_n, \boldsymbol{\psi})$, where $\mathbf{0}_n$ denotes a vector of zeroes that has the length *n* and $\boldsymbol{\psi}$ is the diagonal matrix $\psi = diag_{1\le d\le D}(diag_{1\le t\le T}(\psi_{dt}))$.

The random effects $\mathbf{u}$ are also normally distributed $\mathbf{u} \sim N\{\mathbf{0}_n, \mathbf{G}(\boldsymbol{\delta})\}$, where the covariance matrix **G** can be expressed as the block diagonal matrix of the following form: $\mathbf{G}(\boldsymbol{\delta}) = diag\{\sigma_1^2 \boldsymbol{\Omega}_1(\rho_1), \sigma_2^2 \boldsymbol{\Omega}_2(\rho_2)\}$. The matrices $\boldsymbol{\Omega}_1$and $\boldsymbol{\Omega}_2$ are defined as

$$\boldsymbol{\Omega}_1(\rho_1) = \{(\mathbf{I}_D - \rho_1 \mathbf{W})^T (\mathbf{I}_D - \rho_1 \mathbf{W})\}^{-1} \tag{5}$$

$$\boldsymbol{\Omega}_2(\rho_2) = diag_{1\le d\le D}\{\boldsymbol{\Omega}_{2d}(\rho_2)\}$$

$$\boldsymbol{\Omega}_{2d}(\rho_2) = \frac{1}{1-\rho_2^2} \begin{pmatrix} 1 & \rho_2 & 0 \ldots & \rho_2^{T-2} & \rho_2^{T-1} \\ \rho_2 & 1 & \ddots & 1 & \rho_2^{T-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho_2^{T-2} & & \ddots & 1 & \rho_2 \\ \rho_2^{T-1} & \rho_2^{T-2} & \ldots & \rho_2 & 1 \end{pmatrix} \tag{6}$$

The covariance matrix for the full model (including the sampling error) can be expressed as

$$\mathbf{V}(\boldsymbol{\delta}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T + \boldsymbol{\Psi}$$

The vector **β** and the random effects **u** can be obtained using BLUP estimator $\widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta})$ by means of the following equations, utilizing **X**, **G**, **V** and **Z** matrices:

$$\widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) = \{\mathbf{X}^T \mathbf{V}^{-1}(\boldsymbol{\delta}) \mathbf{X}\}^{-1} \mathbf{X}^T \mathbf{V}^{-1}(\boldsymbol{\delta}) \mathbf{y}$$

$$\widetilde{\mathbf{u}}(\boldsymbol{\delta}) = \mathbf{G}(\boldsymbol{\delta}) \mathbf{Z}^T \mathbf{V}^{-1}(\boldsymbol{\delta}) \{\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta})\}$$

Because $\mathbf{u} = (\mathbf{u}_1^T, \mathbf{u}_2^T)^T$, the second equation given above can be decomposed as follows

$$\widetilde{\boldsymbol{u}}_{\mathbf{1}}(\boldsymbol{\delta}) = \sigma_1^2 \boldsymbol{\Omega}_1(\rho_1) \mathbf{Z}_1^T \mathbf{V}^{-1}(\boldsymbol{\delta}) \{\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta})\} \tag{7}$$

$$\widetilde{\mathbf{u}}_2(\boldsymbol{\delta}) = \sigma_2^2 \boldsymbol{\Omega}_2(\rho_2) \mathbf{V}^{-1}(\boldsymbol{\delta}) \{\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta})\}$$

3. REML estimation method for spatio-temporal model

The method of Restricted Maximum Likelihood (REML) uses maximization for the likelihood function, which corresponds to the joint probability density function as a vector of *n-p* linearly independent contrasts expressed as $\mathbf{F}^{\mathbf{T}}\mathbf{y}$ where **F** is the $n \times (n-p)$ full column rank satisfying the relationships $\mathbf{F}^T\mathbf{F} = \mathbf{I}_{n-p}$ and $\mathbf{F}^T\mathbf{X} = \mathbf{0}_{n-p}$. From the previous conditions, the probability density function of the contrast vectors can be expressed as

$$f_R(\boldsymbol{\delta};\mathbf{y}) = (2\pi)^{-(n-p)/2}|\mathbf{X}^T\mathbf{X}|^{1/2}|\mathbf{V}(\boldsymbol{\delta})|^{-1/2}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}^T\mathbf{P}(\boldsymbol{\delta})\mathbf{y}\right\}$$

where **P** matrix satisfies the condition

$$\mathbf{P}(\boldsymbol{\delta}) = \mathbf{V}^{-1}(\boldsymbol{\delta}) - \mathbf{V}^{-1}(\boldsymbol{\delta})\mathbf{X}\{\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\delta})\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\delta})$$

The matrix **P** satisfies the following relationships $\mathbf{P}(\boldsymbol{\delta})\mathbf{V}(\boldsymbol{\delta})\mathbf{P}(\boldsymbol{\delta}) = \mathbf{P}(\boldsymbol{\delta})$ and $\mathbf{P}(\boldsymbol{\delta})\mathbf{X} = \mathbf{0}_n$. The REML estimator maximizes the log likelihood function $\ell_R(\boldsymbol{\delta};\mathbf{y}) = \log f_R(\boldsymbol{\delta};\mathbf{y})$ using Fisher scoring algorithm. This algorithm utilizes scoring vectors of the form $S_R(\boldsymbol{\delta}) = \partial\ell_R(\boldsymbol{\delta};\mathbf{y})/\partial\boldsymbol{\delta}$ as well as the Fisher information matrix which is $\mathfrak{I}_R(\boldsymbol{\delta}) = -E\left\{\frac{\partial^2\ell_R(\delta;y)}{\partial\delta\partial\delta'}\right\} = (\mathfrak{I}_{rs}^R(\boldsymbol{\delta}))$.

For the spatio-temporal model with four variance components we have the following relationship

$$\frac{\partial\mathbf{P}(\boldsymbol{\delta})}{\partial\delta_r} = -\mathbf{P}(\boldsymbol{\delta})\frac{\partial\mathbf{V}(\boldsymbol{\delta})}{\partial\delta_r}\mathbf{P}(\boldsymbol{\delta})$$

for *r=1,…,4*. The first order derivative of $\ell_R(\boldsymbol{\delta};\mathbf{y})$, with respect to δ_r can be given as below

$$S_r^R(\boldsymbol{\delta}) = -\frac{1}{2}tr\left\{\mathbf{P}(\boldsymbol{\delta})\frac{\partial\mathbf{V}(\boldsymbol{\delta})}{\partial\delta_r}\right\} + \frac{1}{2}\mathbf{y}^T\mathbf{P}(\boldsymbol{\delta})\frac{\partial\mathbf{V}(\boldsymbol{\delta})}{\partial\delta_r}\mathbf{P}(\boldsymbol{\delta})\mathbf{y}$$

so the element indexed by (*r*,*s*) in the Fisher information matrix can be expressed as

$$\mathfrak{I}_{rs}^R(\boldsymbol{\delta}) = \frac{1}{2}tr\left\{\mathbf{P}(\boldsymbol{\delta})\frac{\partial\mathbf{V}(\boldsymbol{\delta})}{\partial\delta_r}\mathbf{P}(\boldsymbol{\delta})\frac{\partial\mathbf{V}(\boldsymbol{\delta})}{\partial\delta_s}\right\}$$

The detailed expressions for the partial derivatives of **V** with respect to the variance components used in the expression for scoring vectors and the Fisher information matrix have the following form:

$$\frac{\partial V(\boldsymbol{\delta})}{\partial\sigma_1^2} = \mathbf{Z}_1\Omega_1(\rho_1)\mathbf{Z}_1^T, \qquad \frac{\partial\mathbf{V}(\boldsymbol{\delta})}{\partial\rho_1} = -\sigma_1^2\mathbf{Z}_1\boldsymbol{\Omega}_1(\rho_1)\frac{\partial\boldsymbol{\Omega}_1^{-1}(\rho_1)}{\partial\rho_1}\boldsymbol{\Omega}_1(\rho_1)\mathbf{Z}_1^T$$

$$\frac{\partial V(\boldsymbol{\delta})}{\partial\sigma_2^2} = \underset{1\le d\le D}{diag}\{\boldsymbol{\Omega}_{2d}(\rho_2)\} \qquad \frac{\partial V(\boldsymbol{\delta})}{\delta\rho_2} = \sigma_2^2\underset{1\le d\le D}{diag}\left\{\frac{\partial\boldsymbol{\Omega}_{2d}(\rho_2)}{\partial\rho_2}\right\}$$

where

$$\frac{\partial \mathbf{\Omega}_1^{-1}(\rho_1)}{\partial \rho_1} = -\mathbf{W} - \mathbf{W}^T + 2\rho_1 \mathbf{W}^T\mathbf{W},$$

$$\frac{\partial \mathbf{\Omega}_{2d}(\rho_2)}{\partial \rho_2} = \frac{1}{1-\rho_2^2}\begin{pmatrix} 0 & 1 & \cdots & \cdots & (T-1)\rho_2^{T-2} \\ 1 & 0 & \ddots & & (T-1)\rho_2^{T-3} \\ \vdots & \ddots & \ddots & & \vdots \\ (T-2)\rho_2^{T-3} & & \ddots & 0 & 1 \\ (T-1)\rho_2^{T-2} & \dots & \dots & 1 & 0 \end{pmatrix} + \frac{2\rho_2 \mathbf{\Omega}_{2d}(\rho_2)}{1-\rho_2^2}$$

Finally, the scoring algorithm assumes that the variance component vector converges to the common value, using the following iterative procedure

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{\delta}^{(k)} + \mathfrak{I}_{rs}^{R}(\boldsymbol{\delta}^{(k)}) S_R(\boldsymbol{\delta}^{(k)})$$

4. Determining the MSE of spatio-temporal estimates using parametric bootstrap method.

The estimation of MSE of spatio-temporal estimator was determined using the parametric bootstrap method implemented in sae package. This method can be summarized as follows:

1. Using direct income estimates and available auxiliary data $\{(\hat{\theta}_{dt}^{DIR}, x_{dt})$, *t=1,..,T, d=1,…,D*}, obtain the estimates of the STFH model described by the equations (1) - (4) together with the estimates of the model parameter **β** and **δ.**

2. Generate bootstrap area effects $\{u_{1d}^{*(b)}$, *d=1,…,D* } from the SAR(1) process given in (4), assuming $(\hat{\sigma}_1^2, \hat{\rho}_1)$ as true values of (σ_1^2, ρ_1)

3. Independently of $\{u_{1d}^{*(b)}\}$ and independently for each d, generate bootstrap time effects $\{u_{2dt}^{*(b)}$, *t=1,…,T* } from the AR(1) process given in (3), with $(\hat{\sigma}_2^2, \hat{\rho}_2)$ acting as true values of the parameters (σ_2^2, ρ_2)

4. Calculate true bootstrap quantities, using the formula

$$\theta_{dt}^{*(b)} = \mathbf{x}_{dt}^T \beta + u_{1d}^{*(b)} + u_{2dt}^{*(b)}$$

5. Generate errors $e_{dt}^{*(b)} \overset{ind}{\sim} N(0, \psi_{dt})$ and obtain bootstrap data from the sampling model

$$\hat{\theta}_{dt}^{DIR*(b)} = \theta_{dt}^{*(b)} + e_{dt}^{*(b)}$$

6. Using the new bootstrap data $\{(\hat{\theta}_{dt}^{DIR*(b)}, x_{dt})$, *t=1,..,T, d=1,…,D*} determine the estimates of STFH model described with equations from (1) to (4) and obtain the bootstrap EBLUPs

$$\hat{\theta}_{dt}^{*(b)} = \mathbf{x}_{dt}^T \hat{\beta}^{*(b)} + \hat{u}_{1d}^{*(b)} + \hat{u}_{2dt}^{*(b)}$$

7. Repeat steps (1)-(6) for *b = 1, ... ,B* , where B is a large number.

8. The parametric bootstrap MSE estimates are given by

$$mse\left(\hat{\theta}_{dt}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{dt}^{*(b)} - \theta_{dt}^{*(b)}\right)^2$$

5. Results and discussion

In the application, we were interested in the estimation of various per capita income components (in particular *income from social-security benefits*) by region NUTS2, based on the micro data coming from the Polish Household Budget Survey (HBS) and regional data obtained from the GUS Local Data Bank. Spatial and spatio-temporal models can fit to this kind of situations as they account for the correlation related to neighbourhood of areas and time-dependency, which both determine the random effects. They are based on cross-sectional and time-series data involving spatial autocorrelation. The model-based approach generally improves the estimation quality due to the use of explanatory variables coming from administrative registers and area random effects, which additionally account for the variability between domains. In current approach we can have extra gains coming from spatial and time dependencies between areas. To obtain better estimates for the year 2013, we decided to utilize historical data coming from ten years before, which enabled “borrowing strength” not only across areas but also over time and space. The results obtained on the basis of these models were compared to the ones obtained from the classical Fay-Herriot model and to direct estimates.

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 (see Westat (2007) for details). In the computations conducted in R-project environment, the packages sae, sae2 and spdep were applied.

At the second stage, the appropriate 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 **Spatio-Temporal model (STFH)** we also estimated the parameters of simpler small area models. In particular, we additionally estimated the parameters of the following small-area models:

- **Rao-Yu model (RY)**, (“borrowing strength” from areas and over time),
- **Spatial Fay-Herriot model (SFH),** (“borrowing strength” only from other areas with proximity matrices, separately for each of them),
- **Fay-Herriot model (FH)**, (“borrowing strength” only from other areas),

and additionally, for comparison purposes, we estimated the unknown parameters using:

- classical spatial econometrics models based on SAR process, including **spatial lag model (lagsar)** and **spatial error model (errorsar)**.

At the third stage, using the model parameters which were estimated at the second stage, we obtained the predictors of per capita income for regions in Poland. In particular, for the STFH, FH, SFH, RY models, the appropriate EBLUPs were obtained, while for the spatial econometric models the appropriate linear predictors were evaluated.

The sae package made it possible to obtain estimation for spatial related model and spatio-temporal model. The sae2 package includes the implementation of the estimation procedure for the Rao-Yu model (see Fay R.E., Li J. (2012), Li J., Diallo M.S., Fay R.E. (2012), Fay, R. E., Diallo, M., (2015)), which provides an extension of the basic type A model to handle time series and cross-sectional data (Rao, Molina (2015)). The spdep package (Bivand, R., Lewin-Koh, N., (2013), Bivand, R., Piras, G., (2015)) was used for estimation of classical spatial econometrics models and also Moran I statistics for the considered variables. Our own R macro was developed, performing calculations for ordinary EBLUP models, spatial EBLUP models, Rao-Yu models, classical econometric models and spatio-temporal models, model diagnostics, as well as the maps for regions.

Table 1. Diagnostics of Rao-Yu and Spatio-Temporal models of income per capita from *social security benefits* based on sample and administrative data

Model/ explanatory variable	Coefficient estimates	Standard error	t-statistics	p-value
1. Rao-Yu model	σ_2^2 = 111.410	σ_1^2 =124.710	ρ_2=0.540	
Intercept	51.1520	16.7460	3.0546	0.0023
Average salary in nat.economy	0.0168	0.0123	1.3680	0.1713
Average retirements pay	0.1424	0.0236	6.0462	1.483E-09
GDP per capita (Poland 100%)	-0.3314	0.1833	-1.8081	0.0706
2. Spatio-temporal model	σ_2^2 = 110.640	σ_1^2 =88.691	ρ_1 =0.856	ρ_2=0.501
Intercept	64.8670	21.3040	3.0448	0.0023
Average salary in nat.economy	0.0261	0.0124	2.1052	0.0353
Average retirements pay	0.1245	0.0237	5.2651	1.402E-07
GDP per capita (Poland 100%)	-0.4945	0.1653	-2.9913	0.0028

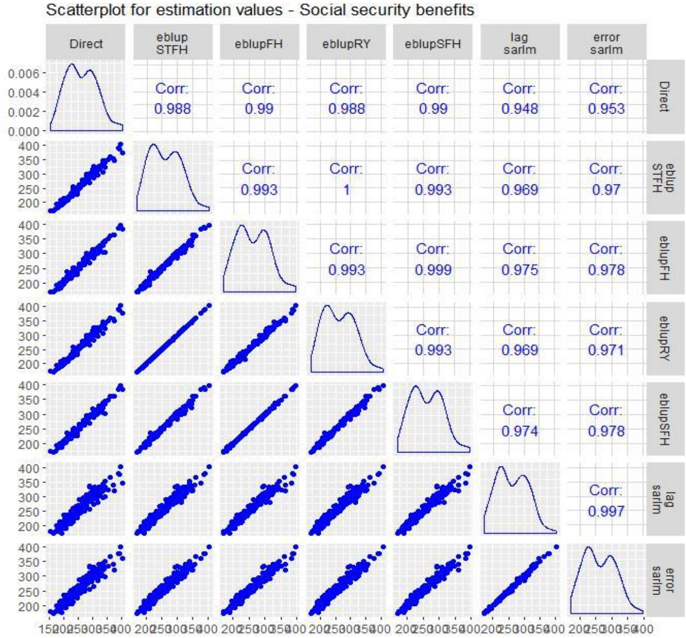
Source: Authors' calculations based on the Polish Household Budget Survey and data from the Local Data Bank.

Table 2. Estimation results for per capita income from *social security benefits* by region in Poland for the year 2013

Region	Direct estimate		1. Rao-Yu estimate			2. Spatio-temporal estimate		
	Value	REE	Value	REE	Time-related random effect	Value	REE	Time-related random effect
	zł.	%	zł.	%	zł.	zł.	%	zł.
Dolnośląskie	401.58	4.28	375.83	2.75	16.936	375.66	2.55	15.951
Kujaw. Pomor.	329.55	3.69	327.55	2.64	9.699	326.43	2.34	10.640
Lubelskie	282.68	3.50	298.15	2.65	-20.880	298.20	2.78	-20.569
Lubuskie	333.49	6.70	346.87	3.24	11.211	346.04	3.17	9.664
Łódzkie	354.50	2.97	350.90	2.31	9.263	350.10	2.14	9.794
Małopolskie	305.97	3.74	320.03	2.63	-13.364	319.93	2.43	-13.158
Mazowieckie	335.47	3.23	339.16	2.41	-7.202	338.44	2.44	-5.301
Opolskie	342.17	5.89	343.94	3.30	-4.882	345.93	2.82	-6.008
Podkarpackie	292.47	3.41	299.82	2.57	-9.169	299.47	2.63	-8.394
Podlaskie	325.53	5.03	323.70	3.00	3.633	322.50	2.95	4.609
Pomorskie	320.82	5.62	326.00	3.25	-5.475	325.55	3.43	-4.576
Śląskie	395.95	2.86	404.36	2.16	-7.007	403.70	2.30	-6.054
Świętokrzyskie	329.61	4.49	329.70	2.89	4.471	329.32	3.07	4.145
Warm.-Mazur.	297.09	5.83	306.51	3.45	-5.081	304.95	3.88	-3.562
Wielkopolskie	290.97	2.48	297.73	2.10	-21.435	298.20	2.23	-22.928
Zach.-pomor.	333.86	7.52	335.91	3.39	-5.610	336.01	3.49	-5.285

Source: Authors' calculations based on the Polish Household Budget Survey and data from the Local Data Bank.

Figure 1. Pair scatterplots of direct and model-based estimates for per capita income from *social security benefits* by region in Poland for the years 2003–2013



Source: Authors' calculations based on the Polish HBS and Local Data Bank.

Table 3. Variance structure parameters of selected small area and econometric models together with their log likelihoods for per capita income from *social security benefits* by region in Poland for the years 2003–2013

Year	Fay-Herriot model		Spatial Fay-Herriot model			Spatial SAR error model		
	σ_1^2	Log likelihood	σ_1^2	ρ	Log likelihood	σ^2	λ	Log likelihood
2003	93.82	-61.95	43.34	0.961	-62.56	92.38	0.572	-59.69
2004	223.00	-66.34	146.24	0.829	-66.02	184.78	0.523	-65.09
2005	118.48	-62.76	116.01	0.371	-62.93	153.29	0.058	-62.97
2006	38.29	-59.62	42.75	0.342	-59.88	117.54	0.309	-61.04
2007	275.76	-68.37	245.69	0.625	-68.24	215.60	0.600	-66.56
2008	261.88	-69.35	205.51	0.826	-69.98	355.35	0.401	-70.03
2009	318.67	-69.02	273.93	-0.411	-68.40	215.58	-0.953	-67.39
2010	202.88	-68.20	196.06	-0.130	-68.07	247.33	-0.646	-67.54
2011	133.95	-66.56	113.34	0.702	-66.78	302.40	0.302	-68.59
2012	558.55	-73.16	551.95	-0.047	-73.09	497.60	-0.341	-72.59
2013	615.95	-74.11	421.61	-0.693	-72.86	369.39	-0.903	-71.51

Source: Authors' calculations based on the Polish HBS and Local Data Bank.

In Table 1 we show estimation results obtained for the Rao-Yu and spatio-temporal STFH model (eq.:(1)-(4)). For each dependent variable the estimates of fixed effects **β** and the parameters of variance-covariance structure of the models denoted as **δ** are presented. The model diagnostics indicate that both parameters related to the variability of random effects (σ) have visible contribution to the variability of the model. This is in contrast with previously presented models (see Jędrzejczak, Kubacki (2016)), where for the model of available income this variability was mostly determined by time-related component. It should be noted that a similar comparison of small area-models of available income also revealed such relationships for the Rao-Yu and spatio-temporal models, which may mean that both these approaches are complementary. Figures 5 and 6 additionally show decompositions of time-related random effects of the Rao-Yu and STFH models and make it possible to observe the impact and the distribution of these effects over time. Figure 1 summarizes dependencies between all pairs of estimates obtained in the study over the analysed period by means of the Pearson correlation coefficients and scatterplots.

In Table 2 there are income estimates, their corresponding relative estimation errors (REE) and time-related random effects u_2. Covariance structure parameters together with log likelihoods for all the years and selected models are given in Table 3. In the table, for comparison purposes, also the spatial SAR error model was provided due to identical assumptions about spatial random effects.

Precision of different estimation methods can be analysed on the basis of the detailed results given in Tables 4-6 and in Figures 2-4. The tables show Consistency Coefficients (CC), Relative Estimation Errors (REE) and REE reduction, respectively. The consistency coefficients presented in Table 4 and Figure 2 can be defined as follows

$$CC_{dt} = (\theta_{dt}^{model} - \hat{\theta}_{dt}^{DIR})/\hat{\theta}_{dt}^{DIR}$$

This measure can be used as a simple approximation of bias of model-based estimates. The results given in Table 4 and Figure 2 indicate that simpler estimation techniques may be less biased than the more complicated ones (Rao-Yu and STFH models). It seems that introducing more assumptions about the random-effects not always reflects the real-world relationships. Special attention should be paid to the values of CC obtained for econometric spatial error (SAR) model, which confirm that the classical spatial econometric models may be insufficient for small area estimation (Figure 2).

In Table 5, REE values for different estimation techniques are summarized, while in Table 6 REE reduction is presented with respect to both: direct and ordinary EBLUP estimates, corresponding to the first and the second column for each model. This approach can be helpful to recognise efficiency gains coming from model-based estimation and additional gains coming from temporal (and spatial) correlation incorporated in more advanced small area models.

