STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 18, Number 3, September 2017

ISSN 1234-7655

Editorial Office
Marek Cierpiał-Wolan, Ph.D., Scientific Secretary
m.wolan@stat.gov.pl
Secretary:
Patryk Barszcz, P.Barszcz@stat.gov.pl
Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec, Technical Assistant
Address for correspondence
GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 369−371, DOI 10. 21307

FROM THE EDITOR

This issue is composed of eleven articles grouped in four sections, as follows: the first group of papers, devoted to *sampling methods and estimation* issues consists of three papers; the next four papers constitute the section of *research articles*, followed by *other articles* (containing just one conference paper), and the whole issue closes with *research communicates and letters* containing two papers of the nature of 'work in progress'. They are briefly characterized below.

In the first paper, ***Population Variance Estimation Using Factor Type Imputation Method,*** **Ranjita Pandey** and **Kalpana Yadav** propose a variance estimator based on factor type imputation in the presence of non-response. They describe properties of the new estimators along with their optimality conditions. The proposed classes of factor type ratio estimators are shown to be more efficient than some of the existing estimators – such as the usual unbiased estimator of variance, ratio-type, dual to ratio type and ratio cum dual to ratio estimators. Their performances in terms of relative efficiencies are illustrated with simulated and real data sets. In particular, one of the proposed estimators is shown to perform best from the point of view of increasing efficiency (but all the three proposed FT type estimators are the best estimators in the sense of having the largest PRE).

Krzysztof Beck's paper, ***Bayesian Model Averaging and Jointness Measures: Theoretical Framework and Application to the Gravity Model of Trade*** discusses the Bayesian model averaging (BMA) along with the benefits due to combining the knowledge generated through the analysis of different models. The BMA structure is described together with its most important statistics (prior parameter proposals, prior model size distributions, and also the jointness measures). Its application is illustrated with the gravity model of trade, where determinants of trade are chosen from the list of nine different variables. It enabled the identification of four robust determinants: geographical distance, real GDP product, population product and real GDP per capita distance. All variables, except for population product, have coefficient signs predicted by the theory. For instance, the complementary relationship between real GDP product and population product allowed one to explain the negative sign of the population product coefficient.

In the paper ***On Asymmetry of Prediction Errors in Small Area Estimation,*** **Tomasz Żądło** starts with an observation that the mean squared error (MSE), which reflects only the average prediction accuracy, is insufficient and even inadequate as a measure of overall quality since we are interested not only in the

average but in the whole distribution of prediction errors. Therefore, the author proposes to use an alternative measure of prediction accuracy in the context of small area estimation, taking into account a modified version of the empirical best predictor based on a generalization of the predictor presented by Molina and Rao (2010). The generalization results from the assumption of a longitudinal model and possible changes of the population and subpopulations in time. The considerations are supported by results of the real data application.

The second part of this issue begins with an article by **Mirosław Krzyśko** and **Łukasz Smaga**, entitled ***An Application of Functional Multivariate Regression Model to Multiclass Classification***. The authors propose the scale response functional multivariate regression model based on possible functions representation of functional predictors and regression coefficients. The proposed functional multivariate regression model is employed to multiclass classification for multivariate functional data. Computational experiments performed on real data sets demonstrate the effectiveness of the proposed method for classification for functional data.

Tomasz Górecki's and **Maciej Łuczak's** article on ***Stacked Regression with a Generalization of the Moore-Penrose Pseudoinverse*** is devoted to the problem of making an optimal selection among available methods of classification. The authors propose a combined method that allows one to consolidate information from multiple sources in a better classifier. They discuss the stacked regression (SR) as a way of forming linear combinations of different classifiers toward improved accuracy of classification through employing the Moore-Penrose (MP) pseudoinverse to find the solution to a system of linear equations. Due to the computational difficulty with a greater number of features, they propose a genetic approach to handle the problem. Experimental results on various real data sets demonstrate that the improvements are efficient and that this approach outperforms the classical SR method, providing a significant reduction in the mean classification error rate.

In the next article, ***An Additive Risks Regression Model for Middle-Censored Lifetime Data,*** **P. G. Sankaran** and **S. Prasad** discuss the middle-censoring data problem arising in situations where the exact lifetime of study subjects becomes unobservable, and whether it happens to fall in a random censoring interval. The authors propose a semiparametric additive risks regression model for analysing middle-censored lifetime data arising from an unknown population. They estimate regression parameters and the unknown baseline survival function by two different methods – the first method uses the martingale-based theory, and the second method is an iterative method. The finite sample behaviour of the estimators is assessed through simulation studies, and the utility of the model with a real life data set is demonstrated in the conclusions.

The article by **Luboš Marek, Stanislava Hronová** and **Richard Hindls,** ***Option for Predicting the Czech Republic's Foreign Trade Time Series as Components in Gross Domestic Product,*** analyses the time series data for the

foreign trade of the Czech Republic (CR), and the issue of predictions in such series using the SARIMA and transfer-function models. The authors' goal is to propose models suitable for describing the time series of the exports and imports of goods and services from/to the CR and to subsequently use these models for predictions in quarterly estimates of the gross domestic product's component resources and utilization. They suggest a class of models with time lag as suitable, allowing for making predictions in the time series of the CR exports and imports several months ahead.

In the next article***, Subjective approach to assessing poverty in Poland – implications for social policy,*** **Leszek Morawski** and **Adrian Domitrz** discuss the effect of adopting a particular weigh system in constructing equivalised household income, such as based on the OECD recommendations concerning such scales. Poland is an interesting case for applying an alternative, subjective approach to calculating equivalent scales due to relatively large average size of households. The overall poverty rates for the two approaches are not distinctly different but they lead to significantly different distributions of poverty. For instance, the subjective approach suggests that one-person households and not large families should be considered most exposed to risk of material poverty. Since the relative positions of different policy-relevant groups of households in the distribution of income differ significantly, the respective programs of social transfers may need to be revised in order to be better targeted.

The section *other articles* includes a conference paper (presented at the Multivariate Statistical Analysis conference held in Łódź, 2016) by **Marek Walesiak** and **Andrzej Dudek,** ***Selecting the Optimal Multidimensional Scaling Procedure for Metric Data with R environment.*** The authors start with an observation that the main decision problem of multidimensional scaling (MDS) procedure for the metric measurement data consists in making selection of the method of normalization of the values of the variables and of distance measure, and finally of a MDS model. The article proposes a solution that allows choosing the optimal multidimensional scaling procedure out of 18 normalization methods included in the analysis and of 5 distance measures for 3 types of MDS models using two criteria: Kruskal's Stress-1 fit measure and Hirschman-Herfindahl HHI index. An empirical example provides illustration of the proposed procedure.

Finally, there are two articles in the last section, *research communicates and letters*. In ***Sample Allocation in Estimation of Proportion in a Finite Population Divided into Two Strata*** **Wojciech Zieliński** and **Dominik Sieradzki** discuss the problem of estimating a proportion of objects with a particular attribute in a finite population Classical estimator is compared with the estimator, which uses the information that the population is distributed among two strata. In the numerical example it was shown that variance of stratified estimator may be smaller by one-fourth compared to variance of classical estimator.

Remarks of the Estimation of Position Parameters by **Czesław Domański** concludes the issue. The author puts under reconsideration the classic problem of the level of accuracy of estimation of random variable parameter due to the lack of an unambiguous procedure to determine the scope of the distance between the value of an estimator and the real value of parameter. Some suggestions on how to deal with the situation when an obtained interval is too wide are provided.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 373, DOI 10. 21307

SUBMISSION INFORMATION FOR AUTHORS

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

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

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

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

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

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

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

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

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 375–392, DOI 10. 21307

POPULATION VARIANCE ESTIMATION USING FACTOR TYPE IMPUTATION METHOD

RANJITA PANDEY[1], **KALPANA YADAV**[2]

ABSTRACT

We propose a variance estimator based on factor type imputation in the presence of non-response. Properties of the proposed classes of estimators are studied and their optimality conditions are derived. The proposed classes of factor type ratio estimators are shown to be more efficient than some of the existing estimators, namely, the usual unbiased estimator of variance, ratio-type, dual to ratio type and ratio cum dual to ratio estimators. Their performances are assessed on the basis of relative efficiencies. Findings are illustrated based on a simulated and real data set.

Key words: auxiliary information, mean squared error, simple random sampling without replacement (SRSWOR).

1. Introduction

Estimation of population variance is of significant importance in the theory of estimation. Efficient variance estimation under auxiliary information has been widely discussed by various authors such as Das and Tripathi (1978), Srivenkatramana (1980), Isaki (1983), Singh et al. (1988), Singh and Katyar (1991), Rao and Shao (1992), Sarndal (1992), Agrawal and Sthapit (1995), Rao and Sitter (1995), Garcia and Cebrain (1996), Arcos et al. (2005), Kadilar and Cingi (2006, 2006a), Solanki and Singh (2013) and Yadav & Kadilar (2013).

A common aspect of data collection is the inability to record all items under a response variable. Amputing incomplete observations from the collected or available data and proceeding with statistical analysis of the restricted complete data set is the most common and convenient approach of handling missing data. However, the process of replacing missing items with plausible values called imputation is popular among data analysts as it enables construction of standard programs based on some probability sampling models, for substituting missing data with a point estimate. Such models have potential to reduce bias and improve

[1] Department of Statistics, University of Delhi, New Delhi. E-mail: ranjitapandey111@gmail.com.
[2] Department of Statistics, University of Delhi, New Delhi. E-mail: kalpana22yadav@gmail.com.

precision to a significant extent in comparison with the amputation approach. Rubin (1976), Fay (1991) and Rao (1996) have reviewed various imputation techniques.

Large sample surveys are mostly accompanied either by *unit* non-response, where a sampled subject refuses/is unable to provide information for some variables, or *item* non-response, where several units on the study variable are missing. Variance estimation after imputation has been studied by Kim et al. (2001), Raghunath and Singh (2006), Beaumont at al. (2011) and Singh and Solanki (2009-2010) using auxiliary information in the presence of random non-response. In the present paper, an improved factor type (FT) estimator of population variance based on an auxiliary variable is proposed, under non-response. Our work is motivated by the theoretical properties of FT estimator introduced by Singh and Shukla (1987).

2. Notations and estimators in literature

Let $\Omega = \{1,2,.....,N\}$ be a finite population of N identifiable units. Let $(y_i, x_i),\ i = 1,2,3,...,N$ be the observed value of study variable and auxiliary variable for i^{th} individual from a finite population Ω. From a finite population of *N* identifiable units, a simple random units sample, *s*, of size *n* is drawn without replacement. *r* denotes the number of responding units in the sample *s*. The remaining *(n-r)* units are non-responding units.

The following notations for the population are defined for study and auxiliary variables respectively: $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i, \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$:

Population mean of the variables *Y* and *X*; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ represent sample means of the study variable *y* and the auxiliary variable *x* respectively; $S^2_{y(N)} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2, S^2_{x(N)} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$: population variances of variables *Y* and *X*; $s^2_{y(n)} = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y})^2, s^2_{x(n)} = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$: sample variances of corresponding to variables *Y* and *X*; $s^2_{y(r)} = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y})^2, s^2_{x(r)} = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 =$ sample variances of responding units in the respective samples; $\mu_{ts} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^t (x_i - P)^s$, $\lambda_{ts} = \frac{\mu_{ts}}{\mu_{t0}^{t/2}\mu_{0s}^{s/2}}$, and the sampling fraction $f = \frac{n}{N}$. *B*(.) represents bias and *M* (.) represents mean squared error of the respective estimators.

To obtain the bias and M.S.E. of existing and suggested estimators we additionally consider

$s^2_{y(r)} = S^2_{y(N)}(1+e_0); s^2_{x(r)} = S^2_{x(N)}(1+e_1)$; $s^2_{x(n)} = S^2_{x(N)}(1+e_2)$ and $|e_i|<1; (i=0,1,2)$.

such that $E(e_0) = E(e_1) = E(e_2) = 0$; $E(e_0^2) = M_1(\lambda_{40}-1)$, $E(e_1^2) = M_1(\lambda_{04}-1)$,

$E(e_2^2) = M_2(\lambda_{04}-1)$;

$E(e_0e_1) = M_1(\lambda_{22}-1), E(e_1e_2) = M_2(\lambda_{04}-1), E(e_0e_2) = M_2(\lambda_{22}-1)$ where

$$M_1 = \frac{1}{r} - \frac{1}{N},\ M_2 = \frac{1}{n} - \frac{1}{N},\ M_3 = M_1 - M_2 = \frac{1}{r} - \frac{1}{n}.$$

The variance of the usual unbiased variance estimator $S^2_{y(N)}$ is given by:

$$V\left(S^2_{y(N)}\right) = M_1 S^4_{y(N)}(\lambda_{40}-1) \tag{1}$$

Isaki (1982) (hereafter *IK*) discussed a *ratio type* variance estimator for estimating population variance and its properties. Under non-response we write

$$t_{IK} = s^2_{y(r)} \frac{S^2_{x(N)}}{s^2_{x(r)}} \tag{2}$$

The estimator t_{IK} is found to be biased and its M.S.E. is given by:

$$M(t_{IK}) = M_1 S^4_{y(N)}\left[(\lambda_{04}-1)+(\lambda_{40}-1)-2(\lambda_{22}-1)\right] \tag{3}$$

Srivenkataramana and Tracy (1980) (hereafter *SV*) have given a *dual to ratio* estimator for variance estimator in sample surveys. Under non-response it can be modified as:

$$t_{SV} = s^2_{y(r)}\left[\frac{S^2_{x(N)} - f s^2_{x(r)}}{(1-f)S^2_{x(N)}}\right] \tag{4}$$

The M.S.E. of t_{SV} is given by:

$$M(t_{SV}) = M_1 S^4_y\left[(\lambda_{40}-1)+g^2(\lambda_{04}-1)-2g(\lambda_{22}-1)\right] \text{where } g = \frac{f}{(1-f)} \tag{5}$$

Yadav and Kadilar (2013) (hereafter *YK*) proposed the *ratio-cum-dual to ratio* type estimator for the population variance of the study variable. The ratio-cum-dual type variance estimator under non-response is given by:

$$t_{YK} = s^2_{y(r)}\left[\alpha\frac{S^2_{x(N)}}{s^2_{x(r)}} + (1-\alpha)\frac{S^2_{x(N)} - f s^2_{x(r)}}{(1-f)S^2_{x(N)}}\right] \tag{6}$$

The M.S.E. of t_{YK} is given by:

$$M(t_{YK}) = M_1 S^4_{y(N)} \left[(\lambda_{40} - 1) + \alpha_1^2 (\lambda_{04} - 1) - 2\alpha_1 (\lambda_{22} - 1) \right] \tag{7}$$

The M.S.E. of the proposed estimator is minimized for the optimum value α as

$$\alpha = \frac{K - g}{1 - g} \text{ such that } K = \frac{\lambda_{22} - 1}{\lambda_{04} - 1}$$

$$M(t_{YK})_{\min} = M_1 S^4_{y(N)} \left[\lambda_{40} - 1 - \frac{(\lambda_{22} - 1)^2}{\lambda_{04} - 1} \right] \tag{8}$$

3. Proposed estimators and their properties

Singh and Shukla (1987) proposed a family of FT ratio estimator of *population mean* for *complete sample* case. Unbiased, ratio, product and dual to ratio estimators are its special cases. An advantage of one-parameter class of estimators is that it requires only knowledge of the quantity $\rho \frac{C_y}{C_x}$ for making the best selection of the parameter. Population correlation coefficient between variables *Y* and *X* is represented by ρ and the respective coefficient of variation by C_y and C_x. The value of function $\rho \frac{C_y}{C_x}$ does not fluctuate considerably in repeated surveys and therefore could be guessed accurately from previous data or past experience or a pilot survey or otherwise [(Murthy (1967); Reddy, (1978)]. The *proposed variance estimator* is constructed as a function of some factors of the parameter termed as Factor-Type (F-T) estimator. This process of factorization makes it possible to yield more than one optimum value of the parameter so that at the same time bias of the estimator can also be controlled. The new class of FT ratio estimator for *population variance* of the study variable *under non-response* is proposed as:

$$t_{SSi} = s^2_{y(r)} \phi_i(k); i = 1,2,3 \tag{9}$$

where $\phi_1(k) = \frac{(A + C) S^2_{x(N)} + fBs^2_{x(n)}}{(A + fB) S^2_{x(N)} + Cs^2_{x(n)}}$; $\phi_2(k) = \frac{(A + C) s^2_{x(n)} + fBs^2_{x(r)}}{(A + fB) s^2_{x(n)} + Cs^2_{x(r)}}$ and

$$\phi_3(k) = \frac{(A + C) S^2_{x(N)} + fBs^2_{x(r)}}{(A + fB) S^2_{x(N)} + Cs^2_{x(r)}}.$$

where, $A = (k-1)(k-2)$, $B = (k-1)(k-4)$, $C = (k-2)(k-3)(k-4)$, $0 \le k \le \infty$.