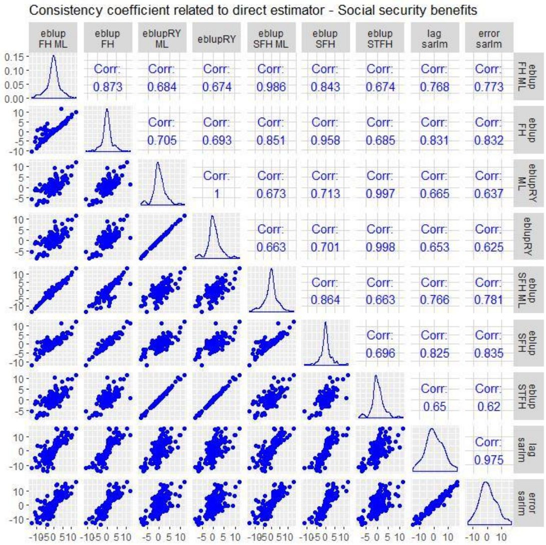
Comparisons of the distributions of REE and REE reduction (Figure 3 and Figure 4) show that all the considered model-based techniques are significantly more efficient than the corresponding direct ones. The Rao-Yu model and spatio-temporal model STFH perform similarly, as compared to the other model-based techniques which have been considered in the study. This regularity can also be observed when a comparison between a spatial, spatio-temporal and Rao-Yu models in terms of REE Reduction related to the ordinary EBLUPs is made.

Table 4. Consistency Coefficients [in %] for model-based estimates related to direct estimates for per capita income from *social security benefits* by region in Poland for the year 2013

Region	Consistency coefficient [in %]			
	Fay-Heriot EBLUP	Spatial Fay-Herriot EBLUP	Rao-Yu EBLUP	Spatio-temporal EBLUP
Dolnośląskie	-4.759	-4.620	-6.413	-6.455
Kujaw. Pomor.	-0.549	-0.049	-0.606	-0.948
Lubelskie	1.052	0.544	5.474	5.492
Lubuskie	-2.020	-3.089	4.013	3.762
Łódzkie	-1.384	-1.501	-1.017	-1.241
Małopolskie	1.520	1.803	4.594	4.563
Mazowieckie	0.263	0.150	1.101	0.885
Opolskie	-0.844	-2.366	0.518	1.098
Podkarpackie	0.397	0.688	2.511	2.395
Podlaskie	-1.370	-1.182	-0.562	-0.931
Pomorskie	1.287	2.550	1.615	1.474
Śląskie	0.251	0.185	2.123	1.958
Świętokrzyskie	-1.437	-1.473	0.028	-0.089
Warm.-Mazur.	1.673	1.149	3.171	2.646
Wielkopolskie	1.115	1.065	2.324	2.485
Zach.-pomor.	-0.752	0.979	0.61	0.645

Source: Authors' calculations based on the Polish HBS and Local Data Bank.

Figure 2. Pair scatterplots for CC obtained for different model-based estimates



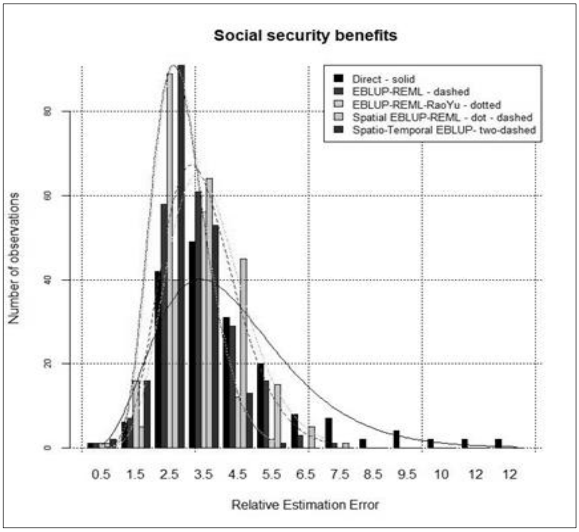
Source: Authors' calculations based on the Polish HBS and Local Data Bank.

Table 5. Estimates of REE [in %] for different estimates of income per capita from *social security benefits* by voivodships for the year 2013

Region	Relative estimation error [in %]				
	Direct	Fay-Heriot EBLUP	Spatial Fay-Herriot EBLUP	Rao-Yu EBLUP	Spatio-temporal EBLUP
Dolnośląskie	4.28	3.96	4.10	2.75	2.55
Kujaw. Pomor.	3.69	3.50	3.67	2.64	2.34
Lubelskie	3.50	3.39	3.63	2.65	2.78
Lubuskie	6.70	5.61	5.91	3.24	3.17
Łódzkie	2.97	2.88	3.04	2.31	2.14
Małopolskie	3.74	3.48	3.68	2.63	2.43
Mazowieckie	3.23	3.27	3.46	2.41	2.44
Opolskie	5.89	4.96	5.39	3.30	2.82
Podkarpackie	3.41	3.28	3.48	2.57	2.63
Podlaskie	5.03	4.58	4.86	3.00	2.95
Pomorskie	5.62	4.85	4.87	3.25	3.43
Śląskie	2.86	2.90	3.07	2.16	2.30
Świętokrzyskie	4.49	4.15	4.48	2.89	3.07
Warm.-Mazur.	5.83	5.04	5.24	3.45	3.88
Wielkopolskie	2.48	2.44	2.52	2.10	2.23
Zach.-pomor.	7.52	5.75	5.72	3.39	3.49

Source: Authors' calculations based on the Polish HBS and Local Data Bank.

Figure 3. Distribution of estimated relative estimation errors (REE) for different estimation methods (direct and model-based)



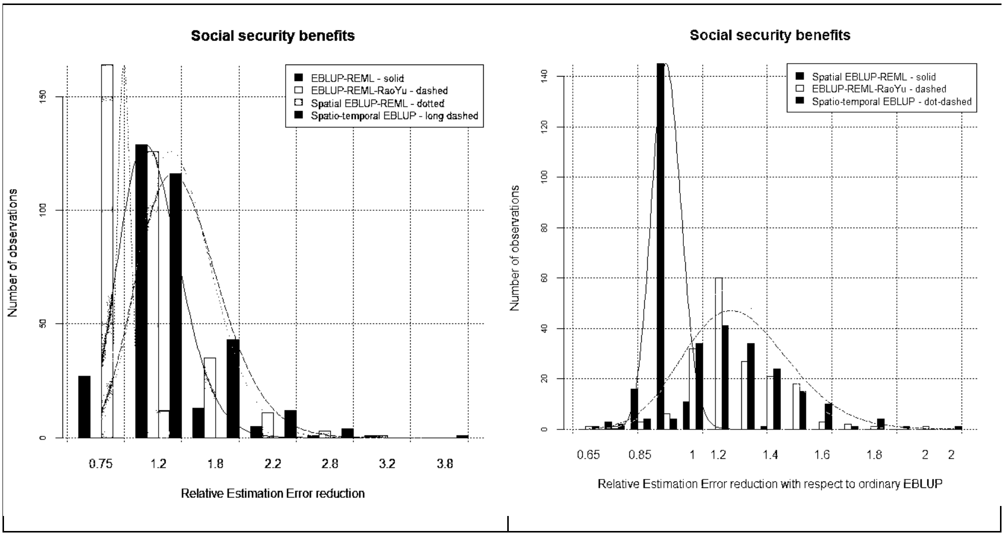
Source: Authors' calculations based on the Polish HBS and Local Data Bank.

Table 6. REE Reduction for income per capita from *social security benefits* related to the direct and FH estimates by voivodships for the year 2013

Region	FH EBLUP	SFH EBLUP		RY EBLUP		STFH EBLUP	
	REE reduction	REE reduction	Spatial REE reduction	REE reduction	Spatio-temporal REE reduction	REE reduction	Spatio-temporal REE reduction
Dolnośląskie	1.0796	1.0431	0.9662	1.5575	1.4427	1.6802	1.5563
Kujaw. Pomor.	1.0549	1.0066	0.9542	1.4010	1.3281	1.5798	1.4976
Lubelskie	1.0342	0.9663	0.9344	1.3231	1.2794	1.2593	1.2177
Lubuskie	1.1946	1.1339	0.9491	2.0684	1.7314	2.1127	1.7685
Łódzkie	1.0312	0.9775	0.9479	1.2860	1.2471	1.3863	1.3444
Małopolskie	1.0756	1.0154	0.9440	1.4191	1.3193	1.5359	1.4279
Mazowieckie	0.9874	0.9330	0.9449	1.3415	1.3587	1.3272	1.3442
Opolskie	1.1858	1.0916	0.9206	1.7817	1.5026	2.0906	1.7631
Podkarpackie	1.0403	0.9813	0.9432	1.3257	1.2743	1.2957	1.2455
Podlaskie	1.0974	1.0335	0.9418	1.6770	1.5281	1.7060	1.5546
Pomorskie	1.1588	1.1536	0.9955	1.7287	1.4918	1.6369	1.4126
Śląskie	0.9882	0.9335	0.9446	1.3267	1.3425	1.2450	1.2599
Świętokrzyskie	1.0812	1.0026	0.9273	1.5520	1.4354	1.4598	1.3501
Warm.-Mazur.	1.1570	1.1122	0.9613	1.6911	1.4617	1.5039	1.2999
Wielkopolskie	1.0156	0.9858	0.9706	1.1819	1.1637	1.1118	1.0947
Zach.-pomor.	1.3075	1.3141	1.0051	2.2186	1.6969	2.1542	1.6476

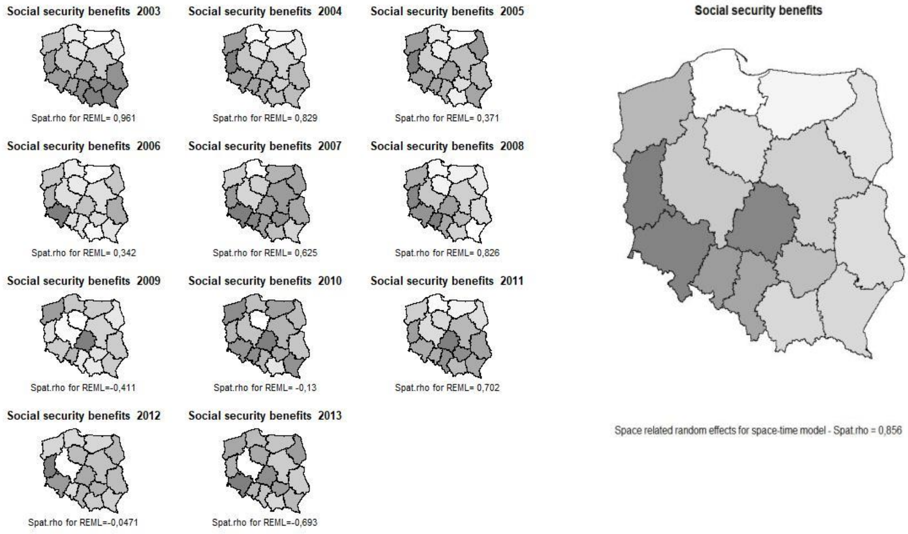
Source: Authors' calculations based on the Polish HBS and Local Data Bank.

Figure 4. Distribution of REE Reduction (left) and REE Reduction due to spatio-temporal effects (right) for different model-based estimation methods



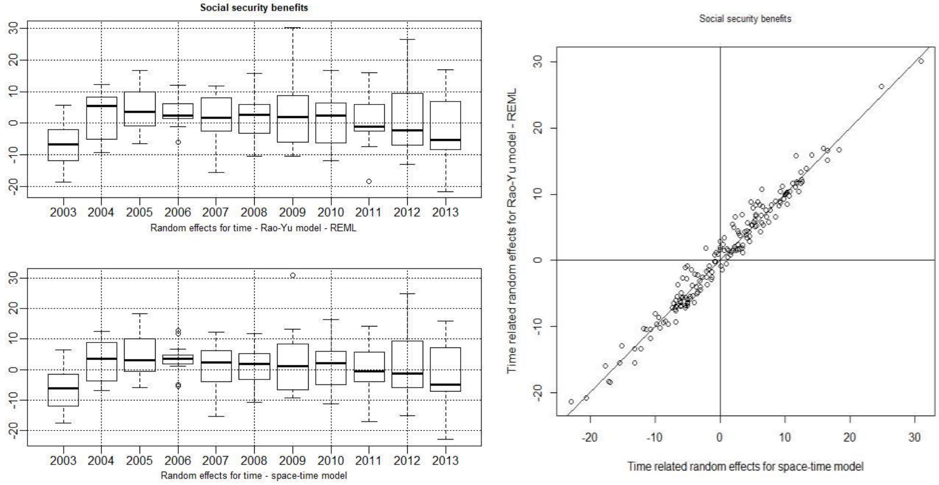
Source: Authors' calculations based on the Polish HBS and Local Data Bank.

Figure 5. Choropleth maps for **spatial random effects** of Spatial Fay-Herriot model (left) and Spatio-Temporal model (right) for *per capita income from social security benefits* by region


Social security benefits 2003
Spat.rho for REML= 0,961
Social security benefits 2004
Spat.rho for REML= 0,829
Social security benefits 2005
Spat.rho for REML= 0,371
Social security benefits 2006
Spat.rho for REML= 0,342
Social security benefits 2007
Spat.rho for REML= 0,625
Social security benefits 2008
Spat.rho for REML= 0,826
Social security benefits 2009
Spat.rho for REML=-0,411
Social security benefits 2010
Spat.rho for REML= -0,13
Social security benefits 2011
Spat.rho for REML= 0,702
Social security benefits 2012
Spat.rho for REML=-0,0471
Social security benefits 2013
Spat.rho for REML=-0,693
Social security benefits
Space related random effects for space-time model - Spat.rho = 0,856

Source: Authors' calculations based on the Polish HBS and Local Data Bank.

Figure 6. Distributions of random **time-related effects** for RY model (top-left), and STFH model (bottom-left) and the scatterplot (right) for time-related effects (*u_2*) for STFH and Rao-Yu model.



Source: Authors' calculations based on the Polish HBS and Local Data Bank.

The maps presented in Figure 5 show the spatial structure of space-related random effects. It can be noticed that for the spatial models under consideration (but not for all years) significant spatial relationships are obtained. Note that such a behaviour becomes evident for the regions where higher spatial autoregression coefficients are observed, i.e. for *north-western* and *south-eastern* regions. This confirms the well-known relationships of regional differences in Poland and is obviously connected with higher industrialization of these voivodships (which may cause larger income from social security benefits). Similar conclusions were made in Kubacki and Jędrzejczak (2016), where spatial relationships for small area models in Polish counties were presented.

Interesting relationships can be observed when comparing the distributions of time-related random effects obtained for the Rao-Yu model and for the spatio-temporal model, as illustrated in Figure 6. Here some consistency between time-related random effects is noticeable. It results from the fact that in the STFH model as well as in the Rao-Yu model they follow the autoregressive process of the first order, AR(1). This regularity can also be observed in the distributions of random effects presented for each year and in the scatterplot obtained for all the years. Note that the values of random effects related to time were determined by different estimators, and calculated using different software (in particular sae and sae2 packages).

Some consistency can also be observed for REE and REE reduction distributions (Fig. 3, Fig. 4), obtained for the Rao-Yu and spatio-temporal models. It should be noted, however, that the methods used to obtain the REE values for these models were also different. In the case of the Rao-Yu model, the method applied for MSE estimation was based on extensions of the Prasad and Rao

(1990) approach, while in the case of spatio-temporal model it was based on the parametric bootstrap technique. Of course, these methods have different implementations, which may indicate that they are convergent. Such relationships were also obtained for other income-related variables.

6. Conclusions

The paper shows a procedure of efficient estimation for small areas based on the application of a spatio-temporal model, i.e. the general linear mixed model with spatially correlated random effects and significant correlation over time. In particular, spatial Simultaneous Autoregressive Process, using spatial neighbourhood as auxiliary information, and AR(1) process for time-related random effects, were incorporated into the estimation.

The presented spatio-temporal model improves the precision of small-area estimates not only in relation to direct estimates, which is easy to obtain, but also in comparison with other indirect techniques based on small-area models, also spatial small area models and sometimes the Rao-Yu model.

The efficiency of the proposed method was proven based on real-world examples prepared for the Polish data coming from the Household Budget Survey and the administrative data. The detailed comparison of relative estimation errors and REE reductions shows that all the considered model-based techniques are significantly more efficient than the direct estimation one, yet the spatio-temporal and the Rao-Yu models provide greater REE reduction than the others. The calculations, where some additional assumptions on the spatial relationships were made, also confirm efficiency gains of the estimators. However, such a correspondence does not always occur for all the years, so one should be conscious that for lower ρ_2 values the benefit of using the spatial method may be ambiguous.

It is worth pointing out that the number of observations used to fit the area-level models was small so the model parameters were estimated with less efficiency and therefore the efficiency gains with respect to direct estimators were obviously smaller than under the unit level models. What is more, the applied models require normality of random effects for MSE estimation and violating this assumption can seriously affect the results.

A more detailed analysis also reveals some correspondence between the Rao-Yu model and spatio-temporal models induced by identical assumptions about the stochastic process for time-related random effects. Further benefits can be expected when time-dependent nonlinear relationships are taken into account, for example nonlinear dependence on explanatory variables. The previously performed analysis of nonlinear models (see Jędrzejczak, Kubacki (2016), Jędrzejczak, Kubacki (2017)) may be a starting point for more detailed comparisons between the Rao-Yu method, nonlinear models and econometric panel models.

Acknowledgement

We thank Mr. Adam Graboś from the Statistics Poland Computing Centre for providing us with the digital maps of borders of the Polish Regions, which were used for visualizations presented in this article.

REFERENCES

BIVAND, R., LEWIN-KOH, N., (2013). maptools: Tools for reading and handling spatial objects, R package version 0.8-25, http://CRAN.R-project.org/package=maptools.

BIVAND, R., PIRAS, G., (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics, Journal of Statistical Software, 63, No. 18, pp. 1–36, http://www.jstatsoft.org/v63/i18/.

CRESSIE, N. A .C., (1991). Small-area prediction of undercount using the general linear model, Proceedings of Statistic Symposium 90: Measurement and Improvement of Data Quality, Ottawa, Statistics Canada, pp. 93–105.

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 for 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.

GRIFFITH, D. A., PAELINCK, J. H. P., (2011). Non-standard Spatial Statistics and Spatial Econometrics, Advances in Geographic Information Science, Springer Berlin Heidelberg.

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, https://czasopisma.uni.lodz.pl/foe/article/view/755/660.

JĘDRZEJCZAK, A., KUBACKI, J., (2017). Analiza rozkładów dochodu rozporządzalnego według województw z uwzględnieniem czasu, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Nr 469, Taksonomia 29, Klasyfikacja i analiza danych – teoria i zastosowania, pp. 69–81, https://www.dbc.wroc.pl/dlibra/publication/41063/edition/37066/content?ref=desc.

KUBACKI, J., JĘDRZEJCZAK, A., (2016) Small Area Estimation of Income Under Spatial SAR Model, Statistics in Transition new series, September 2016, Vol. 17, No. 3, pp. 365–390, https://www.exeley.com/exeley/journals/statistics_in_transition/18/4/pdf/10.21307_stattrans-2017-009.pdf.

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.

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

MOLINA, I., MARHUENDA, Y., (2013). sae: Small Area Estimation, R package version 1.0-2, http://CRAN.R-project.org/package=sae.

MOLINA, I., MARHUENDA, Y., (2015). R package sae: Methodology – sae package vignette: https://cran.r-project.org/web/packages/sae/vignettes/sae_methodology.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.

PETRUCCI, A., SALVATI, N., (2006). Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment, Journal of Agricultural, Biological & Environmental Statistics, 11, No. 2, pp. 169–182, http://dx.doi.org/10.1198/108571106x110531.

PRASAD, N. G. N., RAO, J. N. K., (1990). The estimation of mean squared error of small area estimators. Journal of the American Statistical Association, 85, pp. 163–171, https://www.jstor.org/stable/2289539.

PRATESI, M., SALVATI, N., (2004). Spatial EBLUP in agricultural survey. An application based on census data, Working paper no. 256, Universitá di Pisa, Dipartimento di statistica e matematica applicata all'economia.

PRATESI, M., SALVATI, N., (2005). Small Area Estimation: The EBLUP Estimator with Autoregressive Random Area Effects, Working paper n. 261 Pubblicazioni del Dipartimento di statistica e matematica applicata all'economia.

PRATESI, M., SALVATI, N., (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects, Statistical Methods and Applications, 17, No. 1, pp. 113–141, http://dx.doi.org/10.1007/s10260-007-0061-9.

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, http://www.asasrms.org/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, (2016). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

SAEI, A., CHAMBERS, R., (2003). Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects, M03/15, Southampton Statistical Sciences Research Institute, http://eprints.soton.ac.uk/8165/.

SINGH, B. B., SHUKLA, G. K., KUNDU, D., (2005). Spatio-Temporal Models in Small Area Estimation, Survey Methodology, 31, No. 2, pp. 183–195.

WESTAT, (2007). WesVar® 4.3 User’s Guide.

APPENDIX

Simple R code illustrating the computations

```
# reading the libraries
library(RODBC)
library(sae2)
library(sae)
library(maptools)
library(spdep)

# obtaining the proximity matrix
region.poly <- readShapePoly("Polish_regions")
region.nb   <- poly2nb(region.poly)
W <- nb2mat(wojew.nb, style = "W")
# reading the data from Excel spreadsheet
channel <- odbcConnectExcel2007("Input.xlsx",sep="")
command <- paste("select * from [Sheet1$] order by region,year", sep="")
base <- sqlQuery(channel, command)
d <- cbind(base,desvar=(base[,3])^2)
# please note that position of variance (or standard error) variable in Input file may be different for particular case
# variable for number of domains
D <- 16
# variable for number of time periods
T <- 11
# formula for particular model - see for example Table 1
formula <- "D905_AVG ~ PRZECGOSP + PRZECEMER + PKB_PC"

# obtaining the Rao-Yu estimates
resultT.RY <- eblupRY(as.formula(formula), D, T, vardir =
diag((base[,3])^2),data=base, ids=base$region, method="REML")
# obtaining the decomposition of random effects for Rao-Yu model
```

```
resultT.RY_RE <- eblupRY_randeff_1d(formula, D, T, vardir = diag((base[,3])^2),
data=base, delta = resultT.RY$delta)
# obtaining the spatio-temporal estimates
resultST <- eblupSTFH(as.formula(formula), D, T, desvar, W, data=d)
# obtaining the MSE valus for spatio-temporal model using the parametric
bootstrap procedures
resultPBST <- pbmseSTFH(as.formula(formula), D, T, desvar, W, data=d)
# obtaining the decomposition of random effects for spatio-temporal model
resultST_RE <- eblupSTFH_randeff(as.formula(formula), D, T, vardir =
(base[,3])^2, W, data=d, rho1 = resultST$fit$estvarcomp[2,1], rho2 =
resultST$fit$estvarcomp[4,1], sigma21 = resultST$fit$estvarcomp[1,1], sigma22 =
resultST$fit$estvarcomp[3,1])
.
```

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 135–152, DOI 10.21307/stattrans-2019-038
Submitted – 04.03.2019; Paper ready for publication – 16.09.2019

THE IMPACT OF THE APPLIED TYPOLOGY ON THE STATISTICAL PICTURE OF POPULATION AGEING IN URBAN AREAS IN POLAND – A COMPARATIVE ANALYSIS

Tomasz Klimanek[1], **Sylwia Filas-Przybył**[2]

ABSTRACT

The aim of the paper is to review and compare the processes of population ageing in Polish urban areas. The study presents a novel approach to the problem, because in addition to measuring this phenomenon according to the National Official Register of the Territorial Division of the Country (TERYT) classification, it also measures population ageing according to the classification for urban areas (LAU 2 units) – Degree of Urbanisation (DEGURBA). Several traditional demographic measures for population ageing were applied, such as the median age, parent support ratio, ageing index, elderly dependency ratio, share of people aged 65 and older, and total dependency ratio. Also, Chu's alternative measure of population ageing accompanied by a dynamic version of ageing index was computed. The values of these indicators for 2016 were compared with those for 2010. The authors carried out a more detailed analysis of the differences between the ageing of populations in urban areas according to the degree of urbanization (DEGURBA), and compared the outcome with the results of the TERYT-based measurement (the traditional administrative territorial division). The comparison of the outcomes of both the above-mentioned ways of measuring the phenomenon of population ageing showed discrepancies, namely the ageing process measured according to the DEGURBA typology proved to be less intensive than the same process assessed according to the TERYT typology. This indicates that there are differences between the statistical pictures of population ageing in urban areas depending on whether demographical and morfological aspects are taken into consideration or not.

Key words: TERYT classification, DEGURBA typology, urban statistics, urban ageing.

1. Introduction

One of the most important demographic problems encountered by every country in the world is population ageing. It is commonly defined as the increasing share of older persons in the population. According to *World Population*

[1] Statistical Office in Poznań. E-mail: t.klimanek@stat.gov.pl
[2] Statistical Office in Poznań. E-mail: s.filas@stat.gov.pl

Prospects: the 2017 Revision "the global population aged 60 years or over numbered 962 million in 2017, was more than twice as large as in 1980 when there were 382 million older persons worldwide. The number of older persons is expected to double again by 2050, when it is projected to reach nearly 2.1 billion". When one takes into account the rural-urban perspective it is worth pointing out that "the number of older persons is growing faster in urban areas than in rural areas. At the global level between 2000 and 2015, the number of people aged 60 years or over increased by 68% in urban areas, compared to a 25% increase in rural areas. As a result, older persons are increasingly concentrated in urban areas. In 2015, 58% of the world's people aged 60 years or over resided in urban areas, up from 51% in 2000. Those aged 80 years or over are even more." (United Nations, 2017).

According to *Population Projection 2014-2050* (GUS, 2014) the share of population aged 65 years and over in Poland will amount to 26.3% in urban areas in 2035 compared to 22% in rural areas. However, there will be a lot of regional variation in population ageing.

The urban perspective on different phenomena is of special importance for the Centre for Urban Statistics – a unit dealing with statistics related to cities, towns and urban areas in the Statistical Office in Poznań. It was established as part of the specialization strategy, implemented in Polish official statistics at the start of 2009. The main idea of the specialization strategy was to make each regional office responsible for conducting tasks for the whole country within specific fields. In other words, the regional offices were no longer limited to collecting data from a single province. Since its creation, the tasks of the Centre for Urban Statistics have focused on initiating surveys and formulating new methodological proposals for the statistical observation of cities and towns as well as conducting methodological studies aimed at delimiting and surveying areas that do not overlap with the country's administrative division.

One of the most up-to-date challenges for official statistics is to call for a more flexible approach to the perception of the city as an important spatial element, especially now that the shortcomings of spatial analyses based solely on units of administrative division (TERYT system) or statistical division (NUTS classification) can no longer be ignored. In this context, the grid concept (a network of grid squares, with a certain spatial resolution, e.g. 500x500 m, or 1x1 km) is especially relevant, making it possible to depart from the fixed administrative division and analyse phenomena both within urban structures (Dąbrowski et al., 2016; Filas-Przybył et al., 2016) and across administrative city borders – e.g. urban functional zones. For example, the 1x1 km grid network serves as the basis for the European classification of administrative units, which is used to determine the degree of urbanization – DEGURBA (Dijkstra and Poelman, 2014).

The purpose of the article is to compare the statistical picture of the ageing process of the urban population in Poland using two typologies. One of them is based on the definition of a town in the Act of 29 August 2003 on official names of localities and physiographic objects. The second typology used in this study is the European classification of administrative units based on the degree of urbanization – DEGURBA.

In the second part of the article both TERYT and DEGURBA classification were described in more detail. The definition of town/city used in the TERYT classification and the description of DEGURSBA's densely, intermediate and thinly populated areas were introduced. The aim of the next part of the paper was to introduce, present formulas and to discuss some properties of population ageing measures used in the research. These were: median age, parent support ratio, ageing index, elderly dependency ratio, proportion of population aged 65 and over (% of total), total dependency ratio. Additionally, we computed Chu's alternative measure of population ageing (Nath and Islam, 2016) and dynamic version of ageing index proposed by Długosz (Długosz, 1998). The selection of these demographic measures was motivated partly by previous studies on this topic (GUS, 2015). The comparison of the statistical picture of population ageing from TERYT and DEGURBA perspective was presented in the fourth part of the paper. In order to track changes over time, the data for 2016 are compared with those for 2010 (assuming that in 2010 the DEGURBA classification was the same as in 2016), which, in both cases, come from the Local Data Bank maintained by Statistics Poland. Also, some discussion describing the significance of main findings in light of what was already known about the population ageing was provided at the end of this part of the paper. In the conclusions we included not only the main findings of the research but also possible directions of future works.

2. TERYT and DEGURBA typologies – comparison

Official statistics for cross-classification domains at different levels of territorial aggregation are calculated and presented in Poland on the basis of the TERYT register. It consists of four components: TERC – identifiers and names of territorial units, SIMC – identifiers and names of localities, BREC – statistical regions and enumeration areas, and NOBC – address details of streets, properties, buildings and dwellings. The TERC system contains identifiers and names of units that constitute the three-tier territorial division of the country: province, district, commune (municipality). The territorial code of every commune consists of 7 digits. The first two denote the province where the commune is located, the first four denote the district, while the last digit represents the commune type. Codes, commune types and their descriptions in the TERYT system are presented in detail on the Statistics Poland website https://bdl.stat.gov.pl/BDL/metadane/teryt/rodzaj.

In official statistics the town/city is defined as a unit of territorial division which has been granted town status by a municipal charter (cf. The Act on official names…, 2003 Article 2, Item 3). According to the regulation, this includes urban communes and towns in urban-rural communes. Analogically, rural communes and rural parts of urban-rural communes are classified as villages. However, spatial analyses based solely on the legal/administrative classification of territorial units in the TERYT register often result in a distorted picture of phenomena and processes taking place inside administrative units or between neighbouring units. What is needed, then, is an analytical approach based on the concept of 1x1 km grid and the DEGURBA classification.

DEGURBA – the European classification, based on the degree of urbanization, was first implemented in 1991 in order to characterize areas inhabited by respondents of official statistical surveys. This original DEGURBA typology distinguished between three kinds of areas: densely populated, intermediate and thinly populated (Eurostat, 2011). Their definition was based on population size, population density and geographical contiguity of LAU2 units. However, even at that time it became obvious that the approach based on LAU2 units, whose area varied considerably across EU countries, leads to distorted results and limits the scope of comparative analyses between EU countries.

In 2010 Eurostat introduced a new regional typology, which originated from a method developed by OECD (Berezzi et al., 2011). The method was based on grid square cells of 1 km^2, which in combination with the results of another Urban Audit project, provided an opportunity to revise the definition, borders and number of cities according to the idea of a densely populated area used in the degree of urbanization classification.

The new typology of areas (Dijkstra and Poelman, 2014) based on their degree of urbanization introduces the following classification of LAU2 statistical units (the brackets on the left contain DEGURBA codes):

(1) densely populated area: (alternative name: city/large urban area)
- at least 50% lives in high-density clusters;

(2) intermediate area (alternative name: towns and suburbs/small urban area)
- less than 50% of the population lives in rural grid cells; and
- less than 50% lives in high-density clusters;

(3) thinly populated area (alternative name: rural area):
- more than 50% of the population lives in rural grid cells.

In the above, the following definitions are used:

- Rural grid cells: grid cells outside urban clusters.
- Urban clusters: clusters of contiguous grid cells of 1 km^2 with a density of at least 300 inhabitants per km^2 and a minimum population of 5000.
- High-density cluster: contiguous grid cells of 1 km^2 with a density of at least 1500 inhabitants per km^2 and a minimum population of 50000 (alternative names: urban centre or city centre).

Details of the methodology of establishing the DEGURBA typology can be found in the publications (Dijkstra and Poelman, 2014; Eurostat, 2011).

Application of DEGURBA classification requires an appropriate statistical and IT infrastructure to ensure regular updating of information for the delimitation of densely-populated, intermediate and thinly populated areas. It was assumed that information about changes in LAU borders would be updated annually (Eurostat, 2011). A more challenging problem is how to update the spatial distribution of the population. Censuses, which are the main source of data required for the DEGURBA classification, are conducted every 5 or 10 years. The choropleth maps below show the classification of LAUs (communes) according to the TERYT register and the DEGURBA typology.

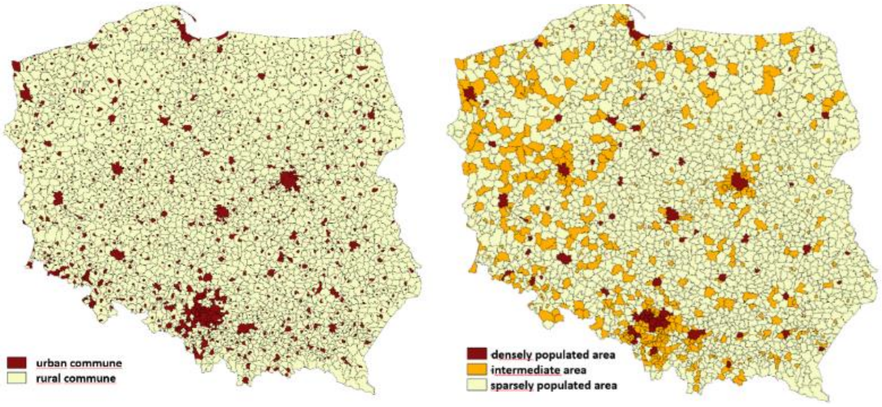


Figure 1. Classification of LAUs (communes) according to the TERYT register (left) and the DEGURBA typology based on data from Census 2011 (right)

A simple comparison of the number of towns according to the definition used in the TERYT register and the number of areas classified as urban according to the DEGURBA typology (densely populated and intermediate areas, codes 1 and 2, respectively) indicates significant discrepancies. There are 930 towns in the TERYT register, compared to only 601 urban units according to the DEGUBRA typology.

3. Population ageing measures

The problem of population ageing in Poland, its regional variation (Kurek, 2008, Stańczak and Szałtys, 2016; Podogrodzka 2016, Majdzińska, 2017), and the way it affects Polish towns (Kurek, 2008, Trzpiot and Ojrzyńska, 2014; Trzpiot and Szołtysek 2015; GUS, 2018) is becoming increasingly relevant in public awareness and discourse. The advancement in population ageing is measured by means of various indicators: either traditional ones, based on the threshold of population ageing and relations between basic age groups, or less common, alternative measures, which take into account changing mortality rates and life expectancy (Abramowska-Kmon, 2011).

In this study, the statistical picture of population ageing in Polish towns according to the TERYT and DEGURBA typologies was compared on the basis of median age and the following demographic measures:

1. Parent support ratio

$$PSR = \frac{L_{85+}}{L_{50-64}} * C \tag{1}$$

where:

PSR – parent suport ratio
L_{85+} – number of people aged 85 and over
L_{50-64} – number of people aged 50-64
C – constant (=100)

2. Ageing index

$$AI = \frac{L_{65+}}{L_{0-14}} * C \tag{2}$$

where:

AI – ageing index
L_{65+} – number of people aged 65 and over
L_{0-14} – number of people aged 0–14
C – constant (=100)

By 2045 *AI* is predicted to exceed 100 in all European countries, which will mean the elderly population will outnumber the youngest one (Kurek, 2008).

3. Elderly dependency ratio

$$EDR = \frac{L_{65+}}{L_{15-64}} * C \tag{3}$$

where:

EDR – elderly dependency ratio
L_{65+} – number of people aged 65 and over
L_{15-64} – number of people aged 15-64
C – constant (=100)

4. Proportion of elderly people

$$PEP = \frac{L_{65+}}{L} * C \tag{4}$$

where:

PEP – proportion of elderly people
L_{65+} – number of people aged 65 and over
L – total population
C – constant (=100)

According to UN data (Abramowska-Kmon, 2011), a population is considered young when the share of people aged 65 and over is below 4%. A population is classified as mature when this share ranges from 4% to 7%, and is regarded as old when it exceeds 7%[3].

5. Total dependency ratio

$$TDR = \frac{L_{0-14} + L_{65+}}{L_{15-64}} * C \tag{5}$$

where:

TDR – total dependency ratio
L_{0-14} – number of people aged 0-14
L_{65+} – number of people aged 65 and over
L_{15-64} – number of people aged 15-64
C – constant (=100)

[3] However, the UN scale seems to have now only historical meaning, as noted by Abramowska-Kmon.

6. Chu's indices based on general formula (Chu, 1997)

$$I_\alpha^P(G:z) = \frac{1}{(\omega-z)^{\alpha-1}} \int_{G(z)}^{1} [G^{-1}(p) - z]^{\alpha-1}\, dp \qquad (6)$$

where:

$G(a)$ – cumulative distribution function for age *a*

$G^{-1}(p)$ – the age of population corresponding to cumulative probability *p* value

z – critical value of peak age, here z = 65

ω – upper limit of age distribution, here ω = 90

For $\alpha = 1$ we have the so-called conventional peak ageing index

$$I_1^P(G:z) = \int_{G(z)}^{1} dp \qquad (7)$$

For $\alpha = 2$ we have the so-called aged gap ageing index

$$I_2^P(G:z) = \frac{1}{(\omega-z)} \int_{G(z)}^{1} [G^{-1}(p) - z]\, dp \qquad (8)$$

And for $\alpha = 3$ we have the so-called age distribution sensitive ageing index

$$I_3^P(G:z) = \frac{1}{(\omega-z)^2} \int_{G(z)}^{1} [G^{-1}(p) - z]^2\, dp \qquad (9)$$

If we have, for example, data in the form of 5-year age groups, the discrete case general formula is the following:

$$I_\alpha^P(G:z) = \frac{1}{(\omega-z)^{\alpha-1}} \sum_j \left(x_j - z\right)^{\alpha-1} p_j \qquad (10)$$

where:

x_j – is the mid-point of *j-th* age group

z – critical value of peak age, here z = 65. Age groups are then 65–69, 70–74, 75–79, 80–84, 85+. For 85+ age group ω = 90 was taken arbitrarily as the upper bound.

From the above mentioned indices we will use only I_2^P. The aged gap ageing index is the weighted proportion of old with weights which are differences between the corresponding age of these old and the critical age z.

7. Długosz's dynamic version of ageing index (Długosz, 1998; Długosz 2003)

$$W_{SD} = [U(0-14)_t - U(0-14)_{t+n}] + [U(>65)_{t+n} - U(>65)_t] \qquad (11)$$

where:

W_{SD} – ageing index

$U(0-14)_t$ – share of population aged 0–14 in 2010

$U(0-14)_{t+n}$ – share of population aged 0–14 in 2016

$U(>65)_t$ – share of population aged 65 and over in 2010

$U(>65)_{t+n}$ – share of population aged 65 and over in 2016

Długosz's dynamic version of ageing index indicates the differences in the percentage share of the youngest and the oldest group in the period studied. It could be useful for building the typology of population based on the mutual relation between the changes in the share of population in the age groups of 0–14 and >65. There are eight theoretical types (A-H) of population ageing (for example type A denotes population becoming younger due to the domination of an increase in the share of population aged 0–14 over the increase in the share of population aged>65, while type H denotes population ageing due to the domination of the increase in the share of population aged >65 over the increase in the share of population aged 0–14).

4. Statistical picture of population ageing in Polish towns – regional comparison and discussion

The impact of the typology used (TERYT vs DEGURBA) on the statistical picture of population ageing in Polish towns was evaluated taking into account the following assumptions:

- as regards the TERYT classification, the analysis included urban communes and towns in urban-rural communes (TERYT code = 1 or 4);
- as regards the DEGURBA classification, the analysis included urban areas (communes) characterized as densely populated or intermediate areas (DEGURBA code = 1 or 2);
- the selected territorial units (towns according to TERYT, urban areas according to DEGURBA) were characterized in terms of median age and selected measures of population ageing (parent support ratio, ageing index, elderly dependency ratio, proportion of elderly people, total dependency ratio, Chu's measure of population ageing and dynamic version of ageing index);
- analysis was based on official statistics from the Local Data Bank maintained by Statistics Poland;
- the analysis of population ageing was conducted for 2010 and 2016;
- the same DEGURBA typology was used for both reference years based on population counts established in the last census (NSP 2011) and LAUs updated in 2016.

Table 1. Classification of communes according to TERYT and DEGURBA in 2016

	Urban commune	Rural commune	Urban-rural commune	Total
Densely populated area	74	0	0	74
Intermediate area	207	84	236	527
Thinly populated area	22	1475	380	1877
Total	303	1559	616	2478

Source: Own elaboration.

According to the TERYT classification, there were 919 towns out of a total of 2478 communes, 303 of which were urban communes, while 616 were towns located in urban-rural communes. The number of urban areas according to the DEGURBA classification amounted to 601 (74 LAU2 were classified as densely populated areas and 527 were classified as intermediate areas).

Table 2. Comparison of population ageing measures in towns and urban areas for 2010 and 2016

Demographic measure	YEAR			
	2010		2016	
	Urban areas (DEGURBA)	Towns (TERYT)	Urban areas (DEGURBA)	Towns (TERYT)
Median age	39.0	39.3	41.1	41.5
PSR – formula (1)	5.8	5.7	9.4	9.4
AI – formula (2)	97.0	100.4	120.3	125.7
EDR – formula (3)	19.1	19.3	25.5	26.1
PEP – formula (4)	13.8	13.9	17.4	17.8
TDR – formula (5)	38.8	38.5	46.7	46.9
I_2^P – formula (10)	0.049	0.050	0.059	0.060
W_{SD} – formula (11)	x	x	0.033	0.036

Source: Own elaboration.

Between the two reference years all measures of demographic ageing increased. This is true for urban areas defined according to the DEGURBA classification and for towns listed in the TERYT register. In 2016 half of the population living in urban areas was older than 39 years. The change in median age reflected by the index calculated in reference to the base year 2010 was 105.4 and was lower compared to that for towns in the TERYT register, where median age in 2016 was equal to 41.5 years, which is over 2 years more than in 2010.

Parent support ratios for both reference years, regardless of the classification, are below 10%. This means that the generation of parents (people aged 85 and over), which requires direct support and care, accounted for less than 10% of the subsequent generation of “children” (people aged 50-64). In this case, the values of the *PSR* measure for urban areas are almost the same as for towns. One particularly worrying trend is the intensity of change in the *PSR* indicator, which is the highest of all measures of population ageing presented. In 2016 *PSR* for urban areas increased by 62% relative to 2010, in the case of towns, the increase was even higher – as much as 65%.

The ageing index also underwent significant changes: 24% in the case of urban areas and 25% in the case of towns. Moreover, between the 2 reference years the indicator exceeded the level of 100, which means that both in urban areas and in towns, the population aged 65 and over outnumbered the youngest age group (0–14 years).

The second most dynamically growing measure of population ageing was the elderly dependency ratio, which in the case of urban areas rose by 33.5% between 2010 and 2016 and by 35.2% in the case of towns. This means the growing burden of supporting the post-working age population by people of the working age.

The proportion of elderly people in 2010 already amounted to the level of almost 14%, only to sky rocket in 2016 to 17.4% in urban areas and 17.8% in towns. In terms of the terminology adopted by the UN, this means that the urban population in Poland in 2010 could already be classified as old (above the threshold of 7%).

Despite an increase of more than 20% in over the reference period, the total dependency ratio in towns was more or less similar for both classifications and in 2016 was equal to 46.7 and 46.9 for urban areas and towns respectively. Interestingly, in 2010 it was slightly lower in towns (38.5 compared to 38.8 in urban areas).

Chu's alternative measure of ageing – the aged gap ageing index – also increased significantly between 2016 and 2010. Defined as a normalized sum of the proportion of old groups in the population (over 65), its dynamics (20% between 2010 and 2016) shows a significant increase of old groups in the population. It is also a quite worrying trend as far as the intensity of change is concerned.

The values of the dynamic version of the ageing index for both TERYT and DEGURBA classification are similar. They indicate that the process of ageing in Polish towns could be synthetically described as representing type H, i.e. population ageing due to the domination of the increase in the share of population aged >65 over the increase in the share of population aged 0–14.

Table 3. Measures of population ageing in urban areas by commune type in 2016

	Urban areas in total (DEGURBA)	Urban communes (densely populated areas)	Urban-rural communes (intermediate areas)	Rural communes (intermediate areas)
Median age	41.1	41.8	39.8	38.5
PSR – formula (1)	9.4	10.6	7.9	7.7
AI – formula (2)	120.3	133.9	99.3	78.4
EDR – formula (3)	25.5	27.5	22.1	19.5
PEP – formula (4)	17.4	18.6	15.3	13.5
TDR – formula (5)	46.7	48.1	44.3	44.3
I_2^P – formula (10)	0.059	0.064	0.050	0.045
W_{SD} – formula (11)	0.033	0.031	0.036	0.021

Source: Own elaboration.

The results presented in the tables above clearly indicate how strongly the levels of population ageing measures in urban areas are affected by the inclusion of areas with intermediate population density according to the DEGURBA classification. The median age, which in densely populated areas amounted to 41.8 years, was lower by 2 years in urban-rural areas classified as intermediate, and in the case of rural communes classified as intermediate areas, was lower by as much as 3.3 years. In many cases these are the communes which have been experiencing an intensive influx of migration for permanent residence from neighbouring big cities. These migration flows consist mainly of young families with small children, who could expect to buy a flat at much lower prices in nearby communes surrounding the big city than in the city itself. This is particularly evident in the case of the other measures shown in the table above. The parent support ratio, equal to 10.4 in densely populated areas, did not exceed 8 in areas with intermediate population density. This difference is particularly large in the case of the ageing index, which is over 130 in densely populated areas but as low as 78.4 in rural areas classified as intermediate. Similar relations can be observed in the case of the elderly dependency ratio, the proportion of elderly population, the total dependency ratio, Chu's ageing index and the dynamic version of the ageing index.

The ageing process shows high regional differentiation. Below are presented the choropleths for selected measures of population ageing in Polish towns on the level of NUTS2 (provinces). The arrangement of the maps for every presented measure of population ageing (the choice of the median, parent support ratio and Chu's ageing index is due to the limitation of space in the paper) is as follows: first, regional differentiation of the selected measure of population ageing in Polish towns is compared between the general TERYT and DEGURBA approach and also for 2010 and 2016. Then, DEGURBA approach is analysed in more detail by presenting separate choropleths for urban areas from the perspective of different commune types (DEGURBA code 1, DEGURBA code 2 for urban-rural communes, DEGURBA code 2 for rural communes). This arrangement also includes a comparison of spatial distribution for 2010 and 2016.

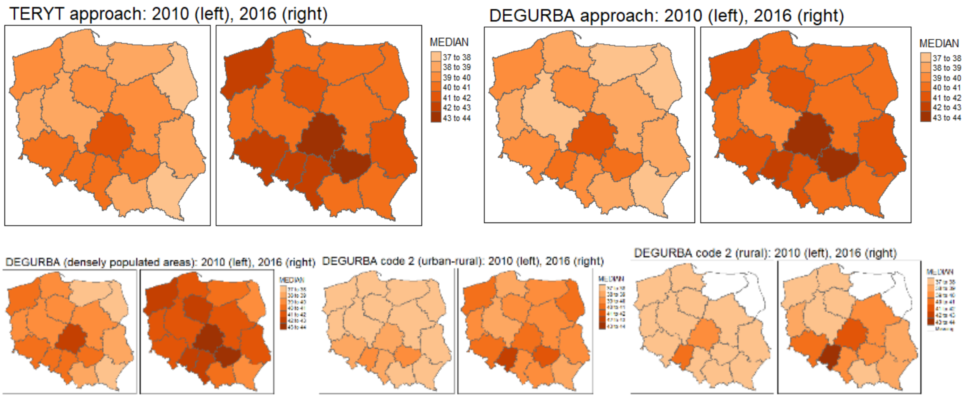

TERYT approach: 2010 (left), 2016 (right)
DEGURBA approach: 2010 (left), 2016 (right)
MEDIAN
37 to 38
38 to 39
39 to 40
40 to 41
41 to 42
42 to 43
43 to 44
DEGURBA (densely populated areas): 2010 (left), 2016 (right)
DEGURBA code 2 (urban-rural): 2010 (left), 2016 (right)
DEGURBA code 2 (rural): 2010 (left), 2016 (right)
Missing

Figure 2. Median age of population in Polish towns according to TERYT and DEGURBA typologies

Source: Own work.