Assume, $\theta_1 = \frac{fB}{A + fB + C}$, $\theta_2 = \frac{C}{A + fB + C}$ and $\theta = \theta_1 - \theta_2$

The properties of the proposed family of estimators are presented through the following theorems:

Theorem 1:

(i) The estimator t_{SS1} for population variance could be written in terms of $e_i ; i = 0,1,2$ as

$$t_{SS1} = S^2_{y(N)}\left[1 + \theta e_2 - \theta\theta_2 e_2^2 + \theta e_0 e_2 + e_0\right] \tag{10}$$

with $B(t_{SS1}) = S^2_{y(N)} M_2 \theta\left[(\lambda_{22}-1) - \theta_2(\lambda_{04}-1)\right]$ (11)

and M.S.E. $M(t_{SS1}) = S^4_{y(N)}\left[M_1(\lambda_{40}-1) + 2\theta M_2(\lambda_{22}-1) + \theta^2 M_2(\lambda_{04}-1)\right]$ (12)

The corresponding minimum M.S.E. at $\theta = \dfrac{-(\lambda_{22}-1)}{(\lambda_{04}-1)} = -P$ is given by

$$\left[M(t_{SS1})\right]_{\min} = S^4_{y(N)}\left[\frac{M_1(\lambda_{40}-1)(\lambda_{04}-1) - M_2(\lambda_{22}-1)^2}{(\lambda_{04}-1)}\right] \tag{13}$$

(ii) The estimator t_{SS2} in terms of $e_i ; i = 0,1,2$ is

$$t_{SS2} = S^2_{y(N)}\left[1 + e_0 - \theta e_2 + \theta e_1 - \theta\theta_2 e_2^2 + \theta\theta_1 e_2^2 - \theta e_0 e_2 + \theta e_0 e_1 - \theta^2 e_1 e_2\right] \tag{14}$$

with $B(t_{SS2}) = S^2_{y(N)}(M_1 - M_2)\theta\left[(\lambda_{22}-1) - \theta_2(\lambda_{04}-1)\right]$ (15)

and $M(t_{SS2}) = S^4_{y(N)}\left[M_1(\lambda_{40}-1) + \theta M_3\left[2(\lambda_{22}-1) + \theta(\lambda_{04}-1)\right]\right]$ (16)

The minimum mean squared error of t_{SS2} at $\theta = \dfrac{-(\lambda_{22}-1)}{(\lambda_{04}-1)} = -P$ is given by

$$\left[M(t_{SS2})\right]_{\min} = S^4_{y(N)}\left[\frac{M_1(\lambda_{40}-1)(\lambda_{04}-1) - M_3(\lambda_{22}-1)^2}{(\lambda_{04}-1)}\right] \tag{17}$$

(iii) The estimator t_{SS3} for population variance could be written in terms of

$e_i ; i = 0,1,2$ as $t_{SS3} = S^2_{y(N)}\left[1 + e_0 + \theta e_1 - \theta\theta_2 e_1^2 + \theta e_0 e_1\right]$ (18)

with $B(t_{SS3}) = S^2_{y(N)} M_1 \theta\left[(\lambda_{22}-1) - \theta_2(\lambda_{04}-1)\right]$ (19)

and $M(t_{SS3}) = S^4_{y(N)} M_1\left[(\lambda_{40}-1) + 2\theta(\lambda_{22}-1) + \theta^2(\lambda_{04}-1)\right]$ (20)

The minimum M.S.E. of t_{SS3} at $\theta = \dfrac{-(\lambda_{22}-1)}{(\lambda_{04}-1)} = -P$ is given by

$$\left[M\left(t_{SS3}\right)\right]_{\min} = S_{y(N)}^4 M_1\left[\frac{\left(\lambda_{40}-1\right)\left(\lambda_{04}-1\right)-\left(\lambda_{22}-1\right)^2}{\left(\lambda_{04}-1\right)}\right] \tag{21}$$

Proof: $t_{SS3} = s_{y(r)}^2 \phi_i(k); i = 1,2,3.$

Substituting the value of $\phi_i(k); i = 1,2,3$ and using the concept of large sample approximation, we get

$$t_{SS1} = S_{y(N)}^2\left(1+e_0\right)\left[\frac{A+C+f\,B+f\,Be_2}{A+C+f\,B+Ce_2}\right]$$

$$= S_{y(N)}^2\left(1+e_0\right)\left(1+\theta_1 e_2\right)\left(1+\theta_2 e_2\right)^{-1}$$

$$t_{SS2} = S_{y(N)}^2\left(1+e_0\right)\left[\frac{A+f\,B+C+Ce_2+f\,Be_1}{A+f\,B+C+Ce_1+f\,Be_2}\right] = S_{y(N)}^2\left(1+e_0\right)\left(1+\theta_1 e_1+\theta_2 e_2\right)\left(1+\theta_1 e_2+\theta_2 e_1\right)^{-1}$$

$$t_{SS3} = S_{y(N)}^2\left(1+e_0\right)\left[\frac{A+C+f\,B+f\,Be_1}{A+C+f\,B+Ce_1}\right] = S_{y(N)}^2\left(1+e_0\right)\left(1+\theta_1 e_1\right)\left(1+\theta_2 e_1\right)^{-1}$$

Using Taylor's expansion and ignoring terms of $o\left(n^{-1}\right)$ and higher order leads to equations (10), (14) and (18).

Since we know that $B\left(t_{SSi}\right) = E\left(t_{SSi} - S_{y(N)}^2\right); i = 1,2,3.$

Therefore, $B\left(t_{SS1}\right) = S_{y(N)}^2 E\left[e_0 + \theta e_2 - \theta\theta_2 e_2^2 + \theta e_0 e_2\right]$

$$B\left(t_{SS2}\right) = S_{y(N)}^2 E\left[e_0 - \theta e_2 + \theta e_1 - \theta\theta_2 e_1^2 + \theta\theta_1 e_2^2 - \theta e_0 e_2 + \theta e_0 e_1 - \theta^2 e_1 e_2\right]$$

and $B\left(t_{SS3}\right) = S_{y(N)}^2 E\left[e_0 + \theta e_1 - \theta\theta_2 e_1^2 + \theta e_0 e_1\right]$

Substituting the values of $e_i; i = 0,1,2$ using section 2, and simplifying, equations (11), (15) and (19) are obtained.

Also, $M\left(t_{SSi}\right) = E\left(t_{SSi} - S_{y(N)}^2\right)^2; i = 1,2,3.$

Substituting the values of estimators and solving it, and ignoring higher order terms, we get

$$M\left(t_{SS1}\right) = S_{y(N)}^4 E\left[e_0^2 + 2\theta e_0 e_2 + \theta^2 e_2^2\right]$$

$$M\left(t_{SS2}\right) = S_{y(N)}^4 E\left[e_0^2 - 2\theta e_0 e_2 + 2\theta e_0 e_1 + \theta^2 e_2^2 - 2\theta^2 e_1 e_2 + \theta^2 e_1^2\right]$$

$$M\left(t_{SS3}\right) = S_{y(N)}^4 E\left[e_0^2 + 2\theta e_0 e_1 + \theta^2 e_1^2\right]$$

Substituting the expectations values of e_0, e_1 and e_2 and solving it, leads to equations (12), (16) and (20).

Now, differentiating these expressions with respect to P and then equating to zero yields $\frac{d\,M(t_{SSi})}{dP} = 0 \Rightarrow P = \frac{(\lambda_{22}-1)}{(\lambda_{04}-1)}$

Substituting the value of P in equation (12), (16) and (20), corresponding expressions for the minimum *M.S.E.s* are obtained.

Remark 1: Multiple choices of *k*:

The optimality condition $\theta = \theta_1 - \theta_2 = -P$ provides the equation

$$(P-1)k^3 + [P(f-8)+(f+9)]k^2 + [P(23-5f)-(26+5f)]k + [(4f-22)P+(4f+24)] = 0, \tag{22}$$

which is a cubic equation in k. Its roots are represented by k_1, k_2, k_3 (say), for which mean squared error is optimum. The best choice criterion for k, which controls the quantum of bias in the corresponding estimator, is outlined in the following algorithm:

Step I: Compute $\left|B(t_{SSi})_{k_j}\right|$ *for* $i, j = 1,2,3$

Step II: For given *i*, choose k_j as $\left|B(t_{SSi})_{k_j}\right| = \min_{j=1,2,3}\left[\left|B(t_{SSi})_{k_j}\right|\right]$.

Remark 2: Factor-type ratio estimator (Singh and Shukla (1987)) for population variance of the study variable (without imputation) is defined as:

$$t_{SS} = \frac{(A+C)S_x^2 + fBs_x^2}{(A+fB)S_x^2 + Cs_x^2} \tag{23}$$

M.S.E. and minimum M.S.E. of estimator t_{SS} at $\theta = \frac{-(\lambda_{22}-1)}{(\lambda_{04}-1)}$ are shown below:

$$M(t_{SS}) = S_y^4 M_1\left[(\lambda_{40}-1)+\theta^2(\lambda_{04}-1)+2\theta(\lambda_{22}-1)\right] \tag{24}$$

$$\left[M(t_{SS})\right]_{\min} = S_y^4 M_1\left[\frac{(\lambda_{40}-1)(\lambda_{04}-1)-(\lambda_{22}-1)^2}{(\lambda_{04}-1)}\right] \tag{25}$$

4. Comparisons

On pair-wise comparison of expressions for *M.S.E.s* (from section 2 and section 3) (i) among the proposed estimators (ii) between the proposed and some of the existing estimators, we obtain theoretical conditions of superiority, which are shown in Table 1 and Table 2.

Table 1. Comparison within Proposed estimators

Estimators (Minimum M.S.E.)	More efficient than (Minimum M.S.E.)	Condition
t_{SS2}	t_{SS1}	$\lambda_{22}>1$
t_{SS3}		
t_{SS3}	t_{SS2}	

Table 2. Comparison within Proposed estimators and Traditional estimators

Estimators (Minimum M.S.E.)	More efficient than (Minimum M.S.E.)	Condition
t_{SS1}	$S^2_{y(N)}$	$\lambda_{22}>1$
t_{SS2}		
t_{SS3}		
t_{SS1}	t_{IK}	$(\lambda_{22}-1)^2>\frac{M_1}{M_2}A$
t_{SS2}		$(\lambda_{22}-1)^2>\frac{M_1}{M_3}A$
t_{SS3}		$(\lambda_{22}-1)^2>A$
t_{SS1}	t_{SV}	$(\lambda_{22}-1)^2>\frac{M_1}{M_2}B$
t_{SS2}		$(\lambda_{22}-1)^2>\frac{M_1}{M_3}B$
t_{SS3}		$(\lambda_{22}-1)^2>B$
t_{SS1}	t_{YK}	$\lambda_{22}>1$
t_{SS2}		$\lambda_{22}>1$
t_{SS3}		is equal to

where $A=(\lambda_{04}-1)[2(\lambda_{22}-1)-(\lambda_{04}-1)]$; $B=(\lambda_{04}-1)g[2(\lambda_{22}-1)-g(\lambda_{04}-1)]$.

5. Simulation study

An artificial population [Source: Shukla and Thakur (2008)] of size $N = 200$ containing values of main variable Y and auxiliary variable X.

Parameters of the population are given as below:

$\overline{Y} = 42.485; \quad \overline{X} = 18.515; \quad S_y^2 = 199.0598; \quad S_x^2 = 48.5375; \quad \rho = 0.8652;$

$f = 0.3; \quad \lambda_{22} = 2.47; \quad \lambda_{04} = 3.74; \quad \lambda_{40} = 2.56, \quad$ n = 60, r = 50

For the above data set, equation (22) provides three k-values: $k_1 = 1.54$; $k_2 = 2.94$; $k_3 = 6.67$

The simulation process comprises the following steps:

Step 1: Draw a random sample of size n = 60 from the population of $N = 200$ by SRSWOR.

Step 2: Discard 10 randomly chosen units from each sample corresponding to Y.

Step 3: Impute these discarded units of Y by the proposed methods and the available methods separately. Compute the value of different estimators and also for the proposed estimators.

Step 4: Repeat the above steps 30,000 times, which provides multiple sample-based estimates $\hat{t}_{s^2(1)}, \hat{t}_{s^2(2)} \ldots\ldots \hat{t}_{s^2(30,000)}$.

Step 5: Bias of $\hat{t}_1$ is obtained by

$$B\left(\hat{t}_{s^2}\right) = \frac{1}{30{,}000} \sum_{i=1}^{30,000} \left[s_{y(r)}^2 - S_{y(N)}^2 \right] .$$

Step 6: Mean squared error of $\hat{y}$ is computed by

$$M\left(\hat{t}_{s^2}\right) = \frac{1}{30{,}000} \sum_{i=1}^{30,000} \left[s_{y(r)}^2 - S_{y(N)}^2 \right]^2 .$$

Step 7: Percentage Relative efficiency (*PRE*) is computed from equation (26) and shown in Table 5:

$$PRE\left(*, t_{SSi}\right)_j = \frac{M[*]}{M\left[t_{SSi\,;\,i=1,2,3}\right]} \times 100 : j = 1,2,3,4; \tag{26}$$

such that * represents different existing methods.

Bias and M.S.E.*s* of the existing and proposed estimators computed from 30,000 repeated samples drawn by SRSWOR from population $N = 200$ are shown in Table 3.

Table 3. Bias, Mean Squared Error of Different Suggested and Traditional Estimators

Traditional Estimators	Bias	M.S.E.	Suggested Estimators		Bias	M.S.E.
$S^2_{y(N)}$	-35.82	2417.27	t_{SS1}	$k_1 = 1.55$	7.03	1572.64
t_{IK}	-40.51	1914.42		$k_2 = 2.94$	3.32	1800.42
t_{SV}	-45.04	2130.76		$k_3 = 6.67$	6.48	1602.47
t_{YK}	-43.86	1934.08	t_{SS2}	$k_1 = 1.55$	20.61	1067.79
				$k_2 = 2.94$	20.37	1047.37
				$k_3 = 6.67$	20.57	1064.34
			t_{SS3}	$k_1 = 1.55$	1.27	1262.77
				$k_2 = 2.94$	-3.79	1511.06
				$k_3 = 6.67$	0.52	1293.07

Computational results for efficiency loss due to imputation is measured as $(LI)_i = \frac{M(t_{SSi})}{M(t_{SS})}$ such that, $M(t_{SSi})$ and $M(t_{SS})$ are the M.S.E.s of the proposed estimators with and without imputation (from Remark 2). The losses are reported in Table 4.