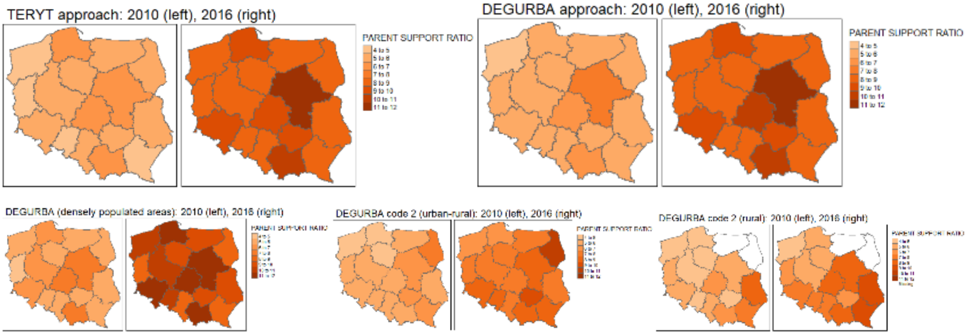

TERYT approach: 2010 (left), 2016 (right)
DEGURBA approach: 2010 (left), 2016 (right)
PARENT SUPPORT RATIO
4 to 5
5 to 6
6 to 7
7 to 8
8 to 9
9 to 10
10 to 11
11 to 12
DEGURBA (densely populated areas): 2010 (left), 2016 (right)
DEGURBA code 2 (urban-rural): 2010 (left), 2016 (right)
DEGURBA code 2 (rural): 2010 (left), 2016 (right)

Figure 3. Parent support ratio of population in Polish towns according to TERYT and DEGURBA typologies

Source: Own work.

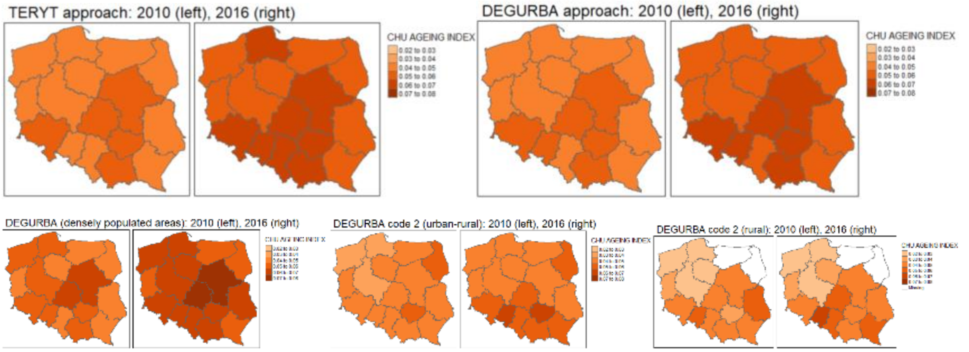

TERYT approach: 2010 (left), 2016 (right)
DEGURBA approach: 2010 (left), 2016 (right)
CHU AGEING INDEX
0.02 to 0.03
0.03 to 0.04
0.04 to 0.05
0.05 to 0.06
0.06 to 0.07
0.07 to 0.08
DEGURBA (densely populated areas): 2010 (left), 2016 (right)
DEGURBA code 2 (urban-rural): 2010 (left), 2016 (right)
DEGURBA code 2 (rural): 2010 (left), 2016 (right)

Figure 4. Chu's ageing index of population in Polish towns according to TERYT and DEGURBA typologies

Source: Own work.

The results presented in the choropleths given above (Figure 2 - Figure 4) exhibit a number of patterns. First, there is quite a significant spatial differentiation in the value of calculated demographic ageing indicators for the urban population. The cities of the central and southern provinces of Poland are the most affected by demographic ageing. In the case of the median age, high values in 2016 were observed for the urban population in Zachodniopomorskie province. Secondly, we can clearly see the acceleration of the demographic ageing process, especially in the oldest age groups between 2010 and 2016, which is reflected primarily by a huge increase in parent support ratio and Chu's index. The decomposition of urban areas according to the DEGURBA classification shows that densely populated areas (DEGURBA code = 1) are most severely affected by demographic ageing. These processes also dynamically occur in urban areas, which are made up of urban-rural and rural communes, but their intensity is significantly lower.

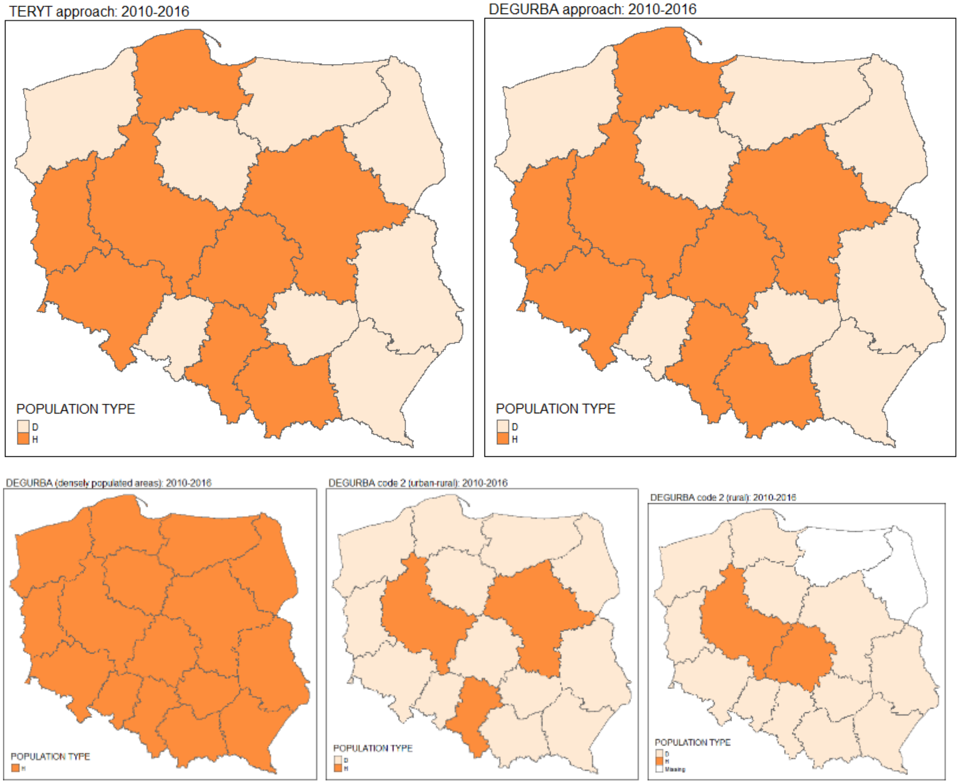

TERYT approach: 2010-2016
POPULATION TYPE
D
H
DEGURBA approach: 2010-2016
POPULATION TYPE
D
H
DEGURBA (densely populated areas): 2010-2016
POPULATION TYPE
H
DEGURBA code 2 (urban-rural): 2010-2016
POPULATION TYPE
D
H
DEGURBA code 2 (rural): 2010-2016
POPULATION TYPE
D
H
Missing

Figure 5. Population types according to Długosz’s dynamic version of the ageing index

Source: Own work.

It is also worth referring to the types of population that can be obtained using Długosz's dynamic index of ageing. It turns out that regardless of whether we use the TERYT classification or the DEGURBA classification, in the period 2010-2016, the population living in cities/urban areas for a given province is of type D, i.e. the population is getting younger owing to the fact that the decrease in the share of the population aged >65 is greater than the decline in the percentage of the population aged 0-14 or of type H, i.e. population ageing is due to the increase in the population aged 0-14 (the same membership in both classifications). In the case of densely populated areas there is always an H-type population, for urban areas which are formed by urban-rural communes in three provinces, there is a H-type population (Wielkopolskie, Śląskie and Mazowieckie provinces), with respect to the rest, we are dealing with a population of type D.

In the case of rural communes in two provinces, we have an H-type population (Wielkopolskie and Łódź provinces); as regards the rest, we have a population of type D.

One of the first comparisons of the two classifications was conducted in Poland by Filas-Przybył (2012), who analysed such characteristics as province area per one town/urban area, area of towns/urban areas per one town/urban area, urban population per one town/urban area, the share of urban population in the total population of the province. The study revealed that the largest number of towns according to the TERYT classification can be found in Wielkopolskie province (109); according to the DEGURBA typology, Śląskie province is the most urbanized region of the country (92 urban units). The biggest urban density was found to exist in Śląskie province, where the proportional share of the province area per one town was equal to 173.7 km^2 (TERYT) and 134.1 km^2 (DEGURBA). It turned out that regardless of the classification used, Śląskie province was also the most urbanized province. This was reflected by the size of an average town both in terms of area and population, which was almost twice as big as the average town size in Poland (TERYT). However, according to the DEGURBA typology, the largest average town in terms of size was found in Wielkopolskie province, while in terms of population – in Zachodniopomorskie province. Regardless of the classification, the smallest average town in terms of area, which was a third of the size of the average town in Poland according to the DEGURBA typology, was found in Warmińsko-Mazurskie province.

In the light of current research on the ageing of the urban population, taking into account the territorial aspect and the results presented in this paper, it seems that the way of looking at the towns/urban areas presented in this article should be developed. On the one hand, we have an approach where urbanity essentially determines only the formal status, on the other hand, we have the DEGURBA classification, coming from population density and the fact of population clustering (kilometre grid clusters), which certainly better reflects urbanization processes than the legal acts or administrative decisions assigning the formal city status.

5. Conclusions

The article provides a comparison of the statistical picture of the ageing process of the urban population of Polish towns according to two typologies – TERYT and DEGURBA – based on selected measures of population ageing for 2010 and 2016. In 2016 all these measures in urban areas identified according to the DEGURBA classification were found to be in general slightly lower than those obtained for the population of towns listed in the TERYT register. This represents a less advanced stage of population ageing in urban areas delimited not on the basis of legal or administrative city/town status but based on a more objective criterion of the degree of urbanization. Such findings, among other things, are the result of the inclusion of some rural communes in areas classified as urban according to the DEGUBRA typology. It is worth noting that in recent years these communes have experienced an intensive influx of migration for permanent residence from nearby big cities. Moreover, these migration flows consist mainly of young families with small children, who decide to move out of flats rented in big cities or shared with their parents. These migration flows have a considerable influence of relations between basic age groups that are the basis for calculating classical measures of population ageing. To a certain degree, the same can be true in the case of urban-rural communes classified as urban areas according to DEGURBA. Of the 616 communes of this type, 236 were classified as urban areas, while the remaining 380 – as thinly populated areas.

It should also be noted that the traditional typology of towns according to the TERYT register makes it more difficult to conduct an in-depth and multidimensional analysis and evaluation of socio-economic phenomena that take place in space. It is therefore recommended that this traditional typology should be complemented by the approach described in this article, which is based on the 1x1 km grid and the DEGURBA classification, making it possible to increase the scope and depth of analysis, also as regards population ageing processes (see Table 4). Given the dynamic changes currently taking place in the demographic structure, it is necessary to apply increasingly sophisticated analytical methods, which rely on a modern statistical infrastructure supported by spatial (geocoded) statistics.

Table 4. Selected aspects of applying TERYT and DEGURBA classifications for the analysis of different phenomena

Aspect	TERYT	DEGURBA	Both TERYT and DEGURBA
Urbanity conceptualisation	Only formal	Selected demographic (population size) and morphological (population density) aspects	Complex formal, demographic and morphological aspects
Possibility of decomposition	Into urban, urban – rural and rural communes, by town size	Into densely populated areas, intermediate areas, thinly populated areas	Multidimensional analysis (by TERYT communes' categories, town size, DEGURBA codes)
Relevance for units facing demographic changes of high dynamics	Poor	Better	Better
Quality of statistical information infrastructure needed	Low	High	High

Source: Own work.

Obviously, the comparative analysis of the impact of the typology used for the statistical picture of population ageing described in this article includes certain limitations and simplifications. It seems, however, that the results are promising and indicate directions for future research, including the extension of the scope of such studies to take into account how the migration of young generation for permanent residence affects the measures of population ageing, and the application of methods of multivariate statistics for identifying similar patterns of population ageing in towns.

REFERENCES

ABRAMOWSKA-KMON, A., (2011). O nowych miarach zaawansowania procesu starzenia się ludności, Studia Demograficzne, 1/159, pp. 3–19.

BREZZI, M., DIJKSTRA, L., RUIZ, V., (2011). OECD Extended Regional Typology: The Economic Performance of Remote Rural Regions, OECD Regional Development Working Papers, 2011/06, OECD Publishing, Paris.

CHU, C. Y. C. (1997). Age-distribution dynamics and aging indexes. Demography, 34(4), pp. 551–563.

DĄBROWSKI, A., FILAS-PRZYBYŁ, S., PAWLIKOWSKI, D., (2016). Identification of specific areas within provincial capital cities and their functional areas in terms of the demographic and economic situation of their inhabitants using GIS-based spatial analysis,
Available at: http://scorus.org/index.php/conferences/2016-2/scorus-conference-in-lisbon-portugal> [Accessed 20 November 2018].

DIJKSTRA, L., POELMAN, H., (2014). A Harmonised Definition of Cities and Rural Areas: the New Degree of Urbanisation, Regional Working Paper, WP 01/2014, European Commission,
Available at: https://ec.europa.eu/regional_policy/sources/docgener/work/2014_01_new_urban.pdf> [Accessed 25 November 2018].

DŁUGOSZ, Z., (1998). Próba określenia zmian starości demograficznej Polski w ujęciu przestrzennym, Wiadomości Statystyczne, No. 3, pp. 15–25.

DŁUGOSZ, Z., (2003). The level and dynamics of population ageing process on the example of demographic situation in Europe, Bulletin of Geography (socio-economic series), No. 2, Toruń: Nicolaus Copernicus University Press, pp. 5–15.

FILAS-PRZYBYŁ, S., (2012). Nowa metodologia klasyfikowania jednostek przestrzennych oparta na stopniu urbanizacji. Unpublished graduation work.

FILAS-PRZYBYŁ, S., KLIMANEK, T., KRUSZKA, K., STACHOWIAK, D., (2016). Identyfikacja obszarów specjalnych wewnątrz miast wojewódzkich – na przykładzie miasta Poznania.
Available at: https://www.arcanagis.pl/identyfikacja-obszarow-specjalnych-wewnatrz-miast-wojewodzkich-na-przykladzie-miasta-poznania> [Accessed 20 November 2018].

KUREK, S., (2008). Typologia starzenia się ludności Polski w ujęciu przestrzennym, Wydawnictwo Naukowe Akademii Pedagogicznej, Prace monograficzne nr 497, Kraków.

MAJDZIŃSKA, A., (2017). Zróżnicowanie terytorialne starzenia się ludności Polski, Acta Universitatis Lodziensis Folia Oeconomica, 5(331), pp. 73–90.

NATH, D., ISLAM, M. D., (2010). New Indices: An Application of Measuring the Aging Process of Some Asian Countries with Special Reference to Bangladesh. Journal of Population Ageing, pp. 23–39.

PODOGRODZKA, M., (2016). Starzenie się ludności Polski w przekroju regionalnym, Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, No. 290, pp. 83–94.

STAŃCZAK, J., SZAŁTYS, D., (2016). Regionalne zróżnicowanie procesu starzenia się ludności Polski w latach 1990–2015 oraz w perspektywie do 2040 roku,
Available at:
https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5468/28/1/1/regionalne_zroznicowanie_procesu_starzenia_sie_ludnosci.pdf> [Accessed 23 December 2018].

TRZPIOT, G., OJRZYŃSKA, A., (2014). Analiza ryzyka starzenia demograficznego wybranych miast w Polsce, Studia Ekonomiczne, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, No. 178, pp. 235–249.

TRZPIOT, G., SZOŁTYSEK, J., (2015). Przemiany demograficzne a mobilność mieszkańców miast, Studia Ekonomiczne, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, No. 223, pp. 121–139.

EUROSTAT, (2011). Correspondence table Degree of Urbanisation (DEGURBA) – Local Administrative Units, Methodological notes – The New Degree of Urbanisation,
Available at: http://ec.europa.eu/eurostat/ramon/miscellaneous/index.cfm?TargetUrl=DSP_DEGURBA> [Accessed 21 November 2018].

GUS, (2014). Prognoza ludności na lata 2014–2050,
Availableat at:
http://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5469/1/5/1/prognoza_ludnosci_na_lata____2014_-_2050.pdf> [Accessed 21 November 2018].

GUS, (2015). Identyfikacja obszarów specjalnych wewnątrz miast wojewódzkich oraz na ich obszarach funkcjonalnych uwzględniających sytuację demograficzną i ekonomiczną ich mieszkańców na podstawie analiz przestrzennych z wykorzystaniem Geographic Information System (GIS),
Available at: http://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultstronaopisowa/5850/1/1/raport_obszary_specjalne_gis_1.pdf> [Accessed 20 November 2018].

GUS, (2018). Miasta w liczbach 2016,
Available at:
https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5499/3/8/1/miasta_w_liczbach_2016.pdf> [Accessed 29 November 2018].

Ustawa z dnia 29 sierpnia 2003 r. o urzędowych nazwach miejscowości i obiektów fizjograficznych, Dz.U. 2003 nr 166 poz. 1612.

United Nations, Department of Economic and Social Affairs, Population Division, (2017). World Population Prospects: The 2017 Revision. New York: United Nations. ST/ESA/SER.A/399.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 153–166, DOI 10.21307/stattrans-2019-039
Submitted – 04.03.2019; Paper ready for publication – 27.09.2019

ON PERMUTATION LOCATION–SCALE TESTS

Dominika Polko-Zając [1]

ABSTRACT

Statisticians are constantly looking for methods of statistical inference that would be both effective and would require meeting as few assumptions as possible. Permutation tests seem to fit here, as using them makes it possible to perform statistical inference in situations where classical parametric tests do not work. Permutation tests appear to be comparably powerful to parametric tests, but require meeting fewer assumptions, e.g. regarding the size of the sample or the from of distribution of the tested variable in a population. The presented tests make it possible to verify the overall hypothesis about the identity of both location and scale parameters in the studied populations. In literature, the Lepage test and the Cucconi test are most often referred to in this context. The paper considers various forms of test statistics, and presents a simulation study carried out to determine the size and power of the tests under normality. As the study demonstrated, the advantage of the proposed method is that it can be applied to small-size samples. A nonparametric, complex procedure was used to assess the overall ASL (achieved significance level) value by applying the permutation principle. For comparative purposes, the results for the permutation Lepage test and the permutation Cucconi test are also presented.

Key words: permutation tests, comparing populations, test power, the Lepage test, the Cucconi test.

1. Introduction

Comparing populations most frequently refers to a comparison of the characteristics of these populations. If it is assumed that population distributions differ only in the central tendency, there are various parametric and nonparametric tests to verify this hypothesis. Many authors undertake to study the power and size of tests for the significance of differences between means or medians in two or more populations, using for this purpose simulation methods based on bootstrap or permutation tests (Janssen and Pauls, 2005; Chang and Pal, 2008; Kończak, 2016; Anderson et al., 2017). The problem of comparing variances in populations is also common in research. For example, comparative studies using simulations were conducted by Hall (1972), Geng, Wang and Miller (1979), Keselman, Games and Clinch (1979), Conover, Johnson and Johnson

[1] Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice. E-mail: dominika.polko@ue.katowice.pl. ORCID ID: https://orcid.org/0000-0003-4098-6647.

(1981), Balakrishnan and Ma (1990), Lim and Loh (1996), Marozzi (2011) and Gogoi and Gogoi (2017).

Pesarin (2001) initiated the approach for the nonparametric testing problem. He considered reducing the scope of the null hypothesis by splitting it into several partial hypotheses. This nonparametric approach is to perform some reasonable tests for each individual partial hypothesis and combine their results with a chosen combining function. A multi–aspect test to location problem was considered in works by Marozzi (2004), Marozzi (2007) or Salmaso and Solari (2005). Nonparametric combination procedure to asses overall *ASL* (*achieved significance level*) value is very useful in the scale problem too (Marozzi 2012a, 2012b).

It is more complicated to test differences between both location parameters and scale parameters of the distribution in the populations studied. A need of simultaneously detecting location and scale changes arises in many areas, for example in financial matters in stock prices analysis (Lunde and Timmermann, 2004), in the analysis of production processes, for example, testing of the process stability (Park, 2015a), climate dynamics analyses (Yonetani and Gordon, 2001) or biomedical researches (Muccioli, et al., 1996).

Lepage (1971) initiated this topic with his proposal by combining the Wilcoxon rank sum and Ansari–Bradley's test statistics for location and scale parameters. A test based on Lepage's proposal but using Mood's test statistic for the scale parameter was presented by Duran et al. (1976). Later, Lepage's procedure was reviewed and discussed extensively by many authors (Murakami, 2007; Neuhauser, Leuchs and Ball, 2011). Marozzi (2008) considered the problem of location and scale using a nonparametric combination procedure proposed by Pesarin (2001). All the reviewed and compared by a simulation study test statistics used quadratic forms and allow one to consider only two–sided alternatives. Park (2015b) excluded the use of the quadratic form for the test statistics to accommodate various types of alternatives. The proposition described in this article also enables the formulation of any types of alternative hypotheses. The purpose of this research is to present several statistical test proposals for joint comparison of location and scale parameters in two populations using a permutation procedure for a multi–aspect testing approach.

The rest of the paper is organized as follows. In Section 2 the research problem is formally defined and two tests known in the literature for simultaneous testing location and scale parameters are presented. In Section 3 the nonparametric combination procedure for location–scale problem is characterized. In Section 4 several test statistics for a joint comparison of the location and scale parameters in two populations using a nonparametric, permutation procedure to assess *ASL* values are proposed. This Section also contains a simulation comparison of their size and power under normality. There are two cases considered in simulations: both partial alternative hypotheses are one– or two–sided. In Section 5 concluding remarks are presented.

2. Simultaneous tests for the location–scale problem

In order to discuss the location–scale problem, let observations $x_{11},...,x_{1n_1}$ and $x_{21},...,x_{2n_2}$ be random samples taken from populations with distribution functions F_1 and F_2 respectively. Populations are of continuous distributions F_i for $i=1$, 2 with unknown parameters. The null hypothesis of comparing two populations is in the form of $H_0 : F_1(x) = F_2(x)$. In the paper, the location–scale problem is considered where μ_1, μ_2 and σ_1, σ_2 are locations and scale parameters of populations 1 and 2 respectively. According to this notation, the null hypothesis can be also written as

$$H_0 : \mu_1 = \mu_2 \wedge \sigma_1 = \sigma_2, \tag{1}$$

versus alternative hypothesis

$$H_1 : \mu_1 \neq \mu_2 \vee \sigma_1 \neq \sigma_2. \tag{2}$$

In the literature, authors most often refer to the Lepage test. However, you can find another test to verify the same hypothesis, proposed earlier, but not so well known Cucconi test (Bonnini, et al., 2014). The Cucconi test (Cucconi, 1968) used in the situations of finding differences in the location and scale parameters uses the statistic of the form (Marozzi, 2009)

$$C = \frac{U^2 + V^2 - 2\rho UV}{2(1-\rho^2)}, \tag{3}$$

where

$$U = \frac{6\sum_{i=1}^{n_1} R_{1i}^2 - n_1(n+1)(2n+1)}{\sqrt{n_1 n_2 (n+1)(2n+1)(8n+11)/5}},$$

$$V = \frac{6\sum_{i=1}^{n_1} (n+1-R_{1i})^2 - n_1(n+1)(2n+1)}{\sqrt{n_1 n_2 (n+1)(2n+1)(8n+11)/5}},$$

$n = n_1 + n_2$,

R_{ji} – rank of x_{ji} in pooled sample $x = (x_1, x_2)$,

and $\rho = \frac{2(n^2-4)}{(2n+1)(8n+11)} - 1$.

Hypothesis H_0 is rejected if C>-lnα, where α is the test size (Marozzi, 2009).

The second test, the Lepage test (1971), refers to the merger of two test statistics. This test is a combination of the Wilcoxon–Mann–Whitney (Mann and Whitney, 1947; Wilcoxon 1949) and Ansari–Bradley (Ansari and Bradley, 1960) test statistics

$$L=\frac{(W-E_{\mathbf{0}}(W))^2}{V_0(W)}+\frac{(A-E_{\mathbf{0}}(A))^2}{V_0(A)}=\tilde{W}^2+\tilde{A}^2\,, \qquad (4)$$

where

W – Wilcoxon–Mann–Whitney test statistics,

A – Ansari–Bradley test statistics,

$E_0(W)=n_1(n+1)/2$, $V_0(W)=n_1n_2(n+1)/12$,

when *n* is even: $E_0(A)=n_1(n+2)/4$, $V_0(A)=n_1n_2(n+2)(n-2)/48/(n-1)$,

when *n* is odd: $E_0(A)=n_1(n+1)^2/4/n$, $V_0(A)=n_1n_2(n+1)(n^2+3)/48/n^2$,

$\tilde{W}$ – Wilcoxon–Mann–Whitney standardized test statistics,

$\tilde{A}$ – Ansari–Bradley standardized test statistics.

Hypothesis H_0 is rejected if the calculated value of the test statistic exceeds the critical value of the test. Tables for the Lepage test can be found in Lepage (1973).

3. Nonparametric combination procedures

The problem of testing complex hypotheses can also be considered as proposed by Pesarin (2001). When the test concerns the location–scale testing problem then two partial hypotheses are taken into account. The null hypothesis in the form of (1) can be written differently as

$$H_0: H_0^{(1)} \wedge H_0^{(2)} \qquad (5)$$

and the corresponding decomposition is

$$H_0^{(1)}: \mu_1=\mu_2 \text{ and } H_0^{(2)}: \sigma_1=\sigma_2\,. \qquad (6)$$

An alternative hypothesis, which is a negation of the null hypothesis, can then be written as

$$H_1: H_1^{(1)} \vee H_1^{(2)}\,, \qquad (7)$$

where

$$H_1^{(1)}: \mu_1 \neq \mu_2\,,\ \ H_1^{(2)}: \sigma_1 \neq \sigma_2\,. \qquad (8)$$

The paper considers a simulation approach based on the permutations of a data set. A nonparametric, complex procedure was used to assess the overall *ASL* (*achieved significance level*) value. The procedure for testing the null hypothesis versus the alternative hypothesis consists of two steps. First, each of the partial null hypotheses is tested. Then, the results of the first step are jointly managed to solve the general problem (Marozzi, 2008).

In the first stage of separate testing of each of the considered partial null hypotheses, *ASL* values are estimated following the traditional permutation method used during the verification of a single parameter hypothesis, i.e.:

1. Assume the level of significance α.
2. Calculate the value of statistic for the sample data (T_0).
3. Perform the permutations of variable *N*–times and calculate the statistic test value (T_k) for each permutation.
4. On the basis of the empirical distribution of statistic, the *ASL* value is determined.

Regarding location–scale testing, two partial aspects may be emphasized. Two permutation tests are performed and an estimate of two *ASL* values are obtained: the first for the equality test of mean or median parameters, the second for the equality test of scale parameters of the form

$$A\hat{S}L_{T^{(1)}}\left(T_0^{(1)}\right)=\frac{0{,}5+\sum_{k=1}^{N} I\left(\left|T_k^{(1)}\right|\geq\left|T_0^{(1)}\right|\right)}{N+1} \tag{9}$$

and

$$A\hat{S}L_{T^{(2)}}\left(T_0^{(2)}\right)=\frac{0{,}5+\sum_{k=1}^{N} I\left(\left|T_k^{(2)}\right|\geq\left|T_0^{(2)}\right|\right)}{N+1}. \tag{10}$$

where *I*(.) denotes the indicator function.

With respect to standard permutation *ASL* estimation, 0.5 and 1 are added to the numerator and denominator of the fraction, respectively. The reason is to obtain estimated *ASL* values in open interval $(0{,}1)$ avoiding computational problems, which may arise in the second step of the nonparametric procedure. However, since large *N* is used, this correction is practically irrelevant (Marozzi, 2008).

The second step of the nonparametric procedure of statistical inference includes calculation of the overall *ASL* value using the combining function (Pesarin, 2001)

$${}_{\varphi 12}T=\varphi\left(ASL_{T^{(1)}}, ASL_{T^{(2)}}\right).$$

There are many forms of combining functions for determining the overall *ASL* value, although the authors the most often used combining functions:

- the Fisher omnibus combining function (Fisher, 1932)
  $C^{(F)}=-2\left[\log\left(A\hat{S}L_{T^{(1)}}\right)+\log\left(A\hat{S}L_{T^{(2)}}\right)\right],$

- the Liptak combining function (Liptak 1958) $C^{(L)} = \Phi^{-1}\left(1 - A\hat{S}L_{T^{(1)}}\right) + \Phi^{-1}\left(1 - A\hat{S}L_{T^{(2)}}\right)$, where Φ denotes the standard normal distribution function,
- the Tippet combining function (Tippet, 1931) $C^{(T)} = \max\left\{1 - A\hat{S}L_{T^{(1)}}, 1 - A\hat{S}L_{T^{(2)}}\right\}$.

The observed statistics value for the sample data can be determined as

$$_{\varphi 12}T_0 = \varphi\left(A\hat{S}L_{T^{(1)}}\left(T_0^{(1)}\right), A\hat{S}L_{T^{(2)}}\left(T_0^{(2)}\right)\right), \tag{11}$$

and its distribution is determined on the basis of the same permutations of the first step of this procedure, for example the *k*–th permutation value of statistics is computed

$$_{\varphi 12}T_k = \varphi\left(A\hat{S}L_{T^{(1)}}\left(T_k^{(1)}\right), A\hat{S}L_{T^{(2)}}\left(T_k^{(2)}\right)\right). \tag{12}$$

Overall *ASL* value of the test is estimated by using the formula

$$A\hat{S}L_{_{\varphi 12}T} = \frac{\sum_{k=1}^{N} I\left(_{\varphi 12}T_k \geq _{\varphi 12}T_0\right)}{N}. \tag{13}$$

where *I*(.) denotes the indicator function.

4. Monte Carlo study

Most often, the statistical inference concerns situations where there are differences between the considered populations without indicating the nature of this difference. An alternative hypothesis of form (2) is then considered. Thanks to permutation tests, it is also possible to consider one–sided alternative hypotheses, for example:

$$H_1^{(1)}: \mu_1 > \mu_2 \text{ or } H_1^{(2)}: \sigma_1 > \sigma_2.$$

The study considered various forms of test statistics (Table 1). The simulations consisted of calculating the size and power of the presented tests using a complex, nonparametric method of testing the location and scale parameters. All 1–8 models were used in the simulation study when partial alternative two–sided hypotheses were considered. To verify the null hypothesis, when partial alternative hypotheses were one–sided hypotheses, models 1–5 were used. Model 6 considers the form of test statistics included in the combination of statistics used in the Lepage test. For comparative purposes, the results for the permutation Lepage test (model 7) and permutation Cucconi test (model 8) were also included when alternative two–sided hypotheses were considered. The nonparametric combination procedure for the estimated overall *ASL* value was used when considering models 1–6.

In the simulation study samples taken from normal distribution with $n_1 = 10, n_2 = 15$ sample sizes were considered. Three situations were analysed:

a) $\mu_1 - \mu_2 > 0$ and $\sigma_1 / \sigma_2 = 1$,

b) $\mu_1 - \mu_2 = 0$ and $\sigma_1 / \sigma_2 > 1$,

c) $\mu_1 - \mu_2 > 0$ and $\sigma_1 / \sigma_2 > 1$.

Parameters of the distribution from which the second sample was taken are $\mu_2 = 0$ and $\sigma_2 = 1$, whereas parameters of the distribution from which the first sample was taken are defined as follows:

a) if $\mu_1 - \mu_2 > 0$ then $\mu_1 \in (0.2, 1.6)$ with the increment 0.2 and $\sigma_1 = 1$,

b) if $\sigma_1 / \sigma_2 > 1$ then $\sigma_1 \in (1.2, 2.6)$ with the increment 0.2 and $\mu_1 = 0$,

c) if $\mu_1 - \mu_2 > 0$ and $\sigma_1 / \sigma_2 > 1$ then parameters of the distribution (μ_1, σ_1) equal from (0.2,1.2) to (1.6,2.6) with the increment 0.2 for each parameter.

Table 1. Statistics used in simulation study

Model	Statistics $T^{(1)}$	Statistics $T^{(2)}$
1	$T_1^{(1)} = \bar{x}_1 - \bar{x}_2$	$T_1^{(2)} = \frac{s_1^2}{s_2^2}$
2	$T_2^{(1)} = m_1 - m_2$	$T_2^{(2)} = \frac{R_1}{R_2},$
3	$T_3^{(1)} = W$	$T_3^{(2)} = \frac{R_1}{R_2},$
4	$T_4^{(1)} = W$	$T_4^{(2)} = M,$
5	$T_5^{(1)} = W$	$T_5^{(2)} = OB,$
6	$T_6^{(1)} = \tilde{W}^2$	$T_6^{(2)} = \tilde{A}^2$
7	$T_7 = L$	
8	$T_8 = C$	

where:

$\bar{x}_1, \bar{x}_2$ – sample means from first and second population respectively,

m_1, m_2 – sample medians from first and second population respectively,

R_1, R_2 – sample ranges from first and second population respectively,

W – Wilcoxon–Mann–Whitney test statistics,

M – Mood test statistics (Mood, 1954),

OB – O'Brien test statistics (O'Brien, 1979),

$\tilde{W}$ – Wilcoxon–Mann–Whitney standardized test statistics,

$\tilde{A}$ – Ansari–Bradley standardized test statistics,
L – Lepage test statistics (4),
C – Cucconi test statistics (3).

Table 2. Size and power estimates when $\mu_1 - \mu_2 > 0$ and $\sigma_1 / \sigma_2 = 1$, $\alpha = 0.05$, for samples $n_1 = 10, n_2 = 15$ (two–sided alternative hypotheses)

Model	Distribution parameters (μ_1, σ_1)								
	$(0,1)$	$(0.2,1)$	$(0.4,1)$	$(0.6,1)$	$(0.8,1)$	$(1,1)$	$(1.2,1)$	$(1.4,1)$	$(1.6,1)$
1	0.046	0.060	0.124	0.240	0.375	0.527	0.684	0.793	0.895
2	0.048	0.065	0.114	0.206	0.307	0.430	0.570	0.689	0.816
3	0.060	0.096	0.188	0.315	0.467	0.626	0.774	0.863	0.940
4	0.054	0.094	0.163	0.295	0.450	0.566	0.734	0.831	0.922
5	0.060	0.104	0.164	0.307	0.464	0.605	0.765	0.849	0.930
6	0.057	0.069	0.111	0.207	0.343	0.502	0.649	0.765	0.889
7	0.052	0.065	0.104	0.210	0.349	0.510	0.653	0.783	0.894
8	0.055	0.072	0.105	0.207	0.347	0.508	0.662	0.769	0.890

Source: Own calculation in R program.

Table 3. Power estimates when $\mu_1 - \mu_2 = 0$ and $\sigma_1 / \sigma_2 > 1$, $\alpha = 0.05$, for samples $n_1 = 10, n_2 = 15$, (two–sided alternative hypotheses)

Model	Distribution parameters (μ_1, σ_1)							
	$(0,1.2)$	$(0,1.4)$	$(0,1.6)$	$(0,1.8)$	$(0,2)$	$(0,2.2)$	$(0,2.4)$	$(0,2.6)$
1	0.115	0.232	0.308	0.440	0.548	0.635	0.750	0.784
2	0.112	0.203	0.305	0.417	0.498	0.604	0.713	0.752
3	0.092	0.176	0.280	0.396	0.455	0.558	0.685	0.722
4	0.079	0.126	0.210	0.282	0.365	0.443	0.527	0.596
5	0.085	0.167	0.276	0.401	0.478	0.580	0.681	0.718
6	0.068	0.133	0.159	0.263	0.316	0.386	0.477	0.545
7	0.077	0.148	0.210	0.307	0.390	0.463	0.573	0.647
8	0.069	0.132	0.172	0.263	0.327	0.399	0.484	0.550

Source: Own calculation in R program.

Table 4. Power estimates when $\mu_1 - \mu_2 > 0$ and $\sigma_1/\sigma_2 > 1$, $\alpha = 0.05$, for samples $n_1 = 10, n_2 = 15$, (two–sided alternative hypotheses)

Model	Distribution parameters (μ_1, σ_1)							
	$(0.2,1.2)$	$(0.4,1.4)$	$(0.6,1.6)$	$(0.8,1.8)$	$(1,2)$	$(1.2,2.2)$	$(1.4,2.4)$	$(1.6,2.6)$
1	0.145	0.269	0.463	0.624	0.742	0.840	0.874	0.930
2	0.127	0.251	0.422	0.579	0.681	0.792	0.850	0.896
3	0.165	0.309	0.471	0.651	0.742	0.833	0.865	0.920
4	0.124	0.239	0.374	0.516	0.646	0.762	0.812	0.846
5	0.131	0.288	0.469	0.642	0.762	0.859	0.889	0.935
6	0.090	0.161	0.263	0.384	0.508	0.600	0.691	0.739
7	0.106	0.191	0.319	0.424	0.577	0.659	0.748	0.785
8	0.095	0.159	0.295	0.375	0.502	0.604	0.683	0.738

Source: Own calculation in R program.

Table 5. Size and power estimates when $\mu_1 - \mu_2 > 0$ and $\sigma_1/\sigma_2 = 1$, $\alpha = 0.05$, for samples $n_1 = 10, n_2 = 15$ (one–sided alternative hypotheses)

Model	Distribution parameters (μ_1, σ_1)								
	$(0,1)$	$(0.2,1)$	$(0.4,1)$	$(0.6,1)$	$(0.8,1)$	$(1,1)$	$(1.2,1)$	$(1.4,1)$	$(1.6,1)$
1	0.047	0.105	0.179	0.293	0.495	0.651	0.771	0.885	0.954
2	0.042	0.096	0.166	0.261	0.402	0.568	0.691	0.830	0.925
3	0.050	0.099	0.169	0.295	0.460	0.645	0.763	0.878	0.953
4	0.047	0.100	0.178	0.282	0.454	0.634	0.752	0.875	0.947
5	0.043	0.097	0.162	0.277	0.447	0.627	0.754	0.870	0.948

Source: Own calculation in R program.

For each of 1000 Monte Carlo simulations, 1000 random permutations of variables and the nominal significance level $\alpha = 0.05$ were considered. The studies used Fisher's combining function to determine the overall *ASL* value. The results of the simulations carried out to determine the size and power of the tests are presented in Tables 2–7. Estimated probabilities of rejection of the hypothesis H_0 when partial two–sided alternative hypotheses were taken under consideration are presented in Tables 2–4. In the case of partial one–sided alternative hypotheses, estimated probabilities are presented in Tables 5–7, respectively.

Table 6. Power estimates when $\mu_1 - \mu_2 = 0$ and $\sigma_1 / \sigma_2 > 1$, $\alpha = 0.05$, for samples $n_1 = 10, n_2 = 15$, (one–sided alternative hypotheses)

Model	Distribution parameters (μ_1, σ_1)							
	$(0,1.2)$	$(0,1.4)$	$(0,1.6)$	$(0,1.8)$	$(0,2)$	$(0,2.2)$	$(0,2.4)$	$(0,2.6)$
1	0.108	0.183	0.320	0.437	0.536	0.616	0.708	0.779
2	0.107	0.171	0.278	0.404	0.462	0.572	0.628	0.705
3	0.102	0.180	0.277	0.398	0.464	0.574	0.627	0.702
4	0.100	0.171	0.272	0.373	0.441	0.509	0.584	0.674
5	0.110	0.196	0.319	0.435	0.528	0.599	0.659	0.737

Source: Own calculation in R program.

Table 7. Power estimates when $\mu_1 - \mu_2 > 0$ and $\sigma_1 / \sigma_2 > 1$, $\alpha = 0.05$, for samples $n_1 = 10, n_2 = 15$, (one–sided alternative hypotheses)

Model	Distribution parameters (μ_1, σ_1)							
	$(0.2,1.2)$	$(0.4,1.4)$	$(0.6,1.6)$	$(0.8,1.8)$	$(1,2)$	$(1.2,2.2)$	$(1.4,2.4)$	$(1.6,2.6)$
1	0.167	0.345	0.526	0.672	0.806	0.865	0.928	0.949
2	0.159	0.322	0.490	0.625	0.745	0.838	0.895	0.933
3	0.158	0.320	0.487	0.616	0.748	0.846	0.894	0.924
4	0.156	0.313	0.473	0.600	0.745	0.811	0.881	0.906
5	0.160	0.352	0.518	0.661	0.801	0.866	0.931	0.948

Source: Own calculation in R program.

The size of the tests is shown in Tables 2 and 5 in the first column. For all models, the obtained estimated probabilities are close to the nominal level of significance $\alpha = 0.05$. The tests considered achieved comparable results in the case of small samples, the size of which was not equal. The tests used in models 1, 3 and 5 were the most powerful. The probabilities of detecting differences between populations increased with increasing differences between the respective location or scale parameters for both considered partial two– and one–sided alternative hypotheses.

5. Conclusions

The simulation research aimed to determine the ability of the presented location and scale tests to maintain the nominal probability of committing the type I error and the ability to obtain a high probability of rejecting a false zero hypothesis in the conditions of changing distribution parameters in populations from which samples were taken.

The tests that verify the hypothesis about the identity of location and scale parameters in the studied populations are presented. The article considered various forms of test statistics. A simulation study to determine the size and power of tests was carried out using permutation tests.

When analysing the results obtained it can be seen that the inference about the significance of differences between populations is possible with the use of the proposed solution. All testing procedures (under normality) ensured control of type I error at the assumed level of significance. The simulation analysis indicated that the proposed tests allowed the inference about the differences in location or scale parameters, as well as differences in both location and scale parameters of distributions. The results for the permutation Lepage test and permutation Cucconi test are also presented where two–sided alternative hypothesis is considered. Higher power of tests was achieved thanks to the use of a nonparametric procedure that uses Fisher's combining functions to evaluate the overall *ASL* value. The observed assessments of the probability of rejection of the null hypothesis were similar for various pairs of test statistics considered in the simulations. One advantage of the procedure presented in the article is also the possibility of formulating an alternative hypothesis in the form of partial directional hypotheses. The method can be used even in the case of small sample sizes. In the research, other forms of combining functions can be considered and a simulation study taking into account the various distributions of the studied variables can be performed. The direction of further research also concerns the extension of the method to a multidimensional case.

REFERENCES

ANDERSON, M. J., WALSH, D. C. I., CLARKE, K.R., GORLEY, R. N., GUERRA–CASTRO, E., (2017). Permutational Multivariate Analysis of Variance (PERMANOVA), Wiley StatsRef: Statistics Reference Online, pp. 1–15.

ANSARI, A. R., BRADLEY, R. A., (1960). Rank–sum tests for dispersions. Annals of Mathematical Statistics 31, pp. 1174–1189.

BALAKRISHNAN, N., MA, C. W., (1990). A comparative study of various tests for the equality of two population variances, Journal of Statistical Computation and Simulation, 35, pp. 41–89.

BONNINI, S., CORAIN, L., MAROZZI, M., SALMASO, L., (2014). Nonparametric Hypothesis Testing Rank and Permutation Methods with Applications in R, John Wiley & Sons, Ltd.

CHANG, C.–H., PAL, N., (2008). A Revisit to the Behrens–Fisher Problem: Comparison of Five Test Methods. Communications in Statistics – Simulation and Computation, 37, (6), pp. 1064–1085.

CONOVER, W. J., JOHNSON, M. E., JOHNSON, M. M., (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data, Technometrics, 23, pp. 351–361.

CUCCONI, O., (1968). Un nuovo test non parametrico per it confronto tra due gruppi campionori. Giornale degli Economisti, XXVII, pp. 225–248.

DURAN, B. S., TSAI, W. S., LEWIS, T. O., (1976). A class of location-scale tests, Biometrika, 63, pp. 173–176.

FISHER, R. A., (1932). Statistical Methods for Research Workers, 4 ed., Edinburgh: Oliver & Boyd.

GENG, S., WANG, W. J., MILLER, C., (1979). Small sample size comparisons of tests for homogeneity of variances by Monte-Carlo. Communications in Statistics – Simulation and Computation, 8, pp. 379–389.

GOGOI, P., GOGOI, B., (2017). Some Tests Procedures for Scale Differences. International Advanced Research Journal in Science, Engineering and Technology, Vol. 4, Issue 11, pp. 155–166.

HALL, I. J., (1972). Some comparisons of tests for equality of variances, Journal of Statistical Computation and Simulation, 1, pp. 183–194.

JANSSEN, A., PAULS, T., (2005). A Monte Carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two–sample problems. Computational Statistics, 20 (3), pp. 369–383.

KESELMAN, H. J., GAMES, P. A., CLINCH, J. J., (1979). Tests for homogeneity of variance. Communications in Statistics – Simulation and Computation, 8, pp. 113–119.

KOŃCZAK, G., (2016). Testy permutacyjne, Teoria i zastosowania, Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

LEPAGE, Y., (1971). A combination of Wilcoxon's and Ansari–Bradley's statistics, Biometrika 58, pp. 213–217.

LEPAGE, Y., (1973). A table for a combined Wilcoxon Ansari–Bradley statistic, Biometrika 60, pp. 113–116.

LIM, T.S., LOH, W. Y., (1996). A comparison of tests of equality of variances, Computational Statistics and Data Analysis, 22, pp. 287–301.

LIPTAK, I., (1958). On the combination of independent tests. Magyar Tudomanyos Akademia Matematikai Kutato Intezenek Kozlomenyei 3, pp. 127–141.

LUNDE, A., TIMMERMANN, A., (2004). Duration dependence in stock prices: an analysis of bull and bear markets, Journal of Business and Economic Statistics, 22, pp. 253–273.

MANN, H., WHITNEY, D., (1947). On a test of whether one of two random variables is stochastically larger than the other, Annals of Mathematical Statistics, 18, (1), pp. 50–60.

MAROZZI, M., (2004). A bi-aspect nonparametric test for the two-sample location problem, Computational Statistics and Data Analysis, 44, pp. 639–648.

MAROZZI, M., (2007). Multivariate tri–aspect non-parametric testing, Journal of Nonparametric Statistics, 19, pp. 269–282.

MAROZZI, M., (2008). The Lepage location–scale test revisited, Far East Journal of Theoretical Statistics 24, pp. 137–155.

MAROZZI, M., (2009). Some notes on the location–scale Cucconi test, Journal of Nonparametric Statistics, 21, 5, pp. 629–647.

MAROZZI, M., (2011). Levene type tests for the ratio of two scales, Journal of Statistical Computation and Simulation, 81, pp. 815–826.

MAROZZI, M., (2012a). A distribution free test for the equality of scales. Communication in Statistics – Simulation and Computation, 41, pp. 878–889.

MAROZZI, M., (2012b). A combined test for differences in scale based on the interquantile range. Statistical Papers, 53, pp. 61–72.

MOOD, A. M., (1954). On the asymptotic efficiency of certain nonparametric two–sample tests. Ann Math Stat 25, pp. 514–522.

MUCCIOLI, C., BELFORD, R., PODGOR, M., SAMPAIO, P., DE SMET, M., NUSSENBLATT, R., (1996). The diagnosis of intraocular inflammation and cytomegalovirus retinitis in HIV–infected patients by laser flare photometry, Ocular Immunology and Inflammation, 4, pp. 75–81.