Table 4. Loss due to Imputation

Optimum k	$k_1 = 1.55$	$k_2 = 2.94$	$k_3 = 6.67$
$(LI)_1$	0.74	0.72	0.75
$(LI)_2$	0.68	0.75	0.70
$(LI)_3$	0.75	0.72	0.75

Table 5. P.R.E. of suggested estimators with respect to different Traditional estimators

Estimators	**Optimum k values**	$PRE\left(S^2_{y(N)}, t_{SSi}\right)_1$	$PRE\left(t_{IK}, t_{SSi}\right)_2$	$PRE\left(t_{SV}, t_{SSi}\right)_3$	$PRE\left(t_{YK}, t_{SSi}\right)_4$
t_{SS1}	$k_1 = 1.55$	153.71	121.73	135.49	122.98
	$k_2 = 2.94$	134.26	106.33	118.35	107.42
	$k_3 = 6.67$	150.85	119.47	132.97	120.69
t_{SS2}	$k_1 = 1.55$	226.38	179.29	199.55	181.13
	$k_2 = 2.94$	**230.79**	182.78	203.44	184.66
	$k_3 = 6.67$	227.11	179.87	200.19	181.72
t_{SS3}	$k_1 = 1.55$	191.43	151.61	168.74	153.16
	$k_2 = 2.94$	159.97	126.69	141.01	127.99
	$k_3 = 6.67$	186.94	148.05	164.78	149.57

5.1. Values of k for Unbiased Estimator $t_{SSi}; i = 1,2,3.$

For unbiased estimator,

$$B\left(t_{SSi\ ;\ i=1,2,3}\right) = 0 \Rightarrow S^2_{y(N)}\,\theta\left[\left(\lambda_{22}-1\right) - \theta_2\left(\lambda_{04}-1\right)\right] = 0$$

$$\Rightarrow \left(\theta_1 - \theta_2\right)\left[\left(\lambda_{22}-1\right) - \theta_2\left(\lambda_{04}-1\right)\right] = 0 \qquad (27)$$

Case 1: $\theta_1 - \theta_2 = 0 \Rightarrow \dfrac{fB - C}{A + fB + C} = 0 \Rightarrow fB - C = 0$

$$\Rightarrow f\left(k-1\right)\left(k-4\right) - \left(k-2\right)\left(k-3\right)\left(k-4\right) = 0$$

$$\Rightarrow \left(k-4\right)\left[f\left(k-1\right) - \left(k-2\right)\left(k-3\right)\right] = 0 \qquad (28)$$

From (27) either $\left(k-4\right) = 0 \Rightarrow k = k_1^{'} = 4 \qquad (29)$

$$\text{or } k^2 - \left(f+5\right)k + \left(f+6\right) = 0$$

the remaining two roots of *k* are

$$k_2^{'} = \frac{\left(f+5\right) + \sqrt{\left(f+5\right)^2 - 4\left(f+6\right)}}{2} \qquad (30)$$

$$k_3^{'} = \frac{\left(f+5\right) - \sqrt{\left(f+5\right)^2 - 4\left(f+6\right)}}{2} \qquad (31)$$

On putting the value of f for the above data set, we get

$$k_2^{'} = 3.5 \tag{32}$$

$$k_3^{'} = 1.8 \tag{33}$$

Case 2: $\left[(\lambda_{22}-1)-\theta_2(\lambda_{04}-1)\right]=0 \Rightarrow \theta_2 = \dfrac{(\lambda_{22}-1)}{(\lambda_{04}-1)}$ (34)

Since we know that $\theta_2 = \dfrac{C}{A+fB+C} = 0$. Then, on equating it with $\theta_2 = \dfrac{(\lambda_{22}-1)}{(\lambda_{04}-1)}$, we get a cubic equation in the form of k as follows:

$$(\lambda_{22}-\lambda_{04})k^3 - [9\lambda_{04} + f(\lambda_{22}-1) - 8\lambda_{22} - 1]k^2$$
$$+ [23\lambda_{22} - 26\lambda_{04} - 5f(\lambda_{22}-1) + 3]k + [24\lambda_{04} - 22\lambda_{22} + 4f(\lambda_{22}-1) - 2] = 0 \tag{35}$$

On putting the values of λ_{22}, λ_{04} and *f* we get three different values of k.

$$k_4^{'} = 1.72,\ k_5^{'} = 2.60,\ k_6^{'} = 6.19 \tag{36}$$

6. Real data analysis

A real data of size *N* = 66 is taken from Indian Institute of Sugarcane Research, which comprises annual production data (in '000 tonnes) represented as the auxiliary variable *X* and the corresponding cultivation area (in '000 ha.) represented as the study variable *Y*, over the time period of *1950-51* to *2015-16*. Parameters of the above population are given as below:

$\overline{Y} = 22.30;\ \overline{X} = 193558.80;\ S_y^2 = 2278933.68;\ S_x^2 = 8658527591;\ \rho = 0.9904;$

$f = 0.3;\ \lambda_{22} = 1.23;\ \lambda_{04} = 1.77;\ \lambda_{40} = 1.35,$ n = 20, r = 10

For the above data set, equation (22) provides three k-values: $k_1 =$ 1.68, $k_2 = 3.09$ and $k_3 =$ 5.23. Initially we selected 10,000 independent random samples of size n =20 from the above population of size $N =$ 66 by SRSWOR.

The empirical bias and M.S.E.*s* of the existing and proposed estimators computed from these repeated samples are shown in Table 6.

Table 6. Bias, Mean Squared Error of Different Suggested and Traditional Estimators

Traditional Estimators	**Bias**	**M.S.E.**	**Suggested Estimators**		**Bias**	**M.S.E.**
$S^2_{y(N)}$	-1.03E+06	1.24E+12	t_{SS1}	$k_1 = 1.68$	-1.01E+06	1.03E+12
t_{IK}	-1.34E+06	1.12E+12		$k_2 = 3.09$	-1.02E+06	1.05E+12
t_{SV}	-1.04E+06	1.08E+12		$k_3 = 5.23$	-1.02E+06	1.05E+12
t_{YK}	-1.12E+06	1.27E+12	t_{SS2}	$k_1 = 1.68$	-1.03E+06	1.04E+12
				$k_2 = 3.09$	-1.03E+06	1.04E+12
				$k_3 = 5.23$	-1.02E+06	1.02E+12
			t_{SS3}	$k_1 = 1.68$	-1.01E+06	1.03E+12
				$k_2 = 3.09$	-1.02E+06	1.05E+12
				$k_3 = 5.23$	-1.01E+06	1.04E+12

Table 7. Loss due to Imputation

Optimum k	$k_1 = 1.68$	$k_2 = 3.09$	$k_3 = 5.23$
$(LI)_1$	0.70	0.72	0.75
$(LI)_2$	0.77	0.76	0.70
$(LI)_3$	0.73	0.77	0.75

Table 8. P.R.E. of suggested estimators with respect to different Traditional estimators

Estimators	**Optimum** k **values**	$PRE(S^2_{y(N)}, t_{SSi})_1$	$PRE(t_{IK}, t_{SSi})_2$	$PRE(t_{SV}, t_{SSi})_3$	$PRE(t_{YK}, t_{SSi})_4$
t_{SS1}	$k_1 = 1.55$	120.39	108.74	104.85	123.30
	$k_2 = 2.94$	118.10	106.67	102.86	120.95
	$k_3 = 6.67$	118.10	106.67	102.86	120.95
t_{SS2}	$k_1 = 1.55$	119.23	107.69	103.85	122.12
	$k_2 = 2.94$	119.23	107.69	103.85	122.12
	$k_3 = 6.67$	121.57	109.80	105.88	**124.51**
t_{SS3}	$k_1 = 1.55$	120.39	108.74	104.85	123.30
	$k_2 = 2.94$	118.10	106.67	102.86	120.95
	$k_3 = 6.67$	119.23	107.69	103.85	122.12

6.1 Values of k for Unbiased Estimator $t_{SSi}; i=1,2,3.$

For unbiased estimator, $B\left(t_{SSi\,;\,i=1,2,3}\right)=0 \Rightarrow S^2_{y(N)}\theta\left[\left(\lambda_{22}-1\right)-\theta_2\left(\lambda_{04}-1\right)\right]=0$

$$\Rightarrow\left(\theta_1-\theta_2\right)\left[\left(\lambda_{22}-1\right)-\theta_2\left(\lambda_{04}-1\right)\right]=0 \qquad \ldots(37)$$

Case 1: $\theta_1-\theta_2=0\Rightarrow\dfrac{fB-C}{A+fB+C}=0\Rightarrow fB-C=0$

$$\Rightarrow f\left(k-1\right)\left(k-4\right)-\left(k-2\right)\left(k-3\right)\left(k-4\right)=0$$

$$\Rightarrow\left(k-4\right)\left[f\left(k-1\right)-\left(k-2\right)\left(k-3\right)\right]=0 \qquad (38)$$

From (28) either $\left(k-4\right)=0\Rightarrow k=k_1^{'}=4$ (39)

$$\text{or} \quad k^2-\left(f+5\right)k+\left(f+6\right)=0$$

the remaining two roots of *k* are

$$k_2^{'}=\frac{\left(f+5\right)+\sqrt{\left(f+5\right)^2-4\left(f+6\right)}}{2} \qquad (40)$$

$$k_3^{'}=\frac{\left(f+5\right)-\sqrt{\left(f+5\right)^2-4\left(f+6\right)}}{2} \qquad (41)$$

On putting the value of f for the above data set, we get

$$k_2^{'}=3.5 \qquad (42)$$

$$k_3^{'}=1.8 \qquad (43)$$

Case 2: $\left[\left(\lambda_{22}-1\right)-\theta_2\left(\lambda_{04}-1\right)\right]=0\Rightarrow\theta_2=\dfrac{\left(\lambda_{22}-1\right)}{\left(\lambda_{04}-1\right)}$ (44)

Since we know that $\theta_2=\dfrac{C}{A+fB+C}=0$. Then, on equating it with $\theta_2=\dfrac{\left(\lambda_{22}-1\right)}{\left(\lambda_{04}-1\right)}$, we get a cubic equation in the form of k as follows:

$$\left(\lambda_{22}-\lambda_{04}\right)k^3-\left[9\lambda_{04}+f\left(\lambda_{22}-1\right)-8\lambda_{22}-1\right]k^2$$
$$+\left[23\lambda_{22}-26\lambda_{04}-5f\left(\lambda_{22}-1\right)+3\right]k+\left[24\lambda_{04}-22\lambda_{22}+4f\left(\lambda_{22}-1\right)-2\right]=0 \qquad (45)$$

On putting the values of λ_{22}, λ_{04} and *f* we get two different values of k.

$$k_4^{'}=0.72 \text{ and } k_5^{'}=7.53 \qquad (46)$$

7. Conclusion

The present paper suggests three new FT variance estimators under item non-response on the study variable, in a bivariate sample data. FT estimator, a generalized class of estimators for ratio, product, dual to ratio and the usual unbiased estimator are found to be more efficient than some existing estimators. The FT variance estimator maintains an optimum balance between reduction of bias and that of reducing M.S.E through *k*. We can choose k values for different pair of (f, P) values. Thus, the FT variance estimator could be made almost unbiased by an appropriate choice of multiple available k values.

Table 5 and Table 8 show P.R.E. of the suggested estimators with respect to different traditional estimators based on simulated and real data. It is observed from these tables that the proposed FT estimators prove to be better than the usual unbiased, ratio, dual to ratio and ratio cum dual to ratio estimators, under non-response. The proposed estimator t_{SS2} performs best among the three proposed estimators from the point of view of increasing efficiency. The three proposed FT type estimators are the best estimators in the sense of having the largest *PRE* among all the prevalent estimators discussed here.

Acknowledgments

Part of the work of the first author is supported by R&D Grant from University of Delhi.

REFERENCES

AGRAWAL, M. C., STHAPIT, A. B., (1995). Unbiased Ratio-Type Variance Estimation, Statistics and Probability Letters, 25, pp. 361–364.

ARCOS, A., RUEDA, M., MARTINEZ, M. D., GONZALEZ, S., ROMAN, Y., (2005). Incorporating the auxiliary information available in variance estimation, Applied Mathematics and Computation, 160, pp. 387–399.

BEAUMONT, J. F., HAZZIA, D., BOCCI, C., (2011). On variance estimation under auxiliary value imputation in sample surveys. Statistica Sinica, 21, pp. 515–537.

BEAUMONT, J. F., HAZIZA, D., BOCCI, C., (2011). On variance estimation under Auxiliary value imputation in sample surveys, Statistica Sinca, 21, pp. 515–537.

DAS, A. K., TRIPATHI, T. P., (1978). Use of auxiliary information in estimating the finite population variance, Sankhya, C, 40, pp. 139–148.

FAY, B. E., (1991). A design based pros, predictive on missing data variance. Proceedings of the annual Research Conference, U.S., Bureau of the census, pp. 429–440.

GARCIA, M. R., CEBRAIN, A. A., (1996). Repeated substitution method: The ratio estimator for the population variance, Metrika, 43, pp. 101–105.

ISAKI, C. T., (1983). Variance estimation using auxiliary information, Journal of the American Statistical Association, 78, pp. 117–123.

KADILAR, C., CINGI, H., (2006). Ratio estimators for the population variance in simple and stratified random sampling, Applied Mathematics and Computation, 173, pp. 1047–1059.

KADILAR, C., CINGI, H., (2006a). Improvement in Variance Estimation using Auxiliary Information, Hacettepe Journal of Mathematics and Statistics, 35 (1), pp. 111–115.

KIM, J. K., BRICK, J. M., FULLER, W. A., KALTON, G., (2001). On the bias of the multiple imputation variance estimator in survey sampling, Journal of the Royal Statistical Society series B, 68, pp. 509–521.

MURTHY, M. N., (1967). Sampling Theory and Methods. Statistical Publishing Society, Calcutta, India.

RAGHUNATH, A. AND SINGH, S. (2006). Estimation of variance from missing data, Metron-International Journal of Statistics, LXIV, 2, pp. 161–177.

RAO, SITTER, (1995). Variance estimation under two phase sampling with application to imputation for missing data. Biometrika, 82, pp. 453–460.

RAO, J.N.K. (1996). On variance estimation with imputed survey data. Journal of the American Statistical Association, 91 (434), pp. 499-506.

RAO, J. N. K., SHAO J., (1992). Jackknife variance estimation with survey data under hot deck imputation. Biometrika, 79 (4), 811–822.

REDDY, V. N., (1978). A Study on the use of Prior Knowledge on Certain Population Parameters in Estimation. Sankhya, C, 40, pp. 29–37.

RUBIN, D. B., (1976). Inference and missing data, Biometrika, 63(3), pp. 581–592.

SARANDAL, C. E., (1992). Methods for estimating the precision of survey estimates when imputation has been used, Survey Methodology, 18, pp. 241–252.

SHUKLA, D., THAKUR, N. S., (2008). Estimation of mean with imputation of missing data using factor-type estimator, Statistics in Transition, 9, 1, pp. 33–48.

SINGH, H. P., UPADHYAYA, L. N., NAMJOSHI, U. D., (1988). Estimation of finite population variance, Current Science, 57, pp. 1331–1334.

SINGH, H. P., KATYAR, N. P., (1991). Variance estimation through the mean square successive differences and sample variance using a priori information, Journal of the Indian Society of Agricultural Statistics, 43, 1, pp. 16–29.

SINGH, H. P., SOLANKI, R. S., (2009). Estimation of finite population variance using auxiliary information in presence of random non-response, Gujarat Statistical Review, 1, pp. 37–637.

SINGH, H. P., SOLANKI, R. S., (2010). Estimation of finite population variance using auxiliary information in presence of random non-response, Gujarat Statistical Review, 2, pp. 46–58.

SINGH, V. K., SHUKLA, D., (1987). One parameter family of factor-type ratio estimator, Metron, 45, 1-2, pp. 273–283.

SOLANKI, R. S., SINGH, H. P., (2013). An improved class of estimators for the population variance, Model Assisted Statistics and Applications 8, 3, pp. 229–238.

SRIVENKATRAMANA, T., TRACY, D. S., (1980). An alternative to ratio method in sample surveys, Annals of the Institute of Statistical Mathematics, 32, pp. 111–120.