MURAKAMI, H., (2007). Lepage type statistic based on the modified Baumgartner statistic, Computational Statistics & Data Analysis, 51, pp. 5061–5067.

NEUHAUSER, M., LEUCHS, A.-K., BALL, D., (2011). A new location-scale test based on a combination of the ideas of Levene and Lepage, Biometrical Journal, 53, pp. 525–534.

O'BRIEN, R. G., (1979). A general ANOVA method for robust test of additive models for variance, Journal of the American Statistical Association, 74, pp. 877–880.

PARK, H-I., (2015a). Simultaneous Tests with Combining Functions under Normality, Communications for Statistical Applications and Methods, Vol. 22, No. 6, pp. 639–646.

PARK, H-I., (2015b). Nonparametric Simultaneous Test Procedures, Revista Colombiana de Estadística, 38(1), pp. 107–121.

PESARIN, F., (2001). Multivariate Permutation Test with Applications in Biostatistics, Chichester: Wiley.

SALMASO, L., SOLARI, A., (2005). Multiple aspect testing for case-control designs, Metrika, 62, pp. 331–340.

TIPPETT, L. H. C., (1931). The Methods of Statistics, London: Williams and Norgate.

WILCOXON, F., (1949). Some rapid approximate statistical procedures, Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation.

YONETANI, T., GORDON, H. B., (2001). Abrupt changes as indicators of decadal climate variability, Climate Dynamics, 17, pp. 249–258.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 167–179, DOI 10.21307/stattrans-2019-040
Submitted – 11.09.2019; Paper ready for publication – 26.11.2019

SUBJECTIVE AND COMMUNITY WELL-BEING INTERACTION IN A MULTILEVEL SPATIAL MODELLING FRAMEWORK [1]

Włodzimierz Okrasa [2], Dominik Rozkrut [3]

ABSTRACT

Analysing the cross-level interaction between individual and community well-being requires a joint involvement of both 'vertical' and 'horizontal' perspectives. While multilevel modelling separates the effects resulting from personal characteristics from those resulting from community features, the need to account for spatial variation and geographic membership proves that space and place matter, too. In this paper, the explicitly-spatial multilevel model has been developed to this effect, namely to identify both types of effects, space and place-related, using the hierarchical (nested) data structure for the smallest administrative units – NUTS5/LAU2, i.e. communes (gminas). In their analysis, the authors employed two methods for measuring well-being: (i) individual (subjective) well-being measure derived from the nation-wide Time Use Survey data, which they occasionally replaced with 'life satisfaction' type of self-reported measures, and (ii) multidimensional index of local deprivation composed of eleven domain-scales. The spatial multilevel modelling has been extended by an attempt to assess what effect spatial interaction has on cross-level relationships. Its inclusion in the discussion with which this paper concludes seems recommendable, as it indicates the need for more systematic efforts towards a spatially-integrated approach to this kind of modelling problems.

Key words: spatial analysis, measuring subjective well-being, community deprivation, social capital.

1. Introduction

There are several reasons to analyse community and individual well-being jointly and, by the same token, to focus on the relationship between them, especially in the local development context. Several aspects of this relation have been recognized and discussed thoroughly in the literature, inspired among others by Stiglitz-Fitoussi-Sen (2009) report challenging the tradition of using

[1] This article is based on the presentation at the 62nd ISI World Statistics Congress, 13-18 August, Kuala Lumpur.

[2] Cardinal Stefan Wyszynski University in Warsaw, and Statistics Poland. E-mail: w.okrasa@stat.gov.pl. ORCID ID: https://orcid.org/0000-0001-6443-480X.

[3] Statistics Poland. ORCID ID: https://orcid.org/0000-0002-0949-8605.

GDP as the main measure of social progress, along with concomitantly growing awareness of the significant role of subjective well-being in economic development (esp. sustainable development, e.g. Helliwell, et al., 2010) at both macro-level (although not in an unambiguous way, e.g. Easterlin, 2010), as well as in connection with community (e.g. Phillip and Wong, 2017).

In the employed modelling approach, an empirical application is preceded by discussion of the measurement and data issues, including the problem of creation an analytical multi-source database (through 'bottom up' integration of units from different surveys) and construction of the major well-being measures: (i) multidimensional index of local deprivation encompassing eleven components, each of them being constructed from public-use data file (Local Data Bank, Statistics Poland), using 'confirmatory' version of factor analysis (for all 2478 communes (gminas)), and (ii) individual (subjective) well-being measure derived from the nation-wide Time Use Survey, which is substituted in some contexts by self-reported measures from national surveys on Social Cohesion or Social Diagnosis.

An empirical application of the multilevel spatial modelling (which constitutes the major portion of the remaining part of the paper) is preceded by searching for main factors and auxiliary covariates affecting individual (subjective) well-being, while looking after the issue of endogeneity.

When expressed in a way analogous to the so-called basic 'life-satisfaction equation', subjective well-being might be treated as a function of residents' income and hours of work vis-a-vis the impact of community well-being (or deprivation) through employing a causal type of reasoning using path analytic version of the structural model. A path-analytic version of the structural model is employed to decompose total effect of the independent variable into the natural direct and indirect effects (Hong, 2015; Okrasa and Rozkrut, 2018).

Another important factor at the community level (referred often to social cohesion) is social capital, the relative impact of which - weighted against individual income - is checked using the 'compensating variation' approach. Social capital, indicated by the intensity of the third sector organizations' presence in a community, can be interpreted as the amount of money required to compensate a person for a possible loss in utility (for instance, like when price is rising). The 'compensating variation' approach to social capital allows one to identify the utility gain derived from a unit increase in social capital (Anand and Montovani, 2018; Okrasa, 2018).

Following exploration of spatial patterning, clustering and spatial dependence (using GeoDa procedures, Fischer and Getis, 2010) a direct assessment of the spatial interaction effect on the cross-level relationships is also attempted (Patuelli and Arbia, 2016) using flow-type data from between-community migration public statistics.

In the concluding section, a spatially integrated approach to vertical (multilevel) and horizontal (across areal units) relationships between individual (subjective) and group (community) measures of well-being is discussed towards elaborating a comprehensive methodological framework, as noted by Arcaya et al., (2012) who analysed area variations in health, accounting for spatial and membership processes simultaneously providing valuable insights (p. 824).

2. Methodology: operationalzation, data and models

The increased focus on well-being (along the *beyond-GDP* paradigm) results also in several guidelines and recommendations offered in the literature on the measurement of subjective well-being in public statistics, such as ONS Report (Dolan et al., 2011); OECD Report (2013, 2015), CNSTAT/Stone and Mackie (2014); Kalton et al. (2015). While there is a consensus regarding individual (subjective) well-being measures that they are supposed to cover all or some aspects of its triadic structure of subjective well-being – evaluation (e.g. *Satisfaction from Life);* experience (*How did you feel yesterday);* and eudaimonic (*Sense of Life)* – the community well-being measurement approaches still await similar elucidation (e.g. Kim and Ludvigs (2017)), although several country-specific approaches have been already well developed within public statistical systems (for instance, in Australia, Canada, USA, and UK).

2.1. **Individual (Subjective) well-being:** Time Use Survey/TUS data-based measures

Since psychometric, self-reported data-based measures of well-being are often criticized by econometricians for their arbitrariness and low reliability, data from time use surveys (collected with day reconstruction method/DRM) are being recommended instead - see Kahneman and Krueger (2006).

Amount of the time *h* spent by respondent on performing an activity with information on emotion (negative-neutral-positive) s/he was associating with this activity (as ‘time in unpleasant state’) can be reflected by the value of U-index (e.g. Krueger et al., 2009, p. 19):

$$\mathbf{U_i = \Sigma_j \left(\mathit{I_{ij} h_{ij}} / \Sigma_j \mathit{h_{ij}} \right)} \text{ (in TUS conducted in 2013: I = -1. 0. +1)} \qquad (1)$$

and

$U = \Sigma_i(\Sigma_j\, I_{ij}h_{ij} / \Sigma_j h_{ij}) / N$ for N-persons / group in population, interpreted as the average proportion of time that the members of the group spend in an unpleasant state.

Such an approach to measuring life satisfaction or happiness is not only more consistent with the concept of utility[4]. The lack of such an underlying concept makes some authors (e.g. Gibson, 2016) full of reservation towards the use of these measures. But it has a direct reference to the *capability approach* according to A. Sen’s interpretation, who stresses that well-being should be conceived directly in terms of *functionings* and *capabilities* instead of resources or utility (e.g. Alkire, 2015). Time use data seems to be one of the most reliable source of information on functioning and capabilities.

2.2. **Community Well-Being (CWB)** is a multifaceted and multilevel concept, hardly covered by standardized procedures of operationalization and measurement. It is a "concept developed by synthesizing research constructs related to resident’s perceptions of the community, … needs

[4] For instance, Gibson (2016) maintained that there is no theoretical justification for maximizing either happiness or life satisfaction due to the fact that neither correspond to utility.

> fulfilments, observable community conditions, and the social and cultural context…" (Sung and Phillips 2016:2 [in Phillpis and Wong 2017). The important features of CWB often include community cohesion (or local, spatial cohesion), which is interpreted here as any of the possible patterns of configuration of the economic cohesion and/or social cohesion and/or territorial cohesion (following Kearns and Forest (2001).

Both types of measures - individual and community well-being – constitute the main input of the Multi-source Analytical Database (MAD). It encompasses Multidimensional Index of Local Deprivation (MILD) for 2478 communes/gminas (NUTS5/LAU2), composed of eleven (pre-selected) domains of deprivation - each characterized by a number of original items: *ecology* – *finance* – *economy* – *infrastructure* – municipal *utilities* – *culture* – *housing* – *social assistance* – *labour market* – education – *health* [65 items].

Other constitutive components of AMD are: Time Use Survey *(TUS2013)* and Social Diagnosis, an independent survey conducted bi-annually by a consortium of universities since 2003 to 2015 (12 352 households or 26 308 persons at age 16+), and data from EU-SILC (Survey of Income and Living Conditions, conducted on a regular basis in member countries of the European Union). Figure 1 presents the structure of MAD, where darkened centre marks the scope of data integrated in the following analysis.

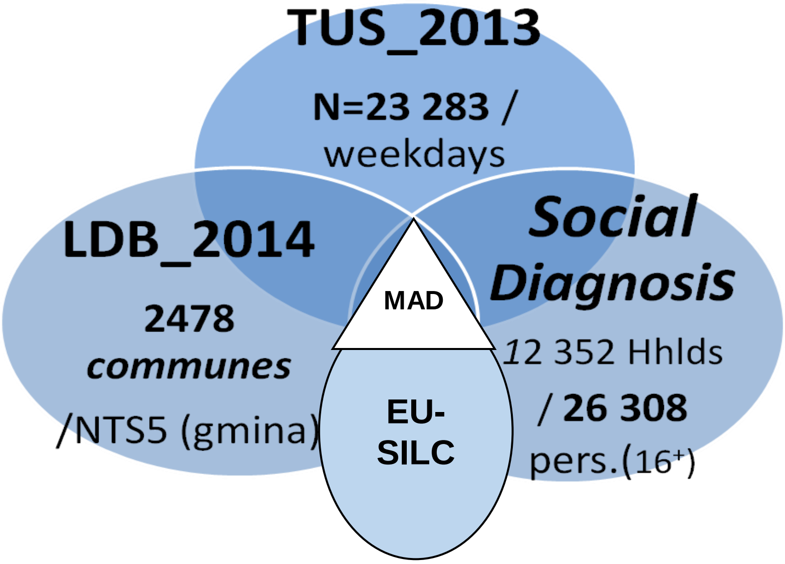


Figure 1. Multisource Analytical Database MAD

The data from Social Diagnosis survey allowed us to construct several compositional types of measures of community well-being. Specifically, measures of the level of *satisfaction* of residents – based on the percentage of 'satisfied' or 'very satisfied' on each of the five scales – were attributed to communes (in which at least 10 households were identified as included into the survey). The following scales were built: (i) satisfaction with *living conditions*; (ii) satisfaction with *living environment;* (iii) satisfaction with *social and family relations*; (iv) satisfaction with *personal situation*, and (v) *disapproval of antisocial* behaviour.

2.3. Individual and community level factors of subjective well-being

Several working hypotheses implied by theoretical considerations or by the results of other research in the literature shown to be subject to verification on the ground of the above sketched MAD. Two of them are briefly checked here. One refers to the extensively discussed trade-off between income from earning and time spend on work (Clark 2018). The second hypothesis concerns the role of social capital in the face of a possible loss of income by household (Anand and Montovani, 2018).

Basic Well-Being Equation – hypothesis of income and time of work trade-off.

Approximation of the so-called in the literature basic well-being equation or ‘life satisfaction equation’ (e.g. Clark 2018) is made here with the following equation:

$$\text{Well-Being} = \beta_1 Y + \beta_2 h + \theta X + \varepsilon \tag{2}$$

where h – time of work; Y earning, and X also auxiliary covariates.

Results are in Table 1 (next section).

The role of social capital – compensating variation approach.
Complementary to the above considerations of work and earning trade-off the role of community's social capital can be tested using the so-called ‘compensating variation ‘ approach (e.g. Anand and Montovani, 2018) .

Formally, a life satisfaction equation can be re-written as:

$$U^0(y^0, SC^0)=U^1(y^0+CV,SC^1) \tag{3}$$

where y is household income, SC stands for social capital, and CV for compensating variation (or CV for y), which can be obtained by identifying the utility gain derived from a unit increase in social capital. Accordingly, the expected utility given any particular value of social capital can be written as:

$$(U_i|SC_i,y_i,X_i) = \beta_0+\beta_y y_i+\beta_{sc}SC_i+\gamma' X_i+\varepsilon_i \tag{4}$$

where *X* represents all additional covariates.

Following Anand and Montovani (2018), CV can be defined as

$$CV=\beta_{SC} / \beta_{y.} \tag{5}$$

(see Anand and Montovani 2018 for details)

These two aspects of relation in which income remains, on the one hand, with time of work and, on the other, with social capital, can be arranged in a joint (extended) well-being equation, with social capital included into the set of predictors. Results are in Table 1.

2.4. Individual well-being and community well-being relationship - a multilevel modelling approach

In order to capture the effect of community for individual well-being, or the so-called membership process, multilevel modelling approach seems to be most appropriate (e.g. Arcaya et al., 2012, Okrasa, 2017). Ideally, it should employ hierarchically nested structure data, which is not the case of data in MADb, where for the selected communes/gminas, the group-level data are complemented by data derived from individual (household) level. However, it suffices to demonstrate the logic of the approach here, albeit with caution in interpretation of detailed results since formally admissible procedure applied to available data of official statistics can only provide an empirical illustration or argument for the appropriateness of such a modelling approach.

Having made the needed reservations, the following model was employed, using notations (e.g. Subramanian, 2010):

- $\boldsymbol{y_{ij}}$; well-being of *i* individual in *j* commune/gmina;
- $\boldsymbol{x_{1ij}}$ predictor of individual – such as: age, education or satisfaction (e.g. from life in a community, family life, etc.)
- predictor of macro-level: *Multideminsonal Index of Local Deprivation* for j-commune/gmina /$MILD_j$

- Model for one-level regression: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + e_{ij}$ (6)

  Let y_{ij} stands for household disposable equivalised income:

  $$y_{ij} = \beta_{0j} + \beta_{1j}\, \textit{ability-to-meet-ends}_{ij} + \beta_{2j}\textit{local-deprivation}_{ij} + e_{ij}$$

  where: $\beta 0j$ – refers to $X0ij$ average score on a well-being scale in j-th commune/gmina (e.g. 'ability to meet ends', X_{0ij} =1);

  β_i – average differentiation of individual well-being associated with individual material status (X_{1ij}) across all territorial units (communes/gminas);

  e_{0ij} – residual term for the level-1.

- Two-level model to account for hierarchical data structure can be specified as two-level regression, to explain the variation of the regression coefficients

β_j through including the level of local deprivation *(alter. local development indicator* $Z_j \equiv MILD(2016)$*):*

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \tag{7}$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} MILD_j + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21} MILD_j + u_{2j}$$

Rearranging terms we obtain:

$$Y_j = \gamma_{00} + \gamma_{10} X_{1ij} + \gamma_{20} X_{2ij} + \gamma_{01} MILD_{jj} + \gamma_{11} X_{1ij} MILD_j + \gamma_{21} X_{2ij} MILD_j$$

$$+ u_{1j} X_{1ij} + u_{1j} X_{2ij} + u_{0j} + e_{ij} \tag{8}$$

- where w_{1j} is a 2-level predictor, i.e. the index of local deprivation, $MILD_{1j}$. Results are in Table 2.

2.5. **Spatial aspects** - checking for spatial dependence

Estimation of the spatial regression model parameters (notation for individual observation):

$$y_i = \rho \sum_{j=1}^{n} W_{ij}\, y_j + \sum_{r=1}^{k} X_{ir} \beta_r + \varepsilon_i \tag{9}$$

where: y_i – the dependent variable for observation *i*; X_{ir} *k* – explanatory variables r = 1. …. k with associated coefficient β_r; ε_i is the disturbance term; ρ is the parameter of the strength of the average association between the dependent variable values for region/observations and the average of them for their neighbours (e.g. LeSage and Pace. 2010. p. 357). The above specification of the spatial regression model assumes that ε_i is meant as the *spatially lagged* term – versus *spatial error* formulation - for the dependent variable (which is correlated with the dependent variable), that is: $\varepsilon_i = \rho W_{i.} y_i + X_{i.} \beta + \epsilon_{i.}$ The latter type of model is used below to check *how* and *why* *'*place' and *'space'* matter.

3. Results

At a glance, results are generally in line with the hypotheses cited above.

As regards the impact of income vs. work time, there are opposite directions of influence of income and work time on well-being measured here by U-index. While greater income is positive for individual well-being (U-index decreases with growing income), the increased amount of time spent on work is negative (U-index increases). Question arises about the point of balance (trade-off between the two factors of well-being – see Kahneman and Deaton (2010) for comparison of the income effect.

Table 1. The *Well-Being Equation* extended by community cohesion – social capital – and individual-level variables

	Unstand. Coefficients		**Stnd. Coeff.**		
Predictors	**B**	**Std. Error**	**Beta**	**t**	**Sign.**
(Constant)	0.029	0.027		1.068	0.285
Job-time (main and additional)	0.004	0.000	**0.285**	24.630	0.000
Income of H'hold *pc* - monthly	-1.841E-05	0.000	**-0.087**	-6.987	0.000
MILD_2014 Local Deprivation	0.000	0.000	0.118	6.630	0.000
Subsidies Real < Simulated/fair	-0.011	0.002	-0.070	-6.887	0.000
Risk assoc. w/depr. Soc.Welfare	-0.036	0.002	-0.649	-15.626	0.000
Risk assoc. w/depr. Lab. Market	0.050	0.003	0.809	18.454	0.000
Ratio 'in-work' to 'not-in-work'	-0.010	0.001	-0.080	-6.900	0.000
Rural	-0.007	0.003	-0.030	-1.978	0.048
U-R mixed	-0.014	0.002	-0.074	-5.547	0.000
Trust in local authority	**-0.002**	**0.001**	**-0.032**	**-3.468**	**0.001**
Satisfaction with living place	-0.002	0.001	-0.017	-1.898	0.058

Adjusted R Square = 0.18; F (11, 10 095) 198.387; p< .000 **CV = -0.032/ -0.087** (= 0.37)

It is worth noting that the measures used here are not exactly of the same type as those analysed in the literature where, for instance, individual earning rather than average income per person in household is used. But, in spite of that the fact that results are consistent with other discussed in the literature confirms usefulness of such an approach, even when public statistics data are used (not necessarily fully comparable with other data).

The second question, concerning the relative impact of social capital vs. income, is also addressed in a simplified version as the former is presented here by positive declaration of trust in local authority. However, there is a substantial 'compensating' effect of the community social capital on individual well-being (acc. to U-index) - living in environment characterized by good relations between residents and public administration is indicative of a possibly cushioning effect for households vulnerable to income shock.

Cross-level relationships depend, however, on both individual and community level factors, and multilevel modelling.

The following model was calculated using also data from EU SILC, with household equivalised disposable income per person as an indicator of individual well-being:

$$\textit{Individual Well-Being}_{H'hld\ eqv.\ disp.\ income} = \gamma_{00} + \gamma_{10}\,\textit{ability}_{\_to\text{-}meet\text{-}ends} + \gamma_{20}\,\textit{time-on-job} + \gamma_{01}\,\textit{MILD} + \gamma_{11}\,\textit{ability}_{\_tme}{*}\textit{MILD}_{j} + \gamma_{21}\,\textit{time-on-job} * \textit{MILD}_{j} + u_{1j}\,\textit{ability\_}_{tme} + u_{2j}\,\textit{time-on-job} + u_{0j} + e_{ij}$$

It is assumed that such a specification of cross-level (between individual and community/gmina measures of well-being, with cross-level *interaction* effect, should ensure robust estimation (e.g. Subramanian. op. cit. p. 521; Hox et al. 2017).

Table 2. Multilevel regression of individual well-being – *household equivalised disposable income per person* – on community and individual level factors with interaction terms.

	Unstand. Coefficients		**Stand. Coeff.**		
Predictors	**B**	**Std. Error**	**Beta**	**t**	**Sign.**
(Constant)	9.687	0.425		22.784	0.000
Ability to ’meet the ends’ (binary)	0.049	0.019	0.041	2.583	0.010
Time on job and commuting	0.028	0.016	0.280	1.698	0.090
MILD /Multidimensional Index of Local Deprivation (2016)	-0.023	0.005	-0.172	-4.199	0.000
Interaction "ability to meet the ends" and local deprivation (MILD2016)	0.003	0.000	0.414	26.174	0.000
Interaction "job-time" and local deprivation (MILD2016)	0.000	0.000	-0.275	-1.688	0.091

Adjusted R Square = .240; F $_{(df\ 5,\ 8496)}$ = 536,381); p< .001

Negative effect of local deprivation (MILD for 2016), both in separation and in interaction with time spend on work – but not with ability to meet the ends, which may offset this effect in better-off households - contrasts with other factors having positive impact on the level of well-being measured here by the household equivalised disposable income. It confirms the role of place and overall quality of the living environment (commune) for individual well-being, which on the other hand significantly depends on such household or person level factors as time spend on work, including commuting.

Spatial autocorrelation and spatial clustering. Moran's I for the below maps (from the left): (a) I=0.20 for local deprivation (MILD); (b) I=0,09 for U-index; (c) I= 0,10 for U-index by MILD

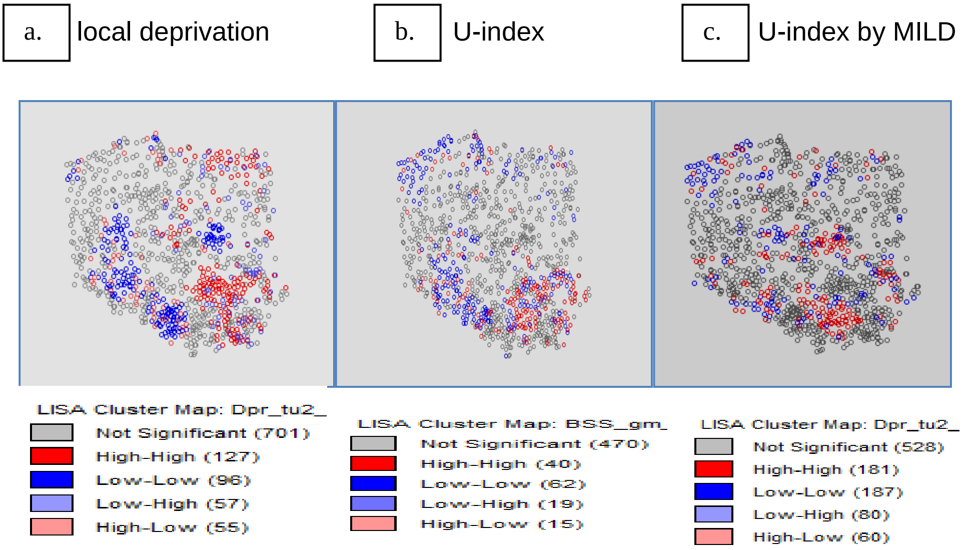


Figure 2. Spatial autocorrelation – Moran's maps

The spatial patterns of local deprivation and subjective well-being (both interpreted in 'negative' terms) show one important feature in common – they both tend to cluster around high or low values of each of these measures in a similar part of the country. In south-east, clusters of high deprived communes (panel a) and also of communes with residents high on the U-scale /'unpleasant state' (panel b) predominate. At the same time, the opposite spatial pattern prevails in the western (especially south-west) part of the country – in communes generally lower on the local deprivation scale live people with a higher level of well-being (lower level of dissatisfaction in the sense of the U-index). The joint spatial distribution of communes (gminas) according to both measures, U-index and MILD, is presented at the panel (c). The overall tendency to spatial concentration is consistent with separately characterized patterns.