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

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 393–412, DOI 10. 21307

BAYESIAN MODEL AVERAGING AND JOINTNESS MEASURES: THEORETICAL FRAMEWORK AND APPLICATION TO THE GRAVITY MODEL OF TRADE

Krzysztof Beck[1]

ABSTRACT

The following study presents the idea of Bayesian model averaging (BMA), as well as the benefits coming from combining the knowledge obtained on the basis of analysis of different models. The BMA structure is described together with its most important statistics, g prior parameter proposals, prior model size distributions, and also the jointness measures proposed by Ley and Steel (2007), as well as Doppelhofer and Weeks (2009). The application of BMA is illustrated with the gravity model of trade, where determinants of trade are chosen from the list of nine different variables. The employment of BMA enabled the identification of four robust determinants: geographical distance, real GDP product, population product and real GDP *per capita* distance. At the same time applications of jointness measures reveal some rather surprising relationships between the variables, as well as demonstrate the superiority of Ley and Steel's measure over the one introduced by Dopplehofer and Weeks.

Key words: Bayesian model averaging, jointness measures, multi-model inference, gravity model of trade.

1. Introduction

In economics, a situation often arises when a vast number of different theories attempt to explain the same phenomenon. Although these theories may complement each other, it is very common that they contradict one another or are even mutually exclusive. In such cases, basing empirical verification on one or a few specifications of an econometric model turns out to be insufficient. Moreover, researchers applying varying specifications will arrive at different, very often incoherent or even contradictory, conclusions. Testing hypotheses on the basis of various economic model specifications can result in a situation in which a variable that is statistically significant in one research specification, may prove to be not significant in another one.

[1] Lazarski University. E-mail: beckkrzysztof@gmail.com.

Brock and Durlauf (2001) draw attention to a problem they called theory open-endedness. It takes place in a situation where two or more competing models propose different explanations of the same phenomenon, and each of the variables proposed as an explanation can be expressed using a different measure. Moreover, some of the theories can complement each other, while other serve as substitutes or even contravene each other. In such a situation, inference based on a single model can lead to contradictory or false conclusions.

The above-mentioned problem is clearly present in the context of the research into the determinants of international trade. The vast body of trade theories offers a great variety of explanations for international trade flows, which can be seen in any international economics textbook. What is more, there is considerable dispute over potential effects of participation in free trade agreements as well as monetary unions on international trade. Even though the gravity model of trade has been the backbone of international trade empirics for over half the century, it is still rather unclear which variables should accompany the core of the model. The literature is full of competing specifications without much attention paid to robustness checks.

For these reasons, this paper pertains to the transition from statistical relevance to basing inference on the robustness of results against a change in the specifications of a model. However, in such a case it is necessary to apply inference and combination of knowledge coming from different model specifications. In such a situation, it is possible to apply BMA, i.e. Bayesian Model Averaging. Through the estimation of all the models within a given set of data, this procedure allows one to determine which variables are robust regressors regardless of the specification. It also allows one to unequivocally establish the direction and strength given regressors possess, and it makes it possible to choose the best models of all possible configurations. Furthermore, using the jointness measures that are available within the BMA framework enables the determination of the substitutional and complementary relationships between the studied variables.

Therefore, for the above-mentioned reasons, BMA and jointness measures are the subject of this study. Theory and structure of Bayesian model averaging is presented in the first section while in the second one jointness measures are discussed. The third section provides an example of BMA application in the analysis of the gravity model of trade and comprises four sub-sections. In the first one, the gravity model of trade is presented, whereas the second shows the variables employed in the verification of the model. The third sub-section presents the results of applying BMA, and the fourth one demonstrates the results of the analysis using jointness measures. The last section provides the summary and conclusions of the article.

2. BMA – Bayesian Model Averaging

For the space of all models, unconditional posterior distribution of coefficient β is given by:

$$P(\beta|y) = \sum_{j=1}^{2^K} P(\beta|M_j, y) * P(M_j|y) \tag{1}$$

where: y denotes data, j (j=1, 2,..,m) is the number of the model, K being the total number of potential regressors, $P(\beta|M_j, y)$is the conditional distribution of coefficient β for a given model M_j, and $P(M_j|y)$ is the posterior probability of the model. Using the Bayes' theorem, the posterior probability of the model (PMP – Posterior Model Probability) $P(M_j|y)$ can be rendered as (Błażejowski *et al.*, 2016):

$$PMP = p(M_j|y) = \frac{l(y|M_j) * p(M_j)}{p(y)}, \tag{2}$$

where PMP is proportional to the product of $l(y|M_j)$ – model specific marginal likelihood – and $P(M_j)$ – model specific prior probability – which can be written down as $P(M_j|y) \propto l(y|M_j) * P(M_j)$. Moreover, because: $P(y) = \sum_{j=1}^{2^K} l(y|M_j) * P(M_j)$, weights of individual models can be transformed into probabilities through the normalization in relation to the space of all 2^K models:

$$P(M_j|y) = \frac{l(y|M_j) * P(M_j)}{\sum_{j=1}^{2^K} l(y|M_j) * P(M_j)}. \tag{3}$$

Applying BMA requires specifying the prior structure of the model. The value of the coefficients β is characterized by normal distribution with zero mean and variance $\sigma^2 V_{oj}$, hence:

$$P(\beta|\sigma^2, M_j) \sim N(0, \sigma^2 V_{oj}). \tag{4}$$

It is assumed that the prior variance matrix V_{oj} is proportional to the covariance in the sample: $(gX_j'X_j)^{-1}$, where g is the proportionality coefficient. The g prior parameter was put forward by Zellner (1986) and is widely used in BMA applications. In their seminal work on the subject of choosing the g prior Fernández *et al.* (2001) put forward the following rule, to choose the best g prior:

$$g = \frac{1}{\max(n, k^2)}, \tag{5}$$

where $\frac{1}{n}$ is known as UIP – unit information prior (Kass and Wasserman, 1995), whereas $\frac{1}{k^2}$ is convergent to RIC – risk inflation criterion (Foster and George, 1994). For further discussion on the subject of g priors see: Ley and Steel (2009, 2012); Feldkircher and Zeugner (2009); and Eicher *et al.* (2011).

Besides the specification of g prior, it is necessary to determine the prior model distribution while applying BMA. For binomial model prior (Sala-I-Martin *et al.*, 2004):

$$P(M_j) \propto \left(\frac{Em}{K}\right)^{k_j} * \left(1 - \frac{Em}{K}\right)^{K-k_j}, \tag{6}$$

where Em denotes the expected model size, while k_j the number of covariate in a given model. When $Em = \frac{K}{2}$ it turns into uniform model prior – priors on all the models are all equal ($P(M_j) \propto 1$). Yet another instance of prior model probability is binomial-beta distribution (Ley, Steel, 2009):

$$P(M_j) \propto \Gamma(1 + k_j) * \Gamma\left(\frac{K - Em}{Em} + K - k_j\right). \tag{7}$$

In the case of binomial-beta distribution with expected model size *K/2*, the probability of a model of each size is the same $\left(\frac{1}{K+1}\right)$. Thus, the prior probability of including the variable in the model amounts to 0.5, for both binomial and binomial-beta prior with $Em = K/2$.

Using the posterior probabilities of the models in the role of weights allows one to calculate the unconditional posterior mean and standard deviation of the coefficient β_i. Posterior mean (PM) of the coefficient β_i, independent of the space of the models, is then given with the following formula (Próchniak, Witkowski, 2012):

$$PM = E(\beta_i|y) = \sum_{j=1}^{2^K} P(M_j|y) * \hat{\beta}_{ij}, \tag{8}$$

where $\hat{\beta}_{ij} = E(\beta_i|y, M_j)$ is the value of the coefficient β_i estimated with OLS for the model M_j. The posterior standard deviation (PSD) is equal to (Próchniak, Witkowski, 2014):

$$PSD = \sqrt{\sum_{j=1}^{2^K} P(M_j|y) * V(\beta_j|y, M_j) + \sum_{j=1}^{2^K} P(M_j|y) * \left[\hat{\beta}_{ij} - E(\beta_i|y, M_j)\right]^2}, \tag{9}$$

where $V(\beta_j|y, M_j)$ denotes the conditional variance of the parameter for the model M_j.

The most important statistic for BMA is posterior inclusion probability (PIP). PIP for the regressor x_i equals:

$$PIP = P(x_i|y) = \sum_{j=1}^{2^K} 1(\varphi_i = 1|y, M_j) * P(M_j|y) \tag{10}$$

where $\varphi_i = 1$ indicates that the variable x_i is included in the model.

PM and PSD are calculated for all models, even those whose value $\varphi_i = 0$, which means that the variable is not present. Due to that fact the researcher can be interested in the value of the coefficient in the models in which a given variable is present. For that purpose, the value of the conditional posterior mean (PMC), that is the posterior mean, can be calculated on condition that a variable is included in the model:

$$PMC = E(\beta_i|\varphi_i = 1, y) = \frac{E(\beta_i|y)}{P(x_i|y)} = \frac{\sum_{j=1}^{2^K} P(M_j|y) * \hat{\beta}_{ij}}{P(x_i|y)}, \tag{11}$$

whereas the conditional posterior standard deviation (PSDC) is given by:

$$PSDC = \sqrt{\frac{V(\beta_j|y) + [E(\beta_i|y)]^2}{P(x_i|y)} - [E(\beta_i|\varphi_i = 1|y)]^2}. \tag{12}$$

Additionally, the researcher can be interested in the sign of the estimated parameter if it is included in the model. The posterior probability of a positive sign of the coefficient in the model [P(+)] is calculated in the following way:

$$P(+) = P[sign(x_i)|y] = \begin{cases} \sum_{j=1}^{2^K} P(M_j|y) * CDF(t_{ij}|M_j), & if\ \ sign[E(\beta_i|y)] = 1 \\ 1 - \sum_{j=1}^{2^K} P(M_j|y) * CDF(t_{ij}|M_j), & if\ sign[E(\beta_i|y)] = -1 \end{cases} \tag{13}$$

where CDF denotes cumulative distribution function, while $t_{ij} \equiv (\hat{\beta}_i/\widehat{SD}_i|M_j)$.

3. Jointness measures

All the statistics cited so far served to describe the influence of regressors on the dependent variable. However, the researcher should also be interested in relationships that emerge between the independent variables. To achieve that, one can utilize the measure of dependence between regressors, which is referred to as *jointness*.

Two teams of scientists came up with jointness measures at the same time. The article by Ley and Steel (2007) was published first; however, in this paper the concept of Doppelhofer and Weeks (2009) shall be presented first due to the fact that Ley and Steel's article constitutes by and large the critique of Dopplehofer and Weeks' concepts. Measures allow the determination of the substitution and complementary relationships between explanatory variables. Below, the focus will be put only on the jointness relationships between pairs of variables. It must also be mentioned, however, that testing the relationships between triplets or even more numerous sets of variables is possible.

We shall define posterior probabilities for the model M_j as:

$$\big(M_j|y\big) = P\big(\varphi_1 = w_1, \varphi_2 = w_2, \dots, \varphi_K = w_K|y, M_j\big) \tag{14}$$

where w_i can assume value 1 (if a variable is present in the model) and 0 if a variable is not present in the model. In the case of analysing two variables x_i and x_h the combined posterior probability of including two variables in the model can be expressed as follows:

$$P(i \cap h|y) = \sum_{j=1}^{2^K} 1\big(\varphi_i = 1 \cap\ \varphi_2 = 1|y, M_j\big) * P\big(M_j|y\big). \tag{15}$$

Table 1. Points of probability mass defined on space $\{0,1\}^2$ for uniform distribution $P(\varphi_i, \varphi_l|y)$.

$P(\varphi_i, \varphi_l\|y)$	$\varphi_h = 0$	$\varphi_h = 1$	Sum
$\varphi_i = 0$	$P\big(\bar{\imath} \cap \bar{h}\|y\big)$	$P(\bar{\imath} \cap h\|y)$	$P(\bar{\imath}\|y)$
$\varphi_i = 1$	$P\big(i \cap \bar{h}\|y\big)$	$P(i \cap h\|y)$	$P(i\|y)$
Sum	$P\big(\bar{h}\|y\big)$	$P(h\|y)$	1

Source: Doppelhofer, Weeks, 2009.

It can be thus stated that $P(i \cap h|y)$ is the sum of the posterior probability of the models, where variables marked by x_i and x_h appear. Doppelhofer and Weeks observe that the relationships between variablesx_i and x_h can be analyzed by comparing posterior probabilities of including these variables separately [$P(i|y)$ and $P(h|y)$] with probability of including and excluding both variables at the same time. The authors justify their reasoning by presenting an analysis of the case of a random vector (φ_i, φ_h) of the combined posterior distribution $P(\varphi_i, \varphi_l|y)$. The points of probability mass defined on space $\{0,1\}^2$ are shown in Table 1.

Table 1 shows distributions related to all the possible realizations of vector (φ_i, φ_h). It is easy to read from the table that the marginal probability of including variable x_i in the model can be calculated as:

$$P(i|y) = P(i \cap h|y) + P\left(i \cap \bar{h}|y\right), \tag{16}$$

whereas the probability of excluding the variable x_i can be rendered as:

$$P(\bar{\imath}|y) \equiv 1 - P(i|y) = P\left(\bar{\imath} \cap \bar{h}|y\right) + P(\bar{\imath} \cap h|y). \tag{17}$$

If there is a correlation between variables x_i and x_h, one should expect that expressions $P(i \cap h|y)$ and $P\left(\bar{\imath} \cap \bar{h}|y\right)$ will get higher values than expressions $P\left(i \cap \bar{h}\backslash y\right)$ and $P(\bar{\imath} \cap h|y)$. On that basis, to follow Whittaker (2009), the authors observe that the natural measure of correlation between two binary random variables φ_i and φ_h is the cross-product ratio (CPR), expressed as:

$$CPR(i, h|y) = \frac{P(i \cap h|y)}{P\left(i \cap \bar{h}|y\right)} * \frac{P\left(\bar{\imath} \cap \bar{h}|y\right)}{P(\bar{\imath} \cap h|y)}. \tag{18}$$

As the realizations of the vector (φ_i, φ_h) for each of the variables can only amount to 1 or 0, $P(i \cap h|y)$ is the binomial distribution of the uniform posterior probability i, which can be rendered as follows:

$$\begin{aligned} P(\varphi_i, \varphi_h|y) = {} & P(i \cap h|y)^{\varphi_i \varphi_h} * P\left(i \cap \bar{h}|y\right)^{\varphi_i(1-\varphi_h)} * \\ & * P(\bar{\imath} \cap h|y)^{(1-\varphi_i)\varphi_h} * P\left(\bar{\imath} \cap \bar{h}|y\right)^{(1-\varphi_i)(1-\varphi_h)} \end{aligned} \tag{19}$$

Logarithmized and put in order, the expressions take the following form:

$$\begin{aligned} ln[P(\varphi_i, \varphi_h|y)] = {} & ln\left[P\left(\bar{\imath} \cap \bar{h}|y\right)\right] + \varphi_h ln\left[\frac{P(\bar{\imath} \cap h|y)}{P\left(\bar{\imath} \cap \bar{h}|y\right)}\right] + \\ & + \varphi_i ln\left[\frac{P\left(i \cap \bar{h}|y\right)}{P\left(\bar{\imath} \cap \bar{h}|y\right)}\right] + \varphi_i \varphi_h ln\left[\frac{P(i \cap h)}{P\left(i \cap \bar{h}\backslash|\right)} * \frac{P\left(\bar{\imath} \cap \bar{h}|y\right)}{P(\bar{\imath} \cap h\backslash|)}\right] \end{aligned} \tag{20}$$

The independence between variables x_i and x_h is possible if and only if $ln[P(\varphi_i, \varphi_h|y)]$ is additive for $P(\varphi_i|y)$ and $P(\varphi_h|y)$. Independence can therefore occur if and only if the natural logarithm of CPR is 0, which means CPR equals 1.

On that basis, Doppelhofer and Weeks derive their jointness measure, which they define as:

$$J_{Dw(ih)} = ln[CPR(i,h|y)] = ln\left[\frac{P(i \cap h|y)}{P(i \cap \bar{h}|y)} * \frac{P(\bar{\imath} \cap \bar{h}|y)}{P(\bar{\imath} \cap h|y)}\right] =$$

$$= ln\left[\frac{P(i|h,y)}{P(\bar{\imath}|h,y)} * \frac{P(\bar{\imath}|\bar{h},y)}{P(i|\bar{h},y)}\right] = \ln\left[PO_{i|h} * PO_{\bar{\imath}|\bar{h}}\right]. \quad (21)$$

The expression $ln\left[PO_{i|h} * PO_{\bar{\imath}|\bar{h}}\right]$ is the natural logarithm of the product of two quotients of posterior odds, where $PO_{i|h}$ indicates posterior odds of including the variable x_i to the model on condition that $x_h$is included, while $PO_{\bar{\imath}|\bar{h}}$ indicates posterior odds of excluding the variable x_i from the model on condition that the variable x_h is excluded.

At this moment, it is worth pointing out that if the probability product of including and excluding both variables $[(P(i \cap h|y) * P(\bar{\imath} \cap \bar{h}|y)]$ is greater than the probability product of including each of the variables one at a time $[P(i \cap \bar{h}|y) * P(\bar{\imath} \cap h|y)]$, then the logarithm assumes positive values. Thus, for the positive values of the measure, complementary relationship has to occur: models that include both variables at the same time or reject both variables at the same time are characterized by the highest posterior probability. If the product of probabilities of including the variables separately is greater than the product of including both or neither at the same time, the logarithm takes negative values. In such an event, a substitutional relationship occurs. To sum up, Doppelhofer and Weeks' jointness measure assumes positive values if there is a complementary relationship between variables, whereas it assumes negative values when this relationship is of substitutional character.

Ley and Steel (2007) set out to develop a jointness measure that would possess the following characteristics:

1) Interpretability – a measure should have a formal statistical or intuitive interpretation.
2) Calibration – values of a measure should be determined on a clearly defined scale based on formal statistical or intuitive interpretation.
3) Extreme jointness – in a situation when two variables appear in all the analyzed models together (e.g. in the case of using MC^3 methods), the maximum value of jointness measure should occur;
4) Definability – jointness should be defined always if at least one of the considered variables is characterized by positive inclusion probability.

Ley and Steel claimed that Doppelhofer and Weeks' jointness measure is faulty as it is not defined in a situation when both regressors are included in all models and when one of the regressors is not taken into consideration in any of the models. Moreover, when the probability of including a variable in the model approaches 1, then the value of the measure is by and large dependent on the limit of the expression $\left[P\left(\bar{\imath} \cap \bar{h}|y\right)\right]/[\,P(\bar{\imath} \cap h|y)]$. This means that a few models, excluding the variable x_i, that are characterized by a very low probability can strongly influence the value of the measure: both in the direction of 0 (if they include the variable x_h) or ∞ (if they do not include the variable x_h). Thus, the measure $J_{Dw(ih)}$ does not contain features 1) and 4).

What is more, the authors pointed out that the interpretation of Doppelhofer and Weeks' measure is not clear enough and, due to this fact, they proposed an alternative measure. This measure is the ratio of probability of including two variables simultaneously to the sum of probabilities of including each of the variables separately, with the exclusion of the probability of including two variables at the same time. This measure meets all the criteria laid out by the authors. Ley and Steel's jointness measure is given by:

$$J_{LS(ih)} = ln\left[\frac{P(i \cap h|\mathrm{y})}{P\left(i \cap \bar{h}|y\right) + P(\bar{\imath} \cap h|y)}\right] = ln\left[\frac{P(i \cap h|y)}{P(i|y) + P(h|y) - 2P(i \cap h|y)}\right]. \tag{22}$$

The advantage of this measure is its interpretative clarity. The expression inside the natural logarithm represents the quotient of posterior odds of models including both variables to the models including each of them separately. Again, the logarithm of this expression takes positive values if the probability of the models including both variables is dominant, which testifies to the complementary relationship. The measure takes negative values if posterior odds of the models including variables separately are higher than in the case where variables appear in the model simultaneously, which testifies to a substitutional relationship.

Doppelhofer and Weeks calculated the limit values of jointness measures, which allow qualifying variables to one of five categories. These values also hold in the case of Lay and Steel's jointness measure. The limit values of jointness measures with their corresponding classifications of relationships between variables are presented in Table 2.

Table 2. Limit values of jointness measures and classification of relationships between variables

Type of the relationship between the variables	Value of the jointness measure (*J*)
Strong substitutes	*J < (-2)*
Significant substitutes	*(-2) < J < (-1)*
Unrelated variables	*(-1) < J < 1*
Significant complements	*1 < J < 2*
Strong complements	*2< J*

Source: Błażejowski, Kwiatkowski, 2015.

4. Application on the example of the gravity model of trade

All the empirical analyses employing BMA were carried out using BMS package for R environment (Zeugner and Feldkircher, 2015). Jointness measures were computed using a package for gretl (Błażejowski and Kwiatkowski, 2015).

4.1. Gravity model of trade

In the simplest form, the equation describing the gravity model of trade (Anderson, 1979, 2011; Egger, 2002; Anderson, Wincoop, 2003) can be shown as:

$$TRADE = \alpha \frac{(RGDPprod)^{\beta_1}}{DIST^{\beta_2}}, \tag{23}$$

which can be easily transformed into a log-linear form:

$$\ln(TRADE) = \ln(\alpha) + \beta_1 \ln(RGDPprod) - \beta_2 \ln(DIST), \tag{24}$$

where *TRADE* stands for the amount of international trade, *RGDPprod* – product of real GDP of the two countries, *DIST* – distance between the countries, whereas α, β_1, β_2 are parameters in the model. However, the model can be expanded by including additional explanatory variables, which was performed in this paper.

4.2. Variables and source of data

Data for 19 European Union countries was used, namely: Austria, Belgium, Cyprus, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, Malta, the Netherlands, Poland, Portugal, Spain, Sweden and the UK. All the variables are expressed bilaterally and as a result the size of the

sample for each variable amounts to 171 pairs of countries. The period of analysis spans the years between 1999 and 2007 for all the variables.

Bilateral trade, which is expressed as logarithmized trade between partners, constitutes the response variable in the model:

$$TRADE_{ij} = ln\left(\frac{1}{T}\sum_{t=1}^{T} Import_{ijt} + Export_{ijt}\right), \tag{25}$$

where *i* and *j* are indexes of partner countries, and the measure itself is a mean for the entire analyzed period (1, 2, …, T). The data on bilateral trade are taken from IMF Directions of Trade.

In the BMA analysis, 9 variables were employed. The first one constitutes the logarithm of the product of real GDPs:

$$RGDPprod_{ij} = ln\left(\frac{1}{T}\sum_{t=1}^{T} GDP_{it} * GDP_{jt}\right), \tag{26}$$

also treated as a mean for the whole period. Data on the subject of real GDP are taken from the Penn World Table. The second of the main gravity variables is the natural logarithm of the distance between the capitals of the countries under consideration, which is marked as *DIST*.

The basic explanatory variables in the gravity model of trade were complemented by additional 7. The first one is the similarity of the production structures measured by Krugman specialization index (1991):

$$KSI_{ij} = \frac{1}{T}\sum_{t=1}^{T}\sum_{l}^{17}\left|v_{it}^{l} - v_{jt}^{l}\right|, \tag{27}$$

where v_{it}^{l} is the value added in the sector *l* expressed as the percentage of the value added in the entire economy of the country *i* in the period *t*, v_{it}^{l} and is the value added in the sector *l* expressed as the percentage of the value added in the entire economy of a country *j* in the period *t*. The mean for the entire period and the division of the economy into 17 sectors were used, whereas the data on them were taken from EU KLEMS. The measure takes values from the interval [0,2], while the growth of the value of the measure is accompanied by the decrease in similarity of production structures.

The next variable added to the gravity model is the average absolute value of the difference of natural log of GDP *per capita* for each pair of countries in the period between 1999 and 2007:

$$RGDPdist_{ij} = \frac{1}{T}\sum_{t=1}^{T}\left|\ln(GDPpercapita_{it}) - \ln\left(GDPpercapita_{jt}\right)\right|. \tag{28}$$

The data on GDP *per capita* comes from the Penn World Table. The similarity of production structures and the distance of GDP *per capita* can be justified by the theory of monopolistic competition adopted by Linder (1961). The theory assumes that there is a tendency that, together with the increasing industrialization, the structures of consumption/production become more similar, which leads to a situation where countries at similar level of affluence will display a high level of intra-industry trade. These conclusions are supported by the works of: Grubel (1971), Grubel and Loyd (1975), Dixit and Stiglitz (1977), Krugman (1979, 1980), Lancaster (1980), Helpman (1981) and Gray (1980).

What is more, averaged binary variables were used in the models in order to reflect the influence of participation in the European Union and Economic and Monetary Union. For the participation in the monetary union (*MU*), the variable takes the value equal to 1 if in a given year both countries were members of the Eurozone, and 0 for other years. Then, a mean for the whole period is calculated. Analogical construction was applied for the participation in the European Union (*EU*):

Another potential determinant of bilateral trade is the natural logarithm of the population product of two analyzed EU countries in the period between 1990 and 2007 – *POPprod.* The data on the size of population come from the Penn World Table. One can expect substitutional relationship between *POPprod* and *RGDPprod*.

Moreover, two additional binary variables were used. They are: *BORDER* - a dummy variable assuming 1 if two countries share a common border, and *LANG* – a binary variable assuming 1 if a pair of countries share at least one official language.

4.3. The results of applying BMA

Below one can find the results of applying BMA after employing Fernández *et al.* (2001) *Benchmark Prior*, which dictated the choice of unit information prior (UIP). Additionally, uniform model size prior was applied. This combination of priors was recommended by Eicher *et al.* (2011). The prior probability of including a given regressor is 0.5. As 9 regressors were used, the space of the model consists of $2^K=2^9=512$ elements, and the inference itself was carried out on the basis of all models. The results of applying BMA are presented in Table 3.

The results indicate that 5 variables were qualified as robust determinants of bilateral trade: geographical distance, product of real GDPs, population product, GDP *per capita* distance, and common language. The remaining four display lower posterior than the prior probability of inclusion, which is 0.5. A stable sign of the coefficient among all the analyzed models also characterizes all the variables that were qualified as robust, and it is in accordance with expectations of the theory, with an exception of population product, which is characterized by negative posterior mean. *DIST* and *RGDPprod* turned out to be the most robust

determinants of trade – models including these variables take the lion's share of posterior probability mass. This ascertains the gravity model of trade capacity to explain international trade flows. *RGDPpc* has a negative impact on trade. This gives support to the theories that suggest a positive relationship between GDP *per capita* and the volume of intra-industry trade. On the other hand, similarity of the production structure is marked as fragile. It will be instructive to look at the value of the jointness measures for *RGDPdist* and *KSI*.

Table 3. BMA statistics with the use of uniform prior model size distribution (dependent variable - bilateral trade).

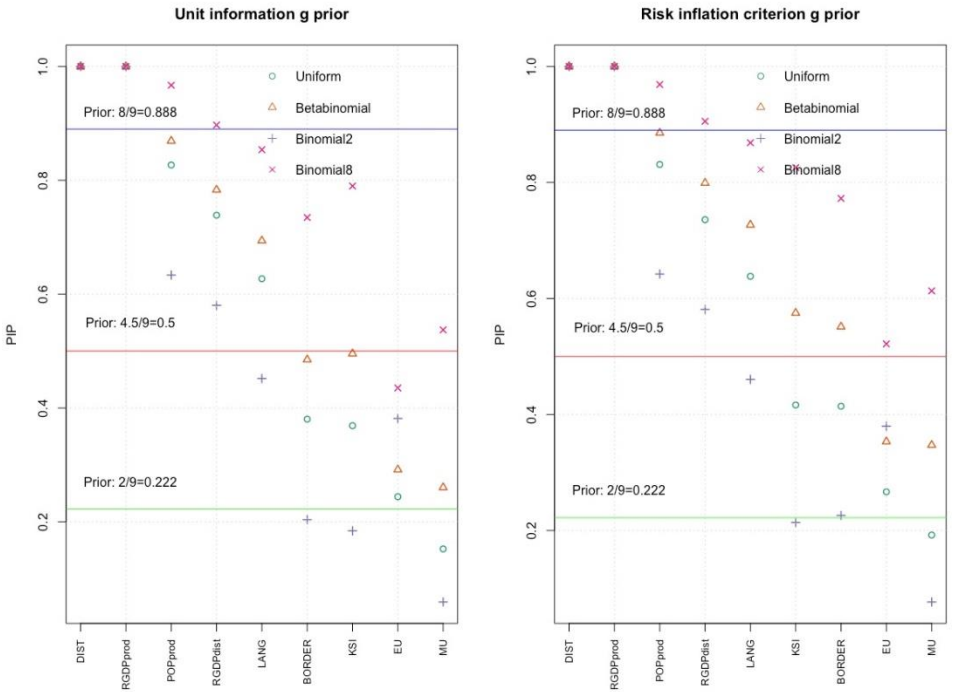
Variable	PIP	PM	PSD	CPM	CPSD	P(+)
DIST	1.000	-0.879	0.097	-0.879	0.097	0.000
RGDPprod	1.000	1.169	0.180	1.169	0.180	1.000
POPprod	0.827	-0.311	0.176	-0.376	0.114	0.000
RGDPdist	0.739	-0.336	0.242	-0.455	0.159	0.000
LANG	0.627	0.299	0.275	0.476	0.190	1.000
BORDER	0.380	0.139	0.209	0.365	0.180	1.000
KSI	0.369	-0.465	0.723	-1.260	0.645	0.000
EU	0.244	0.162	0.364	0.662	0.461	0.916
MU	0.152	0.022	0.069	0.146	0.116	1.000

A cultural similarity captured by the common language dummy proved to have a robust and positive impact on trade. Unexpected result was obtained for the population product. The variable is robust but is characterized by a negative posterior mean. This result is especially surprising when we look at correlation coefficient between *RGDPprod* and *POPprod* – 0.96. This suggests a substitutional dependence between those two variables.

The common border dummy was classified as fragile. This might be explained by potential substitutional relationship with geographical distance or language dummy – these variables most certainly carry the same information. Similarly, the membership in the European Union and the Eurozone are considered fragile. In instances of both of these variables one might expect a substitutional relationship with other regressors, e.g. *RGDPdist* (European Union/Eurozone members are characterized by lower GDP *per capita* distances compared with pairs with countries outside these entities) or *BORDER*.

The next step requires an inquiry on whether the conclusions rely upon the undertaken assumptions. Impact of changing g prior, as well as, model size prior is depicted in the Figure 1. No matter what prior model specification is chosen *DIST*, *RGDPprod*, *POPPprod* and *RGDPdist* are robust determinants of international trade. *LANG* depends on the chosen prior combination, which deem questioning robustness of this variable.

This point shows the superiority of BMA over the classical methods. Applying BMA allows one not only to use knowledge coming from many models but also to check the robustness of the results over the changes in prior specification: both in terms of g prior and model size prior. The classical approach based on statistical significance relies upon the knowledge coming from just one model. Model averaging procedures used in classical econometrics rely on a given specific set of prior assumptions, yet one more time making entire analysis more limited and vulnerable to criticism.



* Uniform, Betabinomial, Binomial2, Binomial8 – denotes uniform, binomial-beta with $Em = 4.5$, binomial with $Em = 2$ and binomial with $Em = 8$ model prior respectively.

Figure 1. Posteriori inclusion probabilities in different specifications of g prior and model size prior

4.4. Jointness measures

To uncover the character of the correspondence between regressors, jointness measures were employed. They were calculated for BMA with unit information prior and uniform model size prior. Results for both measures are shown in Table 4. The values of Doppelhofer and Weeks measures (J_{DW}) are located above the primary diagonal and for Ley and Steel's measure (J_{LS}) above.

Table 4. Jointness measures: J_{DW} (below primary diagonal) and J_{LS} (above primary diagonal)

x	MU	EU	RGDPdist	RGDPprod	POPprod	BORDER	DIST	LANG	KSI
MU	x	-2.48	-1.76	-1.73	-1.72	-2.30	-1.73	-1.97	-1.88
EU	-0.38	x	-1.96	-1.13	-2.48	-1.99	-1.13	-1.09	-1.84
RGDPdist	0.11	-1.85	x	1.03	1.14	-0.72	1.03	-0.02	-1.31
RGDPprod	nan	nan	nan	x	1.56	-0.48	0.00	0.53	-0.53
POPprod	0.24	-5.68	2.03	nan	x	-0.41	1.56	0.07	-0.49
BORDER	-0.45	-0.58	-0.13	nan	0.88	x	-0.48	-1.49	-0.98
DIST	nan	nan	nan	nan	nan	nan	x	0.53	-0.53
LANG	-0.28	0.50	-0.22	nan	-0.74	-1.57	nan	x	-0.86
KSI	0.14	-0.38	-1.72	nan	0.73	0.37	nan	-0.09	x

In Table 4, strong substitutes are highlighted in dark grey, whereas light grey indicates relevant substitutes. Employing the measure J_{DW} allowed the establishing of four pairs of substitutes, one pair of strong substitutes and one pair of complements. *EU* is a strong substitute of *POPprod* and a significant one of *RGDPdist* Border and language dummies are also substitutes, which might be reasonably explained in the following way: countries that are located closer to each other tend to share the same language more often. *KSI* exhibits substitutional relationship with *RGDPpc*. This result might be explained by U-shaped relationship between GDP *per capita* and degree of specialization described by Imbs and Wacziarg (2003): differences in GDP *per capita* are determining specialization patterns, and those in turn determine the patterns of trade. Moreover, using J_{DW} allowed for the identification of one pair of complements marked with the grey font: *POPprod* and *RGDPdist*.

Results in Table 3 reveal a few weaknesses related to the application of J_{DW}, which were mentioned in section 3. First, the measure did not identify many relationships between the variables. Second, an abbreviation "nan" (not a number), which denotes an undefined numeric value, is given in the table. In this case it is the result of the operations in the form of x/0. For that reason, it is worth

employing Ley and Steel's measure (J_{LS}), for which such problems are not present. The values of J_{LS} are located above the primary diagonal in Table 4.

The values of measure J_{LS} better justify the results obtained in section 4.3. The measure identifies 3 pairs of strong substitutes, 14 pairs of significant substitutes and 5 of significant complements. The J_{LS} measure indicates that the participation in the European Union and the Eurozone are either strong or significant substitutes for all the remaining variables. It explains why those variables themselves, despite their strong position in the literature and empirical analyses in the past, turned out to be fragile in the analysis described in section 4.3. Similarly to the J_{DW} measure, J_{LS} classified border and language dummy, as well as real GDP *per capita* and similarity of production structures as significant substitutes. Geographical distance was labelled complement of *POPprod* and *RGDPdist*.

Finally, J_{LS} captured the complementary relationship between *RGDPprod*, *POPprod* and *RGDPpc*. This might help provide two explanations for the negative coefficient on *POPprod*. Firstly, the higher the real GDP product, the bigger the economies and the greater their capacity to trade. At the same time, the higher the population product, the lower GDP *per capita*, and capacity for purchasing of individuals, which could explain negative coefficient on *POPprod*. This effect is present only if *RGDPprod* and *POPprod* are both present in the model. In this instance, *RGDPdist* allows one to control for structural similarity (in terms of both production and consumption) and participation in the EU or the Eurozone.

The second explanation relies upon economies of scale: the bigger the countries, the higher their capacities to explore economies of scale internally and lower the need to trade with outside world. In that instance, *RGDPprod* captures countries capacity to trade and *POPprod* captures their capacity to explore economies of scale internally. In this case, *RGDPdist* additionally allows for controlling differences in welfare between nations.

Therefore, the application of the measure allows one to explain all the results that defy the predictions made according to the theory. It also confirms the criticism levelled against Dopplehofer and Weeks' measure by Ley and Steel. J_{LS} is not only free form computational difficulties of J_{DW}, but also provides better explanations to the obtained results.

5. Conclusions

The following study presents the idea of Bayesian approach to statistics and econometrics, as well as the benefits coming from combining knowledge obtained on the basis of analysis of different models. In the first part, the BMA structure was described together with its most important statistics and g prior, as well as prior model proposals. The second part outlined jointness measures that were put forward by Ley and Steel, as well as Dopplehofer and Weeks.

The empirical part presents the results obtained from the analysis of the determinants of bilateral international trade. The application of Bayesian Model Averaging enabled the identification of four robust determinants: geographical distance, real GDP product, population product and real GDP *per capita* distance. Those four variables are robust to changes in both g prior and model size prior. Language and border dummy, similarity of production structures and participation in the EU were classified as robust for some prior specifications of BMA.

The applied procedure also showed that the model that is the closest to the true one is the model containing the following five independent variables: geographical distance, real GDP and population product, real GDP *per capita* distance and the language dummy. All variables, except for population product, have coefficient signs predicted by the theory. Owing to the application of Ley and Steel's jointness measure, it was possible to explain why some variables firmly rooted in theory were classified as fragile. Participation in the EU and the Eurozone are characterized by substitutional relationship with all other variables. Fragile border dummy and similarity of production structures are substitutes with language dummy and real GDP *per capita* distance respectively, *ergo* contained the same information as the variables classified as robust.

Finally, the complementary relationship between real GDP product and population product enabled two possible explanations of the negative sign of the population product coefficient to be proposed. The first uses the welfare effect reflected in real GDP *per capita*, and the second points to the exploitation of internal economies of scale. It is worth mentioning that the performed exercise demonstrated the superiority of Ley and Steel's jointness measure over the one introduced by Dopplehofer and Weeks.

REFERENCES

ANDERSON, J., (1979). A Theoretical Foundation for the Gravity Model, The American Economic Review. 69 (1), pp. 106–116.

ANDERSON, J., (2011). The Gravity Model, Annual Review of Economics, 3, pp. 133–160.

ANDERSON, J., VAN WINCOOP, E., (2003). Gravity with Gravitas: A Solution to the Border Puzzle, The American Economic Review, 93 (1), pp. 170–192.

BŁAŻEJOWSKI, M., KWIATKOWSKI, J., (2015). Bayesian Model Averaging and Jointness Measures for gretl, Journal of Statistical Software, 68 (5), pp. 1–24.

BŁAŻEJOWSKI, M., KWIATKOWSKI, J., (2016). Bayesian Model Averaging in the Studies on Economic Growth in the EU Regions – Application of the gretl BMA package, Economics and Sociology, 9(4), pp. 168–175.

BROCK, W. A., DURLAUF, S. N., (2001). Growth Empirics and Reality, World Bank Economic Review, 15 (2), pp. 229–272.

DOPPELHOFER, G., WEEKS, M., (2009). Jointness of Growth Determinants, Journal of Applied Econometrics, 24 (2), pp. 209–244.

DIXIT, A., STIGLITZ, J., (1997). Monopolistic Competition and Optimum Product Diversity, The American Economic Review, 67 (3), pp. 297–308.

EGGER, P., (2002). An Econometric View on the Estimation of Gravity Models and the Calculation of Trade Potentials, The World Economy, 25 (2), pp. 297–312.

EICHER, T., PAPAGEORGIOU, C., RAFTERY, A. E., (2011). Determining Growth Determinants: Default Priors and Predictive Performance in Bayesian Model Averaging, Journal of Applied Econometrics, 26 (1), pp. 30–55.

FERNÁNDEZ, C., LEY, E., STEEL, M., (2001). Benchmark priors for Bayesian model averaging, Journal of Econometrics, 100 (2), pp. 381–427.

FOSTER, D., GEORGE, E., (1994). The Risk Inflation Criterion for Multiple Regression, The Annals of Statistics, 22 (4), pp. 1947–1975.

FELDKIRCHER, M., ZEUGNER, S., (2009). Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging, IMF Working Paper, 202, pp. 1–39.

GRAY, H., (1980). The Theory of International Trade among Industrial Countries, Weltwirtschaftliches Archiv, 16 (3), pp. 447–470.

GRUBEL, H., LLOYD, P., (1971). The Empirical Measurement of Intra-Industry Trade, The Economic Record, 47 (120), pp. 494–517.

GRUBEL, H., LLOYD, P., (1975). Intra-industry trade: the theory and measurement of international trade in differentiated products, Wiley, New York.

HELPMAN, E., (1981). International trade in the presence of product differentiation. Economies of scale and monopolistic competition: Chamberlian-Heckscher-Ohlin approach, Journal of International Economics, 11, pp. 305–340.

HESTON. A., SUMMERS, R., ATEN, B., (2012). Penn World Table Version 7.1, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

IMBS, J., WACZIARG, R., (2003). Stages of Diversification, The American Economic Review, 93 (1), pp. 63–86.

KASS, R., WASSERMAN, L., (1995). A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion, Journal of the American Statistical Association, 90 (431), pp. 928–934.

KRUGMAN, P., (1979). Increasing Returns, Monopolistic competition, and International Trade, Journal of International Economics, 9, pp. 469–479.

KRUGMAN, P., (1980). Scale Economies, Product Differentiation and the Pattern of Trade, The American Economic Review, 70 (5), pp. 950–959.

KRUGMAN, P., (1991). Geography and Trade, The MIT Press, Cambridge, MA.

LANCASTER, K., (1980). Intra-Industry Trade under Perfect Monopolistic Competition, Journal of International Economics, 10 (2), pp. 151–175.

LEY, E., STEEL, M., (2007). Jointness in Bayesian variable selection with applications to growth regression, Journal of Macroeconomics, 29 (3), pp. 476–493.

LEY, E., STEEL, M., (2009). On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regressions, Journal of Applied Econometrics, 24 (4), pp. 651–674.

LEY, E., STEEL, M., (2012). Mixtures of g-priors for Bayesian model averaging with economic applications, Journal of Econometrics, 171 (2), pp. 251–266.

LINDER, S., (1961). An Essay on Trade Transformation, Wiley, New York.

MIN, C., ZELLNER, A., (1993). Bayesian and non-Bayesian methods for combining models and forecasts with application to forecasting international growth rates, Journal of Econometrics, 56 (1-2), pp. 89–118.

PRÓCHNIAK, M., WITKOWSKI, B., (2012). Konwergencja gospodarcza typu β w świetle bayesowskiego uśredniania oszacowań, Bank i Kredyt, 43 (2), pp. 25–58.

PRÓCHNIAK, M., WITKOWSKI, B., (2014). The application of Bayesian model averaging in assessing the impact of the regulatory framework on economic growth, Baltic Journal of Economics, 14 (1-2), pp. 159–180.

SALA-I-MARTIN, X., DOPPELHOFER, G., MILLER, R., (2004). Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach, The American Economic Review, 94, (4), pp. 813–835.

WHITTAKER, J., (2009). Graphical Models in Applied Multivariate Statistics, Wiley, Chichester.

ZELLNER, A., (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions. In: Goel PK, Zellner A (eds.). Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics 6. Elsevier, New York, pp. 233–243.

ZEUGNER, S., FELDKIRCHER, M., (2015). Bayesian Model Averaging Employing Fixed and Flexible Priors: The BMS Package for R, Journal of Statistical Software, 68 (4), pp. 1–37.

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 413–432, DOI 10. 21307

ON ASYMMETRY OF PREDICTION ERRORS IN SMALL AREA ESTIMATION

Tomasz Żądło[1]

ABSTRACT

The mean squared error reflects only the average prediction accuracy while the distribution of squared prediction error is positively skewed. Hence, assessing or comparing accuracy based on the MSE (which is the mean of squared errors) is insufficient and even inadequate because we should be interested not only in the average but in the whole distribution of prediction errors. This is the reason why we propose to use different than MSE measures of prediction accuracy in small area estimation. In the prediction accuracy comparisons we take into account our proposal for the empirical best predictor, which is a generalization of the predictor presented by Molina and Rao (2010). The generalization results from the assumption of a longitudinal model and possible changes of the population and subpopulations in time.

Key words: empirical best predictor, prediction errors, small area estimation.

1. Introduction

Nowadays, estimates of the population and large subpopulations characteristics are not sufficient for decision-makers. They require accurate estimates for subpopulations with small or even zero sample sizes. However, because of cost constraints, it is not possible to increase sample sizes continuously to make the estimation of smaller and smaller subpopulations possible using classical methods. The problem is solved using small area estimation methods "borrowing strength" from other subpopulations or time periods. There are three aims of our paper, with the first the main one and the second and the third the supplementary aims.

Firstly, our observation that the distribution of squared prediction errors has strong positive asymmetry (values of the third standardized moments obtained in the simulation study based on real data are presented in Table 1) has become a focus of our attention. It implies that their mean (known as MSE – Mean

[1] University of Economics in Katowice, Faculty of Management, Katowice, Poland. E-mail: tomasz.zadlo@ue.katowice.pl.

Squared Error) does not have to be a good measure of prediction accuracy in terms of the average and, what is in our opinion even more important, their whole distribution should be studied. Hence, the main purpose of the paper is our proposal for assessing the prediction accuracy using new univariate and multivariate prediction measures based on quantiles of the distribution of absolute prediction errors. It will be shown that even if the accuracies of two predictors in terms of the average are similar, the accuracy comparison based on right tails of distributions of absolute prediction errors can give different results (which will be presented in e.g. Figure 4).

Secondly, in the accuracy comparisons resulting from the main aim we will include a generalization of the Empirical Best Predictor (EBP) proposed by Molina and Rao (2010). They proposed the predictor under a model assumed for data from surveys conducted in one period. We will propose a predictor assuming a longitudinal model. It means that in the case of longitudinal surveys we will be able to use information from previous periods to increase the prediction accuracy in the period of interest.

Thirdly, in the proposed longitudinal model we will take into account that the population and subpopulations may change in time. It will cover many longitudinal models known from small area estimation (which are special cases of the general linear mixed models) including models studied by:

- Saei and Chambers (2003), who assume mutually independent two random effects (domain-specific and time-specific) and random components (and the generalization, where AR(1) process is assumed for time-specific random effects),
- Saei and Chambers (2003), where domain-and-time-specific random effects with independent distributions in domains and AR(1) model in time are taken into account,
- Stukel and Rao (1999) and Nissinen (2009) p. 22 with mutually independent two random effects (domain-specific and element-specific) and random components,
- Nissinen (2009) p. 60, who assumes independent domain-specific random effects and autocorrelated (assuming AR(1)) random components in time,
- Molina, Morales, Pratesi, Tzavidis (2010) pp. 143-180 with independent domain-specific random effects, independent for domains and autoccorelated (assuming AR(1)) in time domain-and-time-specific random effects and heteroscedastic random components.

In the simulation study the properties of the proposed predictor (in terms of MSE and the proposed accuracy measures) will be studied under the proposed model taking into account the model misspecification as well.

2. Alternative prediction accuracy measures

In the case of positive asymmetry usually the mean is not the only measure used to describe the distribution. The MSE is the mean of squared errors (which have positive or even strong positive asymmetry) and it is usually used as the only accuracy measure. Moreover, a better predictor is usually defined as the one with smaller MSE. Żądło (2013) proposed a new measure of prediction accuracy Quantile of Absolute Prediction Error defined for the problem of prediction in the *d*th domain as follows:

$$QAPE(p) = \inf\left\{x : P\left(\left|U_d\right| \le x\right) \ge p\right\}, \tag{1}$$

where $U_d = \hat{\theta}_d - \theta_d$ is the prediction error of $\hat{\theta}_d$, which is the predictor of θ_d in the *d*th domain. It means that (1) is the quantile of order *p* of $\left|U_d\right|$. It means that at least $p100\%$ of realizations of absolute prediction errors in the *d*th domain are smaller or equal to $QAPE(p)$. In Żądło (2013) it was used to measure prediction accuracy of the empirical best linear unbiased predictor.

Żądło (2015) proposes multivariate versions of (1), which allow us to measure and compare accuracy in the case of simultaneous prediction in all of domains. It can be treated as the alternative to the average mean squared error studied, e.g. by Fabrizi and Trivisano (2010). Let prediction errors in D domains be denoted by $U_d = \hat{\theta}_d - \theta_d$, where $d = 1,2,...,D$. Let us define the multivariate version of $QAPE$ as follows:

$$MQAPE(p) = \inf\left\{x : \sum_{d=1}^{D} P\left(\left|U_d\right| \le x\right) \ge Dp\right\}. \tag{2}$$

It means that it is the quantile of order *p* of a distribution of a mixture of random variables $\left|U_1\right|,...,\left|U_d\right|,...,\left|U_D\right|$ with equal weights. It means that at least $p100\%$ of realizations of absolute prediction errors in all domains are smaller or equal to $MQAPE(p)$.

Let relative prediction errors be denoted by $W_d = \frac{U_d}{\theta_d} = \frac{\hat{\theta}_d - \theta_d}{\theta_d}$, where $d = 1,2,...,D$. Let us define relative $MQAPE$ as follows:

$$rMQAPE(p) = \inf\left\{x : \sum_{d=1}^{D} P\left(\left|W_d\right| \le x\right) \ge Dp\right\}. \tag{3}$$

It means that it is the quantile of order *p* of a distribution of a mixture of random variables $\left|W_1\right|,...,\left|W_d\right|,...,\left|W_D\right|$ with equal weights. It means that at least $p100\%$ of realizations of moduli of relative prediction errors in all domains are smaller or equal to $rMQAPE(p)$.

The estimation of (1), (2) and (3) is possible using a well-known parametric bootstrap method studied, e.g. by González-Manteiga et al. (2007, 2008) and

Molina and Rao (2010). Using the method, the estimator of the MSE is given by the mean of squared bootstrap realizations of prediction errors. Similarly, by computing quantiles of bootstrap realizations of:

- moduli of prediction errors in one of domains we can estimate (1),
- moduli of prediction errors in all of domains we can estimate (2) and
- moduli of relative prediction errors in all of domains we can estimate (3).

3. Model and predictor

We consider longitudinal data in periods $t=1,2,...,M$, where the population of size N_t in the period t is denoted by Ω_t. The population is divided into D disjoint subpopulations (domains) Ω_{dt} each of size N_{dt}, where $d=1,2,...,D$. A sample in the period t of size n_t is denoted by s_t. Let $s_{dt}=s_t\cap\Omega_{dt}$ and $\overline{\overline{s}}_{dt}=n_{dt}$. The d*th domain of interest in the period of interest t* will be denoted by $\Omega_{d^*t^*}$. Let $\Omega_{rdt}=\Omega_{dt}-s_{dt}$, $N_{rdt}=N_{dt}-n_{dt}$, $\bigcup_{t=1}^{M}\Omega_t=\Omega$, $\overline{\overline{\Omega}}=N$, $\bigcup_{t=1}^{M}\Omega_{dt}=\Omega_d$, $\overline{\overline{\Omega}}_d=N_d$, $\bigcup_{t=1}^{M}\Omega_{rdt}=\Omega_{rd}$, $\overline{\overline{\Omega}}_{rd}=N_{rd}$, $\bigcup_{t=1}^{M}s_t=s$, $\overline{\overline{s}}=n$, $\bigcup_{t=1}^{M}s_{dt}=s_d$, $\overline{\overline{s}}_d=n_d$.

Let M_{id} be the number of periods when the ith population element belongs to the dth domain and m_{id} – the number of periods when the ith population element (which belongs to the dth domain) is observed. Let $M_{rid}=M_{id}-m_{id}$. It is assumed that the population may change in time and that one population element may change its domain affiliation in time. Hence, sets of population elements Ω_d (where $d=1,2,...,D$) may overlap.

The assumption that one population element may change its domain affiliation in time is very important in practice of longitudinal surveys. For example, let us consider the population of households and the division of the population into domains made according to the household size. In this case we should assume that some households can change their sizes in time, which causes the change of the domain affiliation. If a human population is under the study one may be interested in its characteristics for subpopulations defined according to some social or economic criteria. In the case of business surveys the population of firms may be divided into subpopulations according to some economic or financial criteria, what can imply even stronger changes of domains affiliations.

Values of the variable of interest (or the variable of interest after a transformation) are realizations of Y_{idj}'s for the ith population element, which belongs to the dth domain in the period t_{ij}, where $i=1,2,...,N$; $j=1,2,...,M_{id}$;

$d = 1,2,...,D$. The vector $\mathbf{Y_{id}} = \left[Y_{idj} \right]_{M_{id} \times 1}$ will be called the profile and the vector $\mathbf{Y_{sid}} = \left[Y_{idj} \right]_{m_{id} \times 1}$ will be called the sample profile. Let the vector $\mathbf{Y_{rid}} = \left[Y_{idj} \right]_{M_{rid} \times 1}$ be the profile for non-observed realizations of random variables.

Let us introduce assumptions of the following longitudinal model, which is a special case of the general linear mixed model (e.g. Datta and Lahiri, 2000). The difference is introduced in the sizes of matrices, which allows us to take into account longitudinal data and possible changes in population and subpopulations in time. We assume that

$$\begin{cases} \mathbf{Y} = \mathbf{X\beta} + \mathbf{Zv} + \mathbf{e} \\ D^2_{\xi}(\mathbf{v}) = \mathbf{G}(\boldsymbol{\delta}) \\ D^2_{\xi}(\mathbf{e}) = \mathbf{R}(\boldsymbol{\delta}) \\ Cov_{\xi}(\mathbf{v}, \mathbf{e}) = \mathbf{0} \end{cases}, \tag{4}$$

where ξ is the superpopulation model, $\mathbf{Y} = col_{1 \le d \le D} col_{1 \le i \le N_d}(\mathbf{Y_{id}})$, $\mathbf{e} = col_{1 \le d \le D} col_{1 \le i \le N_d}(\mathbf{e_{id}})$, where $\mathbf{e_{id}}$ is the $M_{id} \times 1$ vector of random components, $\mathbf{X} = col_{1 \le d \le D} col_{1 \le i \le N_d}(\mathbf{X_{id}})$, where $\mathbf{X_{id}}$ is the known matrix of size $M_{id} \times p$, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown parameters, $\mathbf{Z}$ is the known matrix of size $\sum_{i=1}^{N} \sum_{d=1}^{D} M_{id} \times h$, $\mathbf{v}$ is the vector of random effects of size $h \times 1$, $\boldsymbol{\delta}$ is the vector of q unknown in practice parameters called variance components.

Let us consider the following decomposition of the vector $\mathbf{Y}$:

$$\mathbf{Y} = \left[\mathbf{Y}_s^T \quad \mathbf{Y}_r^T \right]^T, \tag{5}$$

where $\mathbf{Y}_s$ is the vector of size $\sum_{i=1}^{N} \sum_{d=1}^{D} m_{id} \times 1$ of random variables, whose realizations are known, and $\mathbf{Y}_r$ is the vector of size $\sum_{i=1}^{N} \sum_{d=1}^{D} M_{rid} \times 1$ of random variables, which are not observed in the longitudinal survey. Then,

$$D^2_{\xi}(\mathbf{Y}) = D^2_{\xi} \begin{bmatrix} \mathbf{Y}_s \\ \mathbf{Y}_r \end{bmatrix} = \mathbf{V}(\boldsymbol{\delta}) = \begin{bmatrix} \mathbf{V}_{ss}(\boldsymbol{\delta}) & \mathbf{V}_{sr}(\boldsymbol{\delta}) \\ \mathbf{V}_{rs}(\boldsymbol{\delta}) & \mathbf{V}_{rr}(\boldsymbol{\delta}) \end{bmatrix}, \tag{6}$$

where under (4):

$$\mathbf{V}(\boldsymbol{\delta}) = \mathbf{ZG}(\boldsymbol{\delta})\mathbf{Z}^T + \mathbf{R}(\boldsymbol{\delta}). \tag{7}$$

Let us consider the problem of predicting any given function of the random vector $\mathbf{Y}$ denoted by $\theta(\mathbf{Y})$ or shortly by θ. Among predictors $\hat{\theta}$ of θ, the Best Predictor (BP) is defined as the one, which minimizes (e.g. Molina and Rao 2010):

$$MSE_{\xi}(\hat{\theta}) = E_{\xi}(\hat{\theta} - \theta)^2 . \quad (8)$$

Hence, it is given by:

$$\hat{\theta}_{BP} = E_{\xi}(\theta \mid \mathbf{Y}_s), \quad (9)$$

which means that it may be obtained as a conditional expected value of θ assuming that the conditional distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$ is known.

We assume that the conditional distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$ can be derived (the example is presented in *Remark 1* in this section). In practice, it depends on the vector of unknown parameters, which will be denoted by $\boldsymbol{\tau}$. If we replace the parameters by their estimators, we obtain the Empirical Best Predictor (EBP) denoted by $\hat{\theta}_{EBP}$. Hence, the value of the EBP of $\theta(\mathbf{Y})$ can be obtained through the Monte Carlo approximation algorithm presented below (for prediction in surveys conducted in one period see Molina and Rao 2010).

(a) We estimate $\boldsymbol{\tau}$ based on the realization of $\mathbf{Y}_s$ and we obtain the value of the estimator denoted by $\hat{\boldsymbol{\tau}}$.

(b) Assuming that the distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$ can be derived, we generate L vectors $\mathbf{Y}_r$ (denoted by $\mathbf{Y}_r^{(l)}$, where $l = 1,2,...,L$) from the distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$, where the unknown vector $\boldsymbol{\tau}$ is replaced by $\hat{\boldsymbol{\tau}}$.

(c) We make L vectors denoted by $\mathbf{Y}^{(l)}$, where $\mathbf{Y}^{(l)} = \left[\mathbf{Y}_s^T \quad \mathbf{Y}_r^{(l)T}\right]^T$ and $l = 1,2,...,L$, what means that L vectors $\mathbf{Y}^{(l)}$ include the same realization of $\mathbf{Y}_s$ and different realizations of $\mathbf{Y}_r$.

(d) The value of the EBP of $\theta(\mathbf{Y})$ is obtained as follows: $\hat{\theta}_{EBP} = L^{-1}\sum_{l=1}^{L}\theta(\mathbf{Y}^{(l)})$.

Due to the estimation of an unknown in practice vector of model parameters denoted by $\boldsymbol{\tau}$, the resulting predictor generally is not unbiased and it does not minimize the MSE (as the BP) but its value should be very close to the BP. Its MSE estimator, which takes into account the uncertainty resulting from the estimation of $\boldsymbol{\tau}$, can be obtained using parametric bootstrap method as in Molina and Rao (2010), where their model is replaced by (4).

Remark 1. If we additionally assume that the vector $\mathbf{Y}$ (which may be the vector of the variable of interest after a transformation) is normally distributed, which can be written as follows $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\delta}))$ (where under (4) $\mathbf{V}(\boldsymbol{\delta})$ is given by (7)), then $\boldsymbol{\tau} = \left[\boldsymbol{\beta}^T \quad \boldsymbol{\delta}^T\right]^T$ and

$$\mathbf{Y}_r \mid \mathbf{Y}_s \sim N\left(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right), \mathbf{V}_{rr}(\boldsymbol{\delta}) - \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{V}_{sr}(\boldsymbol{\delta})\right). \quad (10)$$

Hence, in the step (b) of the procedure presented above, vectors $\mathbf{Y}_r^{(l)}$, where $l = 1,2,...,L$ are generated based on (10), where parameters are replaced by their estimates.

The idea of using EBPs was presented earlier by Molina and Rao (2010) but for studies conducted in one period. They study the general case assuming the general linear mixed model for studies conducted in one period. In the special case of their considerations $\mathbf{Y}$ is the vector of the variable of interest after the following transformation: $\mathbf{Y} = T(\ddot{\mathbf{Y}})$, where $\ddot{\mathbf{Y}}$ is the variable of interest, and

$$T(\ddot{\mathbf{Y}}) = \ln(\ddot{\mathbf{Y}} + c), \tag{11}$$

where c is a constant. Then, they study the following model

$$Y_{id} = \mathbf{x}_{id}^T \boldsymbol{\beta} + v_d + e_{id}, \tag{12}$$

where $d = 1,2,...,D$; $i = 1,2,...,N$, $v_d \overset{iid}{\sim} N(0, \sigma_v^2)$, $e_{id} \overset{iid}{\sim} N(0, \sigma_e^2)$, e_{id} and v_d are independent, $\boldsymbol{\delta} = \begin{bmatrix} \sigma_e^2 & \sigma_v^2 \end{bmatrix}^T$.

4. Simulation study – real data

We consider data on $N = 378$ Polish poviats (NUTS-4 level) from years 2010-2012 (M=3). We have excluded one observation because of the lack of the data and one outlying observation (Warsaw). The problem of prediction of totals of the sold production of industry in $D = 16$ domains (voivodships – NUTS-2 level) for companies with at least 10 employees is considered. The number of companies with at least 10 employees is the auxiliary variable. In the first period a sample of 38 poviats is drawn at random with probabilities proportional to the values of the auxiliary variable. Sample sizes in the domains are random and equal from 0 to 5 (with mean 2.375). The balanced panel survey is considered – elements sampled in the first period are observed until the end of the longitudinal survey (which gives 114 observations in 3 periods).

Because empirical best predictors are studied, the distribution of the variable of interest must be assumed. We consider the transformation of the variable of interest given by (11) and logarithmic transformation of the auxiliary variable. To test the distribution of the variable of interest, we use the transformation of residuals based on the Cholesky decomposition of the inverse of variance-covariance matrix (see, e.g. Jacqmin-Gadda et al. 2007). For the model chosen based on the AIC and BIC criteria (more details will be presented in the next paragraph) p-values for Shapiro-Wilk, Jarque-Bera and adjusted Jarque-Bera tests obtained for the sample equal 0.2297; 0.6046 and 0.446 respectively. For the considered model but without the transformations of both variables, p-values for the tests of normality were smaller than 10^{-12}. But if we test normality based on the whole population data (based on $M \times N = 3 \times 378 = 1134$ observations) in both cases (with and without transformations of variables) we should reject the

null hypothesis on normality. That is why the problem of model misspecification will be taken into account in the simulation study as well.

To choose the appropriate model we consider different models: classic and mixed linear models with and without the auxiliary variable, with and without constant, nested-error models and models with random slopes. For mixed models with random slope we consider time-specific, domain-specific, time-and-domain-specific and finally profile-specific random effects. In models with nested errors we consider one random effect (time-specific, domain-specific, time-and-domain-specific and profile-specific) or two random effects (firstly: domain-specific and profile-specific; secondly: domain-specific and domain-and-time-specific). The model with the smallest both AIC and BIC criteria was the following model studied earlier by Stukel and Rao (1999) and Nissinen (2009) p. 22:

$$Y_{idt} = x_{idt}\beta_1 + \beta_0 + u_d + v_{id} + e_{idt}, \tag{13}$$

where Y_{idt} is the variable of interest after transformation (11), x_{idt} is the auxiliary variable after logarithmic transformation, $i = 1,2,...,N$, $d = 1,2,...,D$, $t = 1,2,...,M$, u_d, v_{id} and e_{idt} are mutually independent with zero expected values and variances given by σ_u^2, σ_v^2 and σ_e^2 respectively. Permutation test (with the test statistic given by the loglikelihood) was used to test the significance of the model parameters – at the significance level 0.05 tested parameters were significantly different from zero. Good properties of these tests are presented by Krzciuk and Żądło (2014a, 2014b).

The model-based simulation study was prepared using R software (R Core Team 2016). To mimic the real data, values of the variable of interest after transformation (11) are generated based on the model (13) with one auxiliary variable and the constant, where the parameters of the model are replaced by REML estimates obtained based on all of the observations (sampled and unsampled) of the real data. Hence, both random effects and random components are generated with zero expected values and variances σ_u^2, σ_v^2 and σ_e^2 equal REML estimates based on (13) and the whole population data. Random effects and random components u_d, v_{id} and e_{idt} are generated independently from:

- normal distributions,
- shifted exponential distributions (the third standardized moment is equal to 2),
- shifted gamma distributions (with the value of third standardized moment equal to 4) and
- shifted Pareto distributions (with the value of third standardized moment equal to 5).

It means that in the case of the normal and the shifted exponential distributions, the assumed values of the mean and the variance give explicitly values of the parameters of the distributions used in the simulation study. The case of the shifted gamma and the shifted Poisson distributions is more interesting because it

is possible to set the values of the parameters of these distributions to obtain not only the assumed values of the mean and the variance but also the prespecified value of the third standardized moment (4 - for the shifted gamma and 5 - for the shifted Pareto distributions, as listed above).

In each iteration of the simulation study model parameters are estimated using restricted maximum likelihood, which gives consistent estimates even if the normality assumption is not met (Jiang 1996). The number of iterations equals 5000.

We study properties of the following predictors:

- the empirical best predictor based on the longitudinal model (13) under normality of random effects and random components (EBP),
- the empirical best predictor studied earlier by Molina and Rao (2010) based on the model (12) assumed for transformed data (EBP-MR),
- the empirical best linear unbiased predictor based on the Royall (1976) theorem for the longitudinal model with the smallest AIC and BIC criteria assumed for the untransformed data, i.e. for the mixed model with random regression coefficient with the profile-specific random effect (EBLUP),
- the synthetic regression estimator given by (SYNT-REG) given by (e.g. Bracha 1996, p. 260):

$$N_{dt}\left(\sum_{i\in s_t}\pi_{ti}^{-1}\right)^{-1}\sum_{i\in s_t}y_{ti}\pi_{ti}^{-1}+N_{dt}B\left(N_{dt}^{-1}\sum_{i\in\Omega_{dt}}x_{t*i}-\left(\sum_{i\in s_t}\pi_{ti}^{-1}\right)^{-1}\sum_{i\in s_t}x_{ti}\pi_{ti}^{-1}\right),$$

for $d=1,2,...,D$, where

$$B=\frac{\sum_{i\in s_t}\left(x_{ti}-\left(\sum_{i\in s_t}\pi_{t*i}^{-1}\right)^{-1}\sum_{i\in s_t}x_{ti}\pi_{ti}^{-1}\right)\left(y_{ti}-\left(\sum_{i\in s_t}\pi_{ti}^{-1}\right)^{-1}\sum_{i\in s_t}y_{ti}\pi_{ti}^{-1}\right)\pi_{ti}^{-1}}{\sum_{i\in s_t}\left(x_{ti}-\left(\sum_{i\in s_t}\pi_{ti}^{-1}\right)^{-1}\sum_{i\in s_t}x_{ti}\pi_{ti}^{-1}\right)^{2}\pi_{ti}^{-1}},$$

and π_{ti} is the inclusion probability of the ith population element in the period t.

Because of small sample sizes in the domains (in some domains: 0) we study only indirect predictors and estimators. In each out of 5000 Monte Carlo iterations values of both empirical best predictors are computed based on $L=200$ generated population vectors.

Relative prediction biases for the considered estimators and predictors are presented in Figure 1. Each boxplot presents $D=16$ values of biases of a predictor of $D=16$ domains totals. For example, the values presented in the top-left boxplot are from ca 0.3% to ca 2.1%. The value 2.1% means that for one of the domains the relative bias of EBP predictor equals 2.1% (in this domain the value of the predictor is larger than the domain total on average by 2.1%). If the distributions of the random components and random effects for the transformed variables are normal, the biases of all predictors and estimators are small. EBP in

this case is used under the correctly specified longitudinal model – the transformation of the variables, the assumed normal distribution and the assumed formula of the model (13) are correct. EBP-MR is used under the misspecified formula of the model (assumed for one period instead of the longitudinal data) but under the correct transformation of the variables and assuming correct (i.e. normal) distribution. Both EBLUP and SYNT-REG do not take into account the transformation of the variables. If the distribution is asymmetric, the biases are very large in many cases.

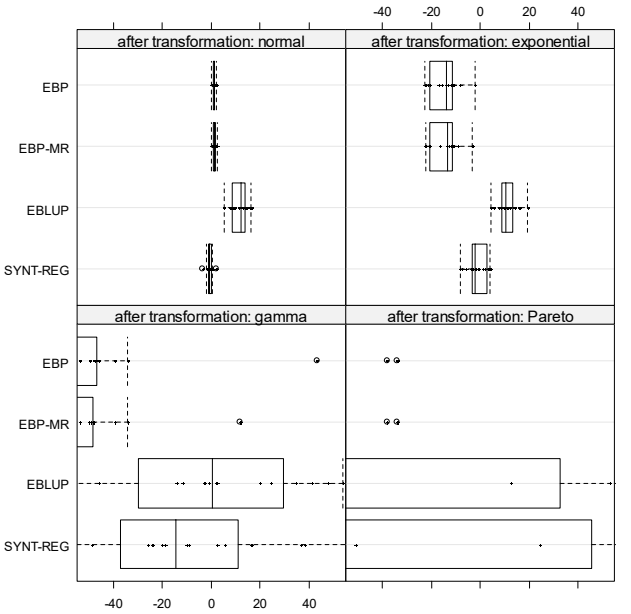


Figure 1. Relative prediction biases (in %) for different predictors and different distributions of random effects and random components (each boxplot presents values for D=16 domains)

Results of the comparisons of the accuracy between the proposed empirical best predictor EBP and other estimators and predictors based on the MSE are presented in Figure 2. Each boxplot presents $D = 16$ values of ratios of the MSE of a predictor to the MSE of EBP for $D = 16$ domains totals. For example, the values presented in the top-left boxplot are from ca 1.02 to ca 1.06. The value 1.06 means that for one of the domains the ratio of the MSE of EBP-MR predictor to the MSE of EBP predictor equals 1.06 (in this domain the value of the MSE of EBP-MR is higher than the MSE of EBP by 6%). If we compare the MSE of EBP-MR to the MSE of the EBP for other distributions, we see that the values of ratios are also very close to 1. In all of the cases the maximum gain in accuracy due to the usage of the proposed predictor measured by the MSE is smaller than 10%. The reasons of the results will be studied in the next section.

What is interesting, in the results presented in Figure 2 is the lack of stability comparing results for different distributions of random effects and random

components. The reason of unstable results is strong positive asymmetry of the distribution of absolute prediction errors, especially if the distribution of random effects and random components is not normal (see values of the third standardized moments of absolute prediction errors presented in Table 1). Because the values of the prediction MSE (the values of the mean of squared errors) are strongly affected by outlying absolute prediction errors, results for alternative measures of prediction accuracy defined in section 2 will be presented as well.

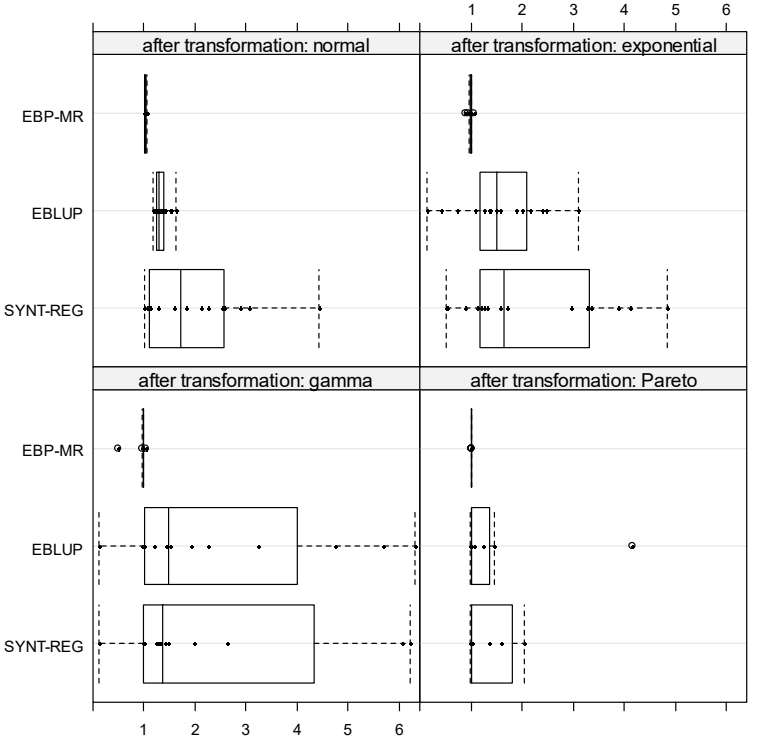


Figure 2. Values of MSE(.)/MSE(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for D=16 domains)

Table 1. Third standardized moments of absolute prediction errors for different predictors and different distributions of random effects and random components (minimum and maximum for D=16 domains)

	After transformation:			
	normal	shifted exponential	shifted gamma	shifted Poisson
SYN-REG	1.3-4.0	8.3-52.7	24.4-70.6	33.1-70.7
EBLUP	1.2-3.6	12.4-55.4	28.7-70.6	34.3-70.7
EBP-MR	1.6-4.4	13.8-61.6	17.5-70.7	44.7-70.7
EBP	1.5-4.5	13.7-61.6	17.5-70.7	44.7-70.7

Firstly, we will compare the accuracy of the predictors based on the same real data. In this case we will use $QAPE(p)$ for $p=(0.5, 0.75, 0.9, 0.95)$ for each domain. It is worth mentioning that the results presented in Figure 3 (and Figures 6-8 in Appendix) are more stable than the results presented in Figure 2. Each boxplot in Figure 3 presents $D=16$ values of ratios of the $QAPE(0.5)$ of a predictor to the $QAPE(0.5)$ of EBP for $D=16$ domains totals. As it was defined and discussed in the section 3, $QAPE(0.5)$ is the median of absolute prediction errors. For example, the values presented in the top-left boxplot are from ca 1 to ca 1.05. The value 1.05 means that for one of the domains the ratio of the $QAPE(0.5)$ of EBP-MR predictor to the $QAPE(0.5)$ of EBP predictor equals 1.05 (in this domain the value of the $QAPE(0.5)$ of EBP-MR is higher than the $QAPE(0.5)$ of EBP by 5%).

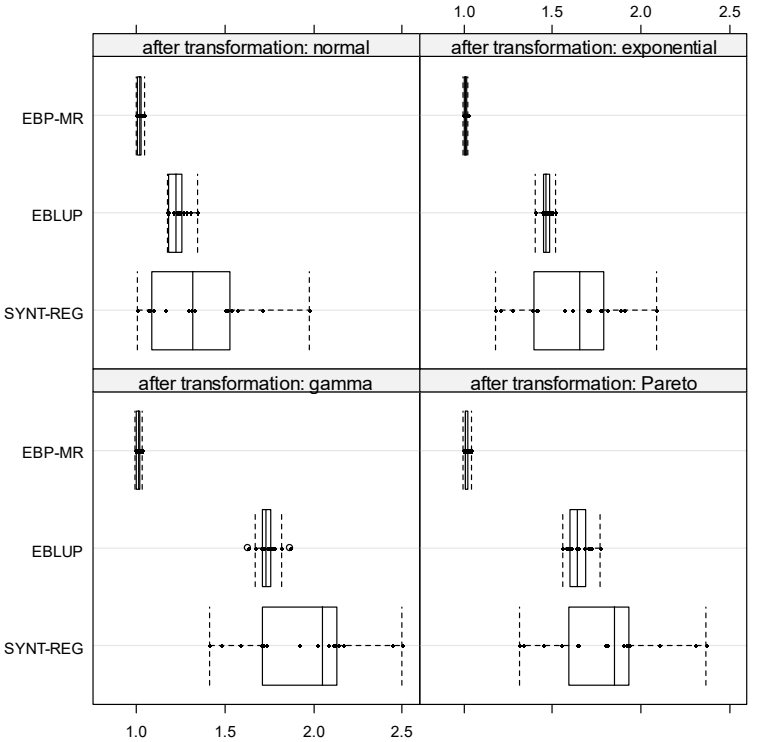


Figure 3. Values of QAPE0.50(.)/QAPE0.50(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for *D*=16 domains)

Additionally, in Figure 4 the values of $rMQAPE(p)$ for $p=(0.5, 0.75, 0.9, 0.95)$ are presented. As an example, we will interpret the value presented by the point in the top-left part of Figure 7 (for EBP under normal

distribution of random effects and random components for the variable of interest after the transformation), which equals $rMQAPE(0.5) = 18.2\%$. It means that at least 50% of moduli of relative prediction errors for all of the domains are smaller or equal to 18.2% and at least 50% of moduli of relative prediction errors for all of the domains are larger or equal to 18.2%. But $rMQAPE(0.5)$ informs only about the average (i.e. median) of absolute prediction errors. If we are interested in the right tail of the distribution of the absolute prediction errors, we can compute $rMQAPE(p)$ for $p > 0{,}5$, e.g. $rMQAPE(0.75) = 32.8\%$, $rMQAPE(0.9) = 51.2\%$ and finally $rMQAPE(0.95) = 67\%$ (see the top-left part of Figure 4 and top-left part of Figure 9 in Appendix).

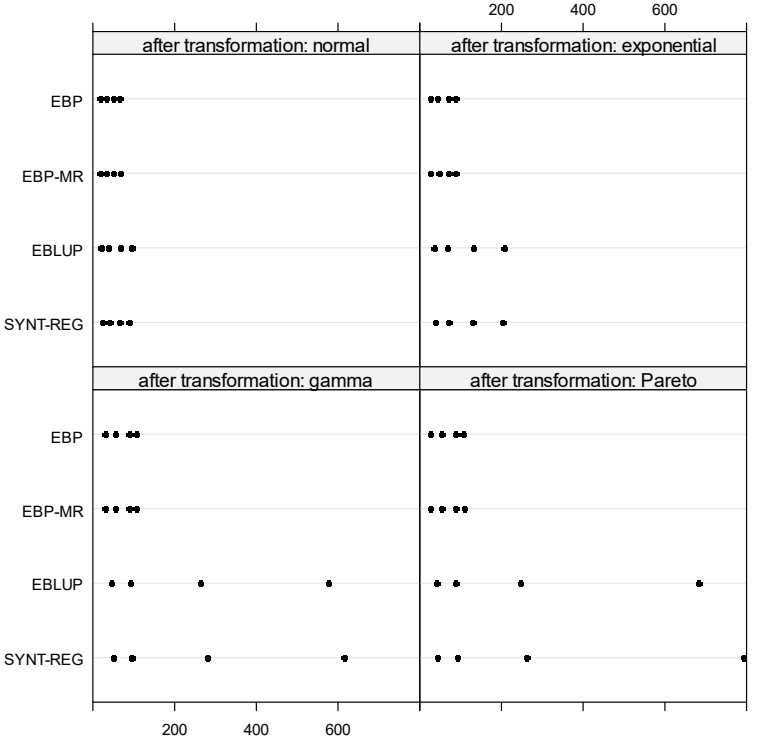


Figure 4. Values of rMQAPE(p) for p=(0.5, 0.75, 0.9, 0.95), different predictors and different distributions of random effects and random components - all results

It should be stressed that although values of $rMQAPE(0.5)$ for each predictor are quite similar even in the case of model misspecification (see the first point for each predictor in Figure 4 or Figure 9 in Appendix), the difference in the accuracy measured in right tails of the distribution of absolute prediction errors by $rMQAPE(0.95)$ can differ substantially especially in the case of model misspecification (see the last point for each predictor in Figure 4). For example in

bottom-right panel in Figure 4, values of $rMQAPE(0.5)$ for EBP and SYNT-REG equal 27% and 45%, respectively, which means that $rMQAPE(0.5)$ for SYNT-REG is 1.67 times higher than for EBP. However, values of $rMQAPE(0.95)$ for EBP and SYNT-REG equal 102% and 776%, respectively, which means that $rMQAPE(0.95)$ for SYNT