It is worth mentioning here the result obtained using an alternative approach, called Functional Data, allowing for taking into account the spatio-temporal property of data and for comparing the spatial patterns of local deprivation (clusters) and subjective well-being for a long-term period (Krzyśko, Okrasa and Wołynski, 2019). The above identified patterns are shown to be even stronger in terms of the autocorrelation coefficient (Moran's I), following the same trends along the East-West geographic axes, and providing useful suggestions for practitioners and decision makers responsible for allocation of public resources for improving both of these areas of concern, i.e. local development and individual well-being.

Table 3. Spatial dependence - spatial regression of Subjective Well-Being on commune's attributes and compositional characteristics

SPATIAL ERROR MODEL – MAXIMUM LIKELIHOOD ESTIMATION

Dependent Variable: ***U –index*** Number of Observations: 937;
Number of Variables: 8; Degrees of Freedom: 929; **Lag coeff. (Lambda): 0.43**;
R-squared: 0.12

Variable	**Coefficient**	**Std.Error**	**z-value**	**Probability**
Constant	0.523731	0.042847	12.2233	0.00000
Monthly income	-0.002730	0.001960	-1.40359	0.16044
Age_avg (%)	-0.014313	0.005653	-2.53177	0.01135 *
Education_hs+ (%)	0.000381	0.000222	1.71849	0.08571 *
Not working pop. (%)	-0.001304	0.000273	-4.77623	0.00000 *
Index of loc.depr.-ecology	0.000560	0.000462	1.21309	0.22510
Index of loc. depr._Soc. Welfare	-0.000415	0.000312	-1.32693	0.18453
Subsidies_pc	1.2323e-005	1.1588e-05	1.06344	0.28758
Lambda	0.431769	0.0677941	6.36883	0.00000

4. Discussion and Conclusion:

Research on individual and community well-being requires data from both individual and community level and both objective and subjective measures in order to explore effectively the relationship in which they remains, and are influenced by such crucial factors as community cohesion, including social capital. As the role of such factors is shown to be important in the local development context, their effects need to be taken into account in the policy about allocation of scarce resources among communes (gminas), especially during the hard time. It might be hopped that communes characterized by a given level of local deprivation but with a higher level of social capita and social cohesion are, on average, less vulnerable to external shocks and are more capable to arrange resources for endogenous, community-based development than others.

Bringing space into analysis gives insight into processes which actually take place on a larger scale than own community – spatial dependency confirms this, suggesting spatio-temporal analytical framework. In particular, for the purpose of rational policy design and evaluation. Individual well-being increases along with greater household income. However, community deprivation reinforces significantly the subjective well-being effect of individual income. Also, deprivation in several domains shows a negative association with U-index (such as risk associated with deprivation in local social welfare).

Working with existing databases, e.g. public files of official statistics, has its advantages and disadvantages, which needs to be recognized to enhance integration procedures in constructing a multi-source analytical database. Nevertheless, geographically referred data provide a promising land of opportunities for policy analysis focused on well-being as the ultimate target of the local development.

REFERENCES

ALKIRE, S., (2015). The Capability Approach and Well-Being Measurement for Public Policy, Oxford Poverty & Human Development Initiative Working Paper No. 94.

ANAND, P., MONTOVANI, I., (2018). The Value of Individual and Community Social Resources, In New Frontiers of the Capability Approach (eds.) F, Fennell S, and Anand, PB Cambridge U. Press.

ARCAYA, M., BREWSTER, M., ZIGLER, C M., SUBRAMANIAN, S. V., (2012). Area variations in health: A spatial multilevel modeling approach, HEALTH PLACE. 2012 JUL, 18(4), pp. 824–831.

CLARK, A. E., (2018). Four Decades of the Economics of Happiness: Where Next? Review of Income and Wealth. Volume 64, Issue 2, https://doi.org/10.1111/roiw.12369

DOLAN, P., LAYARD, R., METCALFE, R., (2011). Measuring Subjective Wellbeing for Public Policy: Recommendations on Measures, 23 Report to the ONS. Special Paper No. 23.

EASTERLIN, R. A., (2010). Happiness and Economic Growth: Does the Cross Section Predict Time Trends? Evidence from Developing Countries [in] Diener E., Kahneman D., Helliwell J., (eds.), International Differences in Well-Being. Published to Oxford Scholarship Online: May 2010,
DOI: 0.1093/acprof:oso/9780199732739.001.0001.

FISCHER, M. M., GETIS, A., (eds.), (2010). Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer.

HELLIWELL, J. F., C. P., BARRINGTON-LEIGH, A., HARRIS, H., HUANG, (2010). “International evidence on the social context of well-being” in E. Diener, J. F. Helliwell and D. Kahneman, eds. International Differences in Well-Being (New York: Oxford University Press).

HONG, G., (2015). Causality in a Social World: Moderation, Mediation and Spill-over, Wiley.

HOX, J. J., MOERBEEK, M., VAN DE SCHOOT, R., (2017). Multilevel Analysis: Techniques and Applications, Third Edition. CRC Francis & Taylor, https://www.crcpress.com/Multilevel-Analysis-Techniques-and-Applications.

KAHNEMAN, D., KRUEGER, A. B., (2006). Developments in the measurement of subjective well-being, Journal of Economic Perspectives, 20, pp. 3–24.

KALTON, G., MACKIE, CH., OKRASA, W., eds., (2015). The Measurement of Subjective Well-Being in Survey Research, Statistics in Transition new series. Vol. 16, No. 3.

KIM, Y., LUDWIGS, K., (2017). Measuring Community Well-Being and Individual Well-Being for Public Policy, In: R. Phillips & C. Wang (eds.), Handbook Of Community Well-Being Research, Springer.

LLOYD, C. D., (2011). Local Models For Spatial Analysis. CRC Press Taylor & Francis Group.

OKRASA, W., (2017). Community well-being, Spatial Cohesion and Individual well-being – towards a multilevel spatially integrated framework, In: W. Okrasa (Ed.) Quality of Life and Spatial Cohesion: Development and well-being in the Local Context, Cardinal Stefan Wyszynski University Press, Warsaw.

OKRASA, W., ROZKRUT, D., (2018). The Time Use Data-based Measures of the Wellbeing Effect of Community Development. Proceedings of the FCSM2018/Federal Committee on Statistical Methodology Research Conference. [in press].

PATUELLI, R., ARBIA, G., (Eds.), (2016). Spatial Econometric Interaction Modelling, Springer.

STIGLITZ, J., A., SEN, J.-P., FITOUSSI, (2009). Report by the Commission on the Measurement of Economic and Social Progress (Paris), www.stiglitz-sen-FITOUSSI.FR

SUBRAMANIAN, S. V., (2010). Multilevel Modeling [in] Fischer and Getis (2019).

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 181–189, DOI 10.21307/stattrans-2019-041
Submitted – 18.01.2018; Paper ready for publication – 16.08.2018

A NEW RATIO ESTIMATOR: AN ALTERNATIVE TO REGRESSION ESTIMATOR IN SURVEY SAMPLING USING AUXILIARY INFORMATION

Mir Subzar[1], S. Maqbool[2], T. A. Raja[2], Prayas Sharma[3]

ABSTRACT

The most dominant problem in the survey sampling is to obtain the better ratio estimators for the estimation of population mean or population variance. Estimation theory is enhanced by using the auxiliary information in order to improve on designs, precision and efficiency of estimators. A modified class of ratio estimator is suggested in this paper to estimate the population mean. Expressions for the bias and the mean square error of the proposed estimators are obtained. Both analytical and numerical comparison has shown the suggested estimator to be more efficient than some existing ones. The bias of the suggested estimator is also found to be negligible for the population under consideration, indicating that the estimator is as good the regression estimator and better than the other estimators under consideration.

Key words: ratio type estimators, auxiliary information, bias, mean square error, simple random sampling, efficiency.

AMS Subject Classification: 62D05

1. Introduction

In sample surveys, auxiliary information on the finite population under study is quite often available from previous experience, census or administrative databases. The sampling literature describes different procedures for using auxiliary information to improve the sampling design and/or obtain more efficient estimators. The use of auxiliary information at the estimation stage has been dealt at great dealt at great length for improving estimation in sample surveys. In sample surveys, auxiliary information is used at selection as well as estimation stages to improve the design as well as obtaining more efficient estimators. Increased precision can be obtained when the study variable Y is highly correlated with auxiliary variable X .

[1] Division of Agricultural Statistics, SKUAST-Kashmir (190025), India. E-mail: subzarstat@gmail.com.

[2] Division of Agricultural Statistics, SKUAST-Kashmir (190025), India.

[3] Department of Decision Sciences, University of Petroleum and Engineering Studies, Dehradun. E-mail: prayassharma02@gmail.com.

Usually, in a class of efficient estimators, the estimators with minimum variance or mean square error is regarded as the most efficient estimator. A good estimator can also be described by the value of its bias. An estimator with minimum absolute bias is regarded as a better estimator among others in the class (Rajesh et al., 2011).

When the population mean of an auxiliary variable is known, so many estimators for the population parameters of study variable have been discussed in literature. The literature on survey sampling describes a great variety of techniques for using auxiliary information by means of ratio, product and regression methods.

If the regression line of the character of interest Y on the auxiliary variable, X is through the origin and when correlation between study and auxiliary variables is positive (high), then the ratio estimate of mean or total may be used (Cochran 1940).

On the other hand, if the regression line used for the estimate does not pass through the origin but makes an intercept along the y-axis, the regression estimation is used (Okafor, 2002). Furthermore, when correlation between study variable on the auxiliary variable passes through a suitable neighbourhood of the origin, in which case, the efficiencies of these estimators are almost equal. When the population parameters of the auxiliary variable X such as population mean, coefficient of variation, coefficient of kurtosis, coefficient of skewness, median are known, a number of modified estimators such as modified ratio estimators, modified product estimators and modified linear regression estimators have been proposed and is widely acceptable in the literature.

This paper is another attempt in solving this problem. An alternative ratio estimator for population mean of the study variable Y (see Sharma and Singh (2014,15), which is more efficient than some of the existing estimators is suggested using the information on one auxiliary variable, X , that is highly correlated with the study variable.

2. Review of the existing estimators

To enhance effective comparison, we summarize below some existing estimators, their biases and mean square errors.

Consider a finite population of N distinct and identifiable units $G=\{G_1,G_2,G_3,...,G_N\}$. Let a sample of size n be drawn from the population by simple random sampling without replacement. Suppose that interest is to obtain a ratio estimate of the mean of a random variable Y from the sample using a related variable X as supplementary information and assuming that the total of X is known from sources outside the survey.

Table 1. Some existing estimators, their biases and mean square errors

S/N	Estimator	Bias	Mean square error
1	$\bar{y}$	0	$\frac{1-f}{n}\bar{Y}^2C_y^2$
2	$\bar{y}_{cl} = \frac{\bar{y}}{\bar{x}}\bar{X}$ Classical ratio	$\frac{1-f}{n}\bar{Y}[C_x^2 - \rho C_x C_y]$	$\frac{1-f}{n}\bar{Y}^2[C_y^2 + C_x^2 - 2\rho C_x C_y]$
3	$\bar{y}_{SK} = \frac{\bar{y}}{\bar{x}+M_d}\bar{X}+M_d$ Subramani and Kumarapandiyan	$\frac{1-f}{n}\bar{Y}[aC_x^2 - \rho a C_x C_y]$	$\frac{1-f}{n}\bar{Y}^2[C_y^2 + C_x^2 a(a-2\theta)]$
4	$\bar{y}_{KC} = \frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}}\bar{X}$ Kadilar and Cingi	$\frac{1-f}{n}\bar{Y}C_x^2$	$\frac{1-f}{n}\bar{Y}^2[C_x^2 + C_y^2(1-\rho^2)]$
5	$\bar{y}_{reg} = \bar{y} + b(\bar{X}-\bar{x})$ Regression estimator	0	$\frac{1-f}{n}\bar{Y}^2C_y^2(1-\rho^2)$

Where

$C_x = \frac{s_x}{\bar{X}};\ C_y = \frac{s_y}{\bar{Y}};$ the coefficients of variation of the auxiliary variable, X and the response variable, Y ;

$\rho = \frac{Sxy}{S_x S_y}$; the correlation coefficient between X and Y ;

$a = \frac{\bar{X}}{\bar{X}+M_d}$, $m = \frac{\bar{X}}{\bar{Y}}$; $B = \frac{S_{xy}}{S_x^2}$; the regression coefficient;

$\theta = \frac{\rho C_y}{C_x}$ and $f = \frac{n}{N}$, where $S_x^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$,

$S_y^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$; the population variances of the auxiliary and study variables respectively;

$S_{xy} = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y});$ the population covariance between X and Y ;

$\bar{X} = N^{-1}\sum_{i=1}^{N} x_i$, $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$; population means of the auxiliary and study variables;

$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$, $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$; sample means of the auxiliary and study variables are respectively defined wherever they appear.

3. Suggested estimator

The proposed ratio estimator is obtained by forming linear combination of Subramani and Kumarapandiyan (2012) and Kadilar and Cingi (2004) estimators as shown below:

$$\bar{y}_{pr} = \frac{\alpha\bar{y}(\bar{X} + M_d)}{\bar{x} + M_d} + \frac{\beta(\bar{y} - b(\bar{X} + \bar{x}))}{\bar{x}}\bar{X} \qquad (1)$$

Such that $\alpha + \beta = 1$

3.1. Bias and mean square error of the proposed estimator

To obtain the approximate expression for the bias and the mean squared error for the proposed ratio estimator, let

$$\bar{x} = \bar{X}(1 + e_x)\,; \quad \bar{y} = \bar{Y}(1 + e_y)\,. \qquad (2)$$

Where

$$e_x = \frac{\bar{x} - \bar{X}}{\bar{X}}\ , \ e_y = \frac{\bar{y} - \bar{Y}}{\bar{Y}}$$

So that,

$$E(e_x) = E(e_y) = 0\ ,\ E(e_x^2) = \frac{(1-f)}{n}C_x^2\ ,\ E(e_y^2) = \frac{(1-f)}{n}C_y^2 \qquad (3)$$

$$E(e_x e_y) = \frac{(1-f)}{n}\rho C_x C_y = \frac{(1-f)}{n}\theta C_x^2$$

Therefore, expressing (1) in terms of (2), we obtain

$$\bar{y}_{pr} = \frac{\alpha\bar{Y}(1+e_y)(\bar{X} + \beta_1)}{\bar{X}(1+e_x) + \beta_1} + (1-\alpha)\left[\frac{(\bar{Y}(1+e_y) + b[\bar{X} - \bar{X}(1+e_x))]}{\bar{X}(1+e_x)}\bar{X}\right]$$

$$= \frac{\alpha\bar{Y}(1+e_y)(\bar{X} + \beta_1)}{(\bar{X} + \beta_1) + \bar{X}e_x} + (1-\alpha)\left[\bar{Y}(1+e_y)(1+e_x)^{-1} - b\bar{X}e_x(1+e_x)^{-1}\right]$$

$$=\alpha\bar{Y}(1+e_y)(1+ae_x)^{-1}+\bar{Y}(1-e_x+e_x^2+e_y-e_ye_x)-\bar{Y}(\alpha-\alpha e_x+\alpha e_x^2+\alpha e_y-\alpha e_ye_x)$$
$$-b\bar{X}e_x+b\bar{X}e_x^2+ab\bar{X}e_x-ab\bar{X}e_x^2$$

By Taylor Series approximation up to order 2, the expression becomes

$$\bar{y}_{pr}=\alpha\bar{Y}(1-ae_x+z^2e_x^2+e_y-we_ye_x)+\bar{Y}(1-e_x+e_x^2+e_y-e_ye_x)$$
$$+\bar{Y}(\alpha e_x-\alpha-\alpha e_x^2-\alpha e_y+\alpha e_ye_x)-B\frac{\bar{X}}{\bar{Y}}e_x+B\frac{\bar{X}}{\bar{Y}}e_x^2+\alpha B\frac{\bar{X}}{\bar{Y}}e_x-\alpha B\frac{\bar{X}}{\bar{Y}}e_x^2$$

$$=\bar{Y}[\alpha-\alpha ae_x+\alpha a^2e_x^2+\alpha e_y-\alpha we_ye_x+1-e_x+e_x^2+e_y-e_ye_x-\alpha+\alpha e_x$$
$$-\alpha e_x^2-\alpha e_y+\alpha e_ye_x-B\frac{\bar{X}}{\bar{Y}}e_x+B\frac{\bar{X}}{\bar{Y}}e_x^2+\alpha B\frac{\bar{X}}{\bar{Y}}e_x-\alpha B\frac{\bar{X}}{\bar{Y}}e_x^2]$$

The expression for the bias of this estimator to first order approximation is obtained as follows:

$$B(\bar{y}_{pr})=E(\bar{y}_{pr}-\bar{Y})$$

$$=E[\bar{Y}(1+e_y+(\alpha BM-BM-\alpha-\alpha a)e_x+(\alpha-\alpha a-1)e_ye_x+(\alpha a^2+1-\alpha+BM-\alpha BM)e_x^2-\bar{Y})]$$

$$=\frac{1-f}{n}\bar{Y}[(\alpha-\alpha a-1)\rho C_yC_x+(\alpha a^2+1-\alpha+BM-\alpha BK)C_x^2 \qquad (4)$$

$$MSE(\bar{y}_{pr})=E(\bar{y}_{pr}-\bar{Y})^2$$

$$=E[\bar{Y}(1+e_y+(\alpha BM-BM-\alpha-\alpha a)e_x+(\alpha-\alpha a-1)e_ye_x+(\alpha a^2+1-\alpha+BM-\alpha BM)e_x^2-\bar{Y})]^2$$

$$=\frac{1-f}{n}\bar{Y}^2[C_y^2+2(\alpha BM-BM-\alpha-\alpha a)\rho C_yC_x+(\alpha BM-BM-\alpha-\alpha a)^2C_x^2]$$

(5)

3.2. Optimal conditions for the proposed estimator

To obtain the value of α that minimizes the MSE, we take partial derivative of equation (5) with respect to α and equate to zero as follows:

$$\frac{\partial MSE(\bar{y}_{pr})}{\partial\alpha}=\frac{1-f}{n}\bar{Y}^2[C_y^2+2(\alpha BM-BM-\alpha-\alpha a)\rho C_yC_x+(\alpha BM-BM-\alpha-\alpha a)^2C_x^2]=0$$

$$\Rightarrow \rho C_xC_y+\alpha(BM+1-a)C_x^2-(BM+1)C_x^2=0$$

$$\Rightarrow \alpha=\frac{(BM+1)C_x^2-\rho C_xC_y}{(BM+1-a)C_x^2} \qquad (6)$$

Substituting for (6) in (5) gives the optimal MSE for $\bar{y}_{pr}$ as:

$$MSE(\bar{y}_{pr}) = \frac{1-f}{n}\bar{Y}^2[C_y^2(1-\rho^2)] \tag{7}$$

4. Efficiency comparison

In order to compare the efficiency of the various existing estimators with that of proposed estimators, we require the expressions of mean square error of these estimators, up to first order approximation. An analytical comparison of the proposed estimator with three of the existing estimators namely: the classical, Subramani and Kumarapandiyan (2012) and Kadilar and Cingi (2004) estimators are carried out.

4.1. Efficiency comparison of proposed and classical

In this section, the analytical condition under which the proposed estimator will be more efficient than classical ratio estimator is established.

$$\begin{aligned} MSE(\bar{y}_{pr}) - MSE(\bar{y}_{cl}) &= \frac{1-f}{n}\bar{Y}^2[C_y^2(1-\rho^2)] - \frac{1-f}{n}\bar{Y}^2[C_y^2 + C_x^2 - 2\rho C_x C_y] \\ &= \frac{1-f}{n}\bar{Y}^2(C_y^2 - C_y^2\rho^2 - C_y^2 - C_x^2 + 2\rho C_x C_y] \\ &= \frac{1-f}{n}\bar{Y}^2(2\rho C_x C_y - C_x^2 - C_y^2\rho^2) \\ &= -\left[\frac{1-f}{n}\bar{Y}^2(C_y\rho - C_x)^2\right] \end{aligned} \tag{8}$$

Since the expression in the square bracket is always positive, we conclude that the proposed estimator will always be more efficient than the classical ratio estimator.

4.2. Efficiency comparison of proposed and Subramani and Kumarapandiyan

$$\begin{aligned} MSE(\bar{y}_{pr}) - MSE(\bar{y}_{SK}) &= \frac{1-f}{n}\bar{Y}^2[C_y^2(1-\rho^2)] - \frac{1-f}{n}\bar{Y}^2[C_y^2 + C_x^2 a(a-2\theta)] \\ &= \frac{1-f}{n}\bar{Y}^2[C_y^2 - C_y^2 - C_x^2 a\left(a - \frac{2\rho C_y}{C_x}\right)] \\ &= \frac{1-f}{n}\bar{Y}^2[-C_y^2\rho^2 - C_x^2 a\left(a - \frac{2\rho C_y}{C_x}\right)] \end{aligned}$$

$$= -\left\{\frac{1-f}{n}\bar{Y}^2[C_x^2\rho^2 + C_x^2 a(a-2\theta)]\right\} \quad (9)$$

Therefore, for the proposed estimator to be more efficient than Yan and Tian (2010), the terms in the second bracket must be positive. This implies that:

$$C_x^2\rho^2 + C_x^2 a(a-2\theta) > 0 \quad (10)$$

4.3. Efficiency comparison of proposed and Kadilar and Cingi

$$MSE(\bar{y}_{pr}) - MSE(\bar{y}_{KC}) = \frac{1-f}{n}\bar{Y}^2[C_y^2(1-\rho^2)] - \frac{1-f}{n}\bar{Y}^2[C_x^2 + C_y^2(1-\rho^2)]$$

$$= -\left[\frac{1-f}{n}\bar{Y}^2 C_x^2\right] \quad (11)$$

Since the expression in the square bracket of equation (11) is always positive, it therefore means that the proposed estimator will always be more efficient than Kadilar and Cingi (2004) estimator of population mean.

5. Numerical comparison

In this section, to study the performance of the estimator presented in this work, we consider empirical population. The source of the population is Singh and Chaudhary (1986) and the values of requisite population parameters are given. We compare the efficiency of the proposed estimator with the existing estimators using the known population data.

Table 2. Data Statistics for population

Parameters	Population	Parameters	Population
N	34	C_x	0.7531
n	20	Md	150
$\bar{Y}$	856.4117	β_1	1.1823
$\bar{X}$	199.4412	$\theta = \rho(C_y/C_x) = BU$	0.50620
ρ	0.4453	$a = \bar{X}/\bar{X} + Md$	0.58204
S_y	733.1407	$M = \bar{X}/\bar{Y}$	0.23288
C_y	0.8561	$B = \rho S_y/S_x$	2.17333
S_x	150.2150	$R = \bar{Y}/\bar{X}$	4.29406

Table 3. Estimators, biases, MSE and % relative efficiency for population.

Population			
Estimator	**MSE**	**Bias of the estimator**	**% Relative efficiency**
$\bar{y}_{cl}$	12557.99	5.658	100.00
$\bar{y}_{SK}$	10236.38	3.293	122.67
$\bar{y}_{KC}$	19977.62	11.457	62.86
$\bar{y}_{pr}$	10165.43	-0.069	123.53

6. Discussion

Optimal mean square error (MSE) of the proposed estimators given in Equation (7) has the same expression as the MSE of the regression estimator which is known to be more efficient than the ratio and the product estimators. The comparison of the suggested estimator with the three existing estimators are derived analytically and these comparisons show that the suggested estimators are more efficient than the classical ratio (1940), Kadilar and Cingi (2004) estimators and preferred over the Subramani and Kumarapandiyan (2012) estimator when the condition stated in the equation (10) is satisfied.

From empirical study, results in the Table 3 reveals that our suggested estimators has lower mean square error than the classical ratio (1940), Kadilar and Cingi (2004) and Subramani and Kumarapandiyan (2012) in the population under consideration, showing that the suggested estimator is more efficient than all the other estimators under consideration. This due to the fact that the suggested estimator is equally as efficient as the regression estimator and confirms Cochran (1940), Robson (1957), Murthy (1967) and Perri (2005) assertion that the regression estimator is generally more efficient than the ratio and product estimators.

Analyses of biases have also shown that the suggested estimator have smallest bias than the all other estimators under consideration. From the Table 3, also from bias point of view, bias is negligible and agrees with the assertion of the Okafor (2002) that any estimator with relative bias less than 10% is considered to have a negligible bias.

7. Conclusion

Since the from the equation (7) the suggested estimator gives the same precision as the regression estimator and is consistently better in terms of bias and efficiency then the three estimators under consideration, the suggested estimator can always be used as an alternative to the regression estimator and gives a better replacement to some existing ratio estimators.

REFERENCES

COCHRAN, W. G., (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. Journal of Agricultural Science, 30, pp. 262–275.

KADILAR, C., CINGI, H., (2004). Ratio estimators in simple random sampling. Applied Mathematics and Computation, 151, pp. 893–902.

MURTHY, M. N., (1967). Sampling theory and methods, Statistical Publishing Society, Calcutta, India.

SINGH, D., CHAUDHARY, F. S., (1986). Theory and Analysis of Sample Survey Designs, New Age International Publisher.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012). Estimation of population mean using known median and co-efficient of skewness. American Journal of Mathematics and Statistics, 2(5), pp. 101–107.

ROBSON, D. S., (1957). Application of multivariate Polykays to the theory of unbiased ratio type estimation. Journal of American Statistical Association, 52, pp. 411–422.

OKAFOR, F. C., (2002). Sample Survey Theory with Applications (1st ed.), N sukka, Nigeria, Afro-Orbis.

PERRI, P. F., (2005). Combining two Auxiliary Variables in Ratio-cum-product type Estimators. Proceedings of Italian Statistical Society. Intermediate meeting on Statistics and Environment, Messina, 21-23 September, pp. 193–196.

RAJESH, T., RAJESH, P., JONG-MIND, K., (2011). Ratio-cum-product Estimators of Population mean using known Parameters of Auxiliary Variables. Communications of the Korean Statistical Society, 18(2), pp. 155–164.

SHARMA, P., SINGH, R., (2014). Improved Ratio Type Estimators Using Two Auxiliary Variables under Second Order Approximation, Mathematical Journal of Interdisciplinary Sciences, Vol. 2, No. 2, pp. 193–204.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 191–194

ABOUT THE AUTHORS

Bilíková Mária is an Associate Professor in the Department of Mathematics and Actuarial Science of the University of Economics in Bratislava. Her research work focuses on actuarial mathematics and model applications in the field of insurance, most of all life insurance. Her main area of research is the stochastic modelling of the various risks to which an insurance company is exposed and their subsequent use to value an insurance company's portfolio within the context of the requirements of the EU Solvency II directive. She has published more than 90 research papers in international/national journals and conferences.

Christian Heumann is a Professor at the Department of Statistics in Ludwig-Maximilians-Universität München (LMU Munich), Germany. His research interests are analysis of data with missing values, especially multiple imputation, model selection, model averaging and models for natural language processing. Professor Heumann has published six books/mongraphs and more than 50 research papers in international journals.

Filas-Przybył Sylwia graduate of Mathematics at Adam Mickiewicz University in Poznań (AMU). Since 2006 she has been a head of Centre for Urban Statistics of Statistical Office in Poznań – a unit which prepared the methodology and then published the results of three editions of study on people commuting to work (2006, 2011, 2016). She has taken part in several key domestic and international projects focused on the challenges of Polish public statistics related to getting access and then processing spatial data on selected socio-economic phenomenon (i.a. Urban Audit, Merging statistical data and geospatial information in Member States, Employing the EU methodology to define labour market areas in Poland, Identification of special areas inside capitals of regions and for their functional areas taking into account socio-demographic situation of their inhabitants based on spatial GIS driven analysis and Students commuting to schools located in provincial cities).

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 Centre of Mathematical Statistics at the regional Statistical Office in Lodz. Her main areas of interest include: income distribution, income inequality and poverty measurement, small area estimation. Currently she is a member of two editorial boards: Statistica & Applicazioni – international statistical journal published in Italy and Folia Oeconomica Acta Universitatis Lodziensis.

Klimanek Tomasz graduate of Poznan University of Economics (now Poznań University of Economics and Business PUEB). Between 1995 and 2014 he worked at the Department of Statistics of PUEB. Since 2009 he has been Deputy Director of the Statistical Office in Poznan. He has taken part in several domestic and international projects aimed at applying new methodologies in official statistics (e.g. EURAREA, ESSnet on Small Area Statistics, ESSnet on Data Integration, MEETS, Poverty Mapping for Poland, Identification of special areas inside capitals of regions and for their functional areas taking into account socio-demographic situation of their inhabitants based on spatial GIS driven analysis and Students commuting to schools located in provincial cities). His main fields of interest are small area estimation, statistical disclosure control, spatial analyses in official statistics and graph data.

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. The Polish Statistician" (Wiadomości Statystyczne).

Maqbool S., is Assistant professor (Statistics) since 2007 at Sher-e- Kashmir University of Agricultural Sciences and Technology-Kashmir (J & K), India. He has published more than 120 research papers in reputed international and national journals. The area of specialization is sampling theory and mathematical programming and attended more than 30 international conferences.

Mishra Madhulika is a research scholar at the Department of Statistics, Banaras Hindu University,Varanasi,India. She is perusing her Ph.D. under the supervision of Prof. Rajesh Singh in the same department. Her research interests are sampling surveys, statistical Inference, demography and data analysis. She has published more than 7 papers in national and international journals and conferences. She is also a member of various statistical societies.

Okrasa Włodzimierz is a Professor and a Head of the Research Methods and Evaluation Unit at the Institute of Sociology, Cardinal Stefan Wyszynski University in Warsaw . He serves as an Advisor to the President of the Statistics Poland and an Editor in Chief of the Statistics in Transition new series. He was teaching and researching in Polish and American universities and was an ASA Senior Research Fellow at the US Bureau of Labor Statistics (1990–1991), a Program Director in the Social Science Research Council, N. Y. (1991–1993), and then worked for the World Bank in Washington, D. C. (1994–2000), analyzing poverty and implementing new household surveys in several 'countries in transition'. He next headed the Social Sciences Unit at the European Science Foundation (2000–2003, in Strasbourg). Elected member of the International Statistical Institute and V-President of the Polish Statistical Association. Author of numerous publications, including books and articles in reputed international journals.

Polko-Zając Dominika works as an assistant in Department of Statistics, Econometrics and Mathematics in University of Economics in Katowice, Poland. Currently she is working on PhD in the field of the methods of comparing populations in economics research. Her main research interest are permutation tests.

Rabe Anasu is currently a Lecturer with the Department of Mathematics and Statistics, Umaru Musa Yaradua University Katsina, Nigeria. He recently obtained a PhD in Statistics from the University of Botswana, UB. His research interests include Statistical modelling of multivariate observations and Linear mixed models, with emphasis on modelling the covariance structure. He is a member of Professional Statisticians Society of Nigeria (PSSN).

Raja Tariq A., is working as Senior Associate Professor (Statistics) at Sher-e-Kashmir University of Agricultural Sciences and Technology-Kashmir (J & K), India. He has published one hundred three research papers in various peer reviewed international and national journals and has two manuals and a book to his credit. Besides being in the advisory committee of more than two hundred PG/Ph.D Scholars, one student has been awarded Ph.D and two are at the verge of completion of their Ph.D. under his guidance. He has chaired six technical sessions and presented twenty-eight research papers in National and International Conferences. He has completed two research Projects and has organized three training programmes at National level. He has received three awards and is executive member of various National and International societies.

Razzak Humera is a research scholar at the Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany. Her main areas of interest include methods for missing data, multivariate data analysis and R programming.

Rozkrut Dominik has been President of Statistics Poland since 2016. He is also a member of the European Statistical System Committee, where has been nominated for the position of Chair of the ESS Partnership Group for 2020–2021. In the framework of his extensive international activity, he also serves as a member of the OECD's Committee on Statistics and Statistics Policy and a bureau member of the UNECE Conference of European Statisticians. Moreover, he represents Poland on the UN Statistical Commission, and has been involved in the work of the ESS Vision Implementation Group, the UN Friends of the Chair Group on the Fundamental Principles of Official Statistics, the UN Global Working Group on Big Data for Official Statistics and the UNECE High-Level Group on Modernisation of Official Statistics. Until 2019, Dr Rozkrut was also the member of the EU Business-to-Government (B2G) Data Sharing expert group.

Dominik Rozkrut's scientific career began at the Department of Statistics and Econometrics at the University of Szczecin. At that time, he worked in the private sector as a member of a market risk management team at a commercial bank. He joined official statistics in 2007, where he served as Director of the regional branch of Statistics Poland, having supervised numerous research and development projects. He is assistant professor at the University of Szczecin, where he received his Ph.D. in statistics and econometrics. He completed an internship at the University of Massachusetts, Lappeenranta University of Technology and UNU-MERIT. Dr. Rozkrut is a member of the Main Council of the Polish Statistical Association.

Sharma Prayas is an Assistant Professor at the Department of Decision Sciences, School of Business, University of Petroleum and Energy Studies, Derhadun. His research interest are Sampling theory, Estimation, Statistical Modeling and Data Analysis in particular. Dr. Sharma has published more than 30 research papers in international/ national journals along with two chapters & one book. Dr. Sharma is member of more than 10 editorial boards including

Investigación Operacional (Cuba), Journal of Reliability and Statistical Studies (India), Cambridge Scholars Publishing (United Kingdom), etc. He is also serving as reviewer for more than 20 reputed journals including Journal of Statistical Theory and Practice, Hacettepe Journal of Mathematics and Statistics, Clinical Epidemiology and Global Health, Statistics and Transition new Series, International Journal of Applied and Computational Mathematics etc. Dr. Sharma is also engased in conselting the private organizations & corpotrates.

Šoltés Erik is an Associate Professor at the Department of Statistics in University of Economics in Bratislava, Slovakia. His research work is focused on applying the theory of credibility in non-life insurance and on the analysis of socio-economic phenomena using various statistical methods such as statistical inference, regression and correlation analysis, general linear models, generalized linear models and multivariate statistical methods (including mainly Principal component analysis, Factor analysis, Cluster analysis and Correspondence analysis). Associate Professor Šoltés has published more than 80 research papers in international/national journals and conferences. He is also the author or the co-author of 5 monographs, 5 textbooks and 5 study manuals.

Subzar Mir is a research scholar in the Division of Agricultural Statistics, Sher Kashmir University of Agricultural Science and Technology - Kashmir, Shalimar, India. His area of research is Sampling Theory and Statistical Inference. He has published 18 research articles in reputed National and international journals of, Agricultural Statistical Sciences, statistical and mathematical sciences. Mir also presented some reputed articles in national and international conferences.

Wesołowski Jacek is a Professor at the Faculty of Mathematics and Information Science of the Warsaw University of Technology and a chief specialist at the Department of Survey Coordination of Statistics Poland. His main research interests are: independence and conditional structures of probability measures and stochastic processes; random matrices; asymptotical statistics and mathematics of survey methodology. He has a well established international scientific collaboration with leading researchers from many countries; e.g. the USA, Canada, France, Spain and Taiwan. Professor Wesołowski is an author of more than 150 papers mostly published in internationally recognized journals such as Annals of Probability, Annals of Statistics, Survey Methodology, Probability Theory and Related Fields, Transactions of the American Mathematical Society, Bernoulli, Journal of Statistical Physics, Journal of Multivariate Analysis, Studia Mathematica and many others. He is also a co-author of a Springer monograph Symmetric Functionals on Random Matrices and Random Matchings Problems.

Zelinová Silvia is a doctoral student in the Department of Mathematics and Actuarial Science of the University of Economics in Bratislava, Slovakia. Her supervisor is Mária Bilíková, Associated professor in the same department. She is doing research on the new standard IFRS 17 and its impact on the work of actuaries in insurance companies. The new standard IFRS 17 requires extensive changes in the accounting, valuation, presentation and disclosure of information about life and non-life insurance and reinsurance contracts. The research is focused on creating applications in accordance with the new standard.

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 195–198

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 2018 to 30 November 2019 as peer-reviewers of manuscripts for the ***Statistics in Transition new series – Volume 20, Numbers 1–4***; the authors` work has benefited from their feedback.

Adebola Solarin Sakiru, Centre for Globalization and Sustainability Research, Multimedia University, Malaysia

Agrawal Bharti, Department of Statistics and Economics, Shri Vaishnav Institute of Management, India

Ahmad Munir, National College of Business Administration & Economics, Pakistan.

Ahmed Shakeel, Department of Statistics, Quaid-i-Azam University, Pakistan

Alao Felix Ilesanmi, Department of Mathematical Sciences, University of Mines and Technology, Ghana

Asar Yasin, Department of Mathematics-Computer, Necmettin Erbakan Üniversitesi, Turkey

Andersen Per Kragh, Department of Biostatistics, University of Copenhagen, Denmark

Awe Olawale O., Department of Mathematical Sciences, Anchor University Lagos, Nigeria

Bednarski Tadeusz, Institute of Economics, University of Wroclaw, Poland

Budnikas Germanas, University of Bialystok, Poland

Bwanakare Second, University of Information Technology and Management (WSIZ) in Rzeszow, Poland

del Mar Rueda García Maria, Department of Statistics, University of Granada, Spain

Dey Sanku, Department of Statistics, King Abdulaziz University, Saudi Arabia

Domański Czesław, Department of Statistical Methods, University of Lodz, Poland

Dudek Andrzej, Department of Statistical Methods, Wrocław University of Economics, Poland

Escabias Machuca Manuel, Department of Statistics, University of Granda, Spain

Espejo Mariano Ruiz, Department of Health Sciences, The Catholic University of Saint Anthony (UCAM), Spain

Fasoranbaku Olusoga Akin, The Federal University Of Technology Akure, Nigeria

Filipiak Katarzyna, Institute of Mathematics, Poznan University of Technology, Poland

Fischer Mischa, Institute for Social Research, Survey Research Center, University of Michigan, USA

Frenkel Ilia B., Chair - Center for Reliability and Risk Management, Israel

Friedrich Peter, Institute of Economics, University of Tartu, Estonia

Fu-Kwun Wang, Department of Industrial Management, National Taiwan University of Science and Technology, Taiwan

Gamrot Wojciech, Department of Statistics, University of Economics in Katowice, Poland

Gatnar Eugeniusz, Department of Economic Analysis and Finance, University of Economics in Katowice and The National Bank of Poland, Poland

Getka-Wilczynska Elzbieta, Institute of Econometrics, Warsaw School of Economics | SGH, Poland

Gupta Sat, Department Head Fellow of the American Statistical Association, University of North Carolina at Greensboro, USA

Hasegawa Hikaru, Department of Economics, Hokkaido University, Japan

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, USA

Kalton Graham, Westat, USA

Kittaneh Omar, Effat University, Israel

Kordos Jan, former Professor of Warsaw School of Economics, Poland

Kozek Andrzej, Department of Mathematics and Statistics, Macquarie University, Australia

Král' Pavol, Department of Quantitative Methods and Information Systems, Matej Bel University in Banská Bystrica, Slovakia

Krzyśko Mirosław, Faculty of Probability and Mathematical Statistics, Adam Mickiewicz University, Poland

Lapiņš Janis, Statistics Department, Bank of Latvia, Latvia

Lawson Nuanpan, Department of Applied Statistics, King Mongkut's University of Technology North Bangkok, Thailand

Lehtonen Risto, University of Helsinki, Finland

Lencina Viviana Beatriz, Faculty of Economics, Statistical Research Institute, National University of Tucumán, Argentina

Lesaoana Maseka, Department of Statistics and Operations Research, University of Limpopo, RPA

Lukman Adewale F., Landmark University, Nigeria

Machado J. A. Tenreiro, Polytechnic Institute of Porto, Portugal

Madigu Godfrey, Centre for Applied Research in Economics, Strathmore University, Kenya

Marchetti Stefano, Department of Economics and Management, University of Pisa, Italy

Marion Glenn, Biomathematics and Statistics Scotland, Process and Systems Modelling, Scotland

Młodak Andrzej, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland

Münnich Ralf, University of Trier, Germany

Okrasa Włodzimierz, University of Cardinal Stefan Wyszyński in Warsaw, and Statistics Poland, Poland

O'Muircheartaigh Colm A., University of Chicago, USA

Osaulenko Oleksandr H., National Academy of Statistics, Accounting and Audit, National Academy of Statistics, Kiev, Ukraine

Ozgul Nilgun, Department of Statistics, Hacettepe University, Turkey

Pacakova Viera, Institute of Mathematics and Quantitative Methods, University of Pardubice, Czech Republic

Panek Tomasz, Institute of Statistics & Demography, Warsaw School of Economics, Poland

Paradysz Jan, University of Economics in Poznan, Poland

Pekasiewicz Dorota, Department of Statistical Methods, University of Lodz, Poland

Piontek Krzysztof, Department of Financial Investments and Risk Management, Wroclaw University of Economics, Poland

Pociecha Józef, Department of Statistics, Cracow University of Economics, Poland

Reddy Krishna, Faculty of Management, Osmania University, India

Retkute Renata, The Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research (SBIDER), University of Warwick, Great Britain

Riaz Saba, Department of Mathematics and Statistics, Riphah International University, Pakistan

Rozkrut Dominik, President of Statistics Poland, Poland

Rubil Ivica, The Institute of Economics, Croatia

Shittu Olanrewaju Ismail, Faculty of Science, University of Ibadan, Nigeria

Shukla Upasana, Department of Statistics, University of Pretoria, South-Africa

Shy-Der Lin, Department of Applied Mathematics Chung, Yuan Christian University, Taipei

Singh Poonam, Department of Statistics, Banaras Hindu University, India

Smaga Łukasz, Faculty of Mathematics and Computer Science, Adam Mickiewicz University | UAM, Poland

Sokołowski Andrzej, Department of Statistics Cracow University of Economics, Poland

Šoltés Erik, Department of Statistics, University of Economics in Bratislava, Slovakia

Šoltesova Tatiana, University of Economics in Bratislava, Slovakia

Strahl Danuta, Wrocław University of Economics, Branch in Jelenia Góra, Poland

Suchecka Jadwiga, University of Lodz, Poland

Sznajder Roman, Department of Mathematics, Bowie State University, USA

Szreder Mirosław, University of Gdansk, Poland

Sztaundynger Jan Jacek, Department of Econometrics, University of Lodz, Poland

Tarczyński Waldemar, University of Szczecin, Poland

Tiensuwan Montip, Department of Mathematics, Mahidol University, Thailand

Van Hoa Tran, Vietnam and East Asia Summit Research Program, Victoria University, Australia

Verma Hemant K., Registrar General of India, Patna, India

Walesiak Marek, Department of Econometrics and Computer Science, Wrocław University of Economics, Poland

Wanat Stanisław, Department of Mathematics, Cracow University of Economics, Poland

Wolter Kirk, Principal Statistical Advisor and Senior Fellow, NORC, University of Chicago, USA

Wołyński Waldemar, Faculty of Mathematics and Computer Science, Adam Mickiewicz University, PolandŽiberna Aleš, Centre for Methodology and Informatics, University of Ljubljana, Slovenia

Zmyślony Roman, Faculty of Mathematics, Computer Science and Econometrics, University of Zielona Góra, Poland

Żądło Tomasz, Department of Statistics, Econometrics and Matematics, University of Economics in Katowice, Poland

STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 199–202

INDEX OF AUTHORS, VOLUME 20, 2019

Abduljaleel M., *Outlier detection in the analysis of nested Gage R&R, random effect model*

Abu Awwad R. R., *Statistical inference of exponential record data under Kullback-Leiber divergence measure*

Abufoudeh G. K., *see under Abu Awwad R. R.*

Adegoke H. M., *see under Yaya O. S.*

Adepoju A. A., *Economic growth and its determinant: a cross-country evidence*

Akintande O. J., *see under Yaya O. S.*

Arcos A., *see under Arnab R.*

Arnab R., *Nonrandomized response model for complex survey designs*

Bdair O. M., *see under Abu Awwad R. R.,*

Ben-Hur D., *see under Pfeffermann D.*

Bilíková M., *see under Šoltes E.*

Blum O., *see under Pfeffermann D.*

Budny K., *Power generalization of Chebyshev's inequality – multivariate case*

Cibulková J., *The effect of binary data transformation in categorical data clustering*

Chaturvedi A., *Generalized Bayes Estimation of Spatial Autoregressive Models*

Dalei N. N., *see under Sharma A.*

Das K. K., *see under Dhar S.*

Dehnel G., *A comparative analysis of economic efficiency of medium-sized manufacturing enterprises in districts of Wielkopolska province using the hybrid approach with metric and interval-valued data*

Dhar S., *Formulation of the simple Markovian model using fractional calculus approach and its application to analysis of queue behaviour of severe patients*

Dorugade A. V., *Efficient two-parameter estimator in linear regression*

Filas-Przybył S., *see under Klimanek T.*

Górecki T., *Variable selection in multivariate functional data classification*

Heumann Ch., *see under Razzak H.*

Ikoba N. A. *Modelling Language Extinction Using Susceptible-Infectious-Removed (SIR) Model*

Intarapak S., *An alternative matrix transformation to the F test statistic for clustered data*

Janusz B., *Application of the strategy combining monetary unit sampling and a Horvitz-Thompson estimator of error amount in auditing – results of a simulation study*

Jędrzejczak A., *Estimation of income characteristics for regions in Poland using spatio-temporal small area models*

Jolayemi E. T., *see under Ikoba N. A.*

Jurkiewicz T., *see under Wycinka E.,*

Kalyani K., *see under Sivakumar D. C. U.*

Kanaparthi R., *see under Sivakumar D. C. U.*

Karimi M., *see under Abduljaleel M.*

Khare B. B., *Estimation of product of two population means by multi-auxiliary characters under double sampling the non-respondents*

Klimanek T., *The impact of the applied typology on the statistical picture of population ageing in Polish towns - a comparative analysis*

Komorowska O., *Comparative analysis of poverty in families with a disabled child and families with non-disabled children in Poland in the years 2014 and 2016*

Kozioł A., *see under Zmyślony R.*

Kozłowski A., *see under Komorowska O.*

Krzyśko M., *see under Górecki T.*

Kubacki J., *see under Jędrzejczak A.*

Kumar D., *Generalized Pareto distribution based on generalized order statistics and associated inference*

Landmesser J. M., *Decomposition of gender wage gap in Poland using counterfactual distribution with sample selection*

Leśkow J., *An application of functional data analysis to local damage detection*

Mahanta L. B., *see under Dhar S.*

Malik M. R., *see under Kumar D.*

Maqbool S., *see under Subzar M.*

Midi H., *see under Abduljaleel M.*

Mishra M., *see under Singh R.*

Mishra S., *see under Chaturvedi A.*

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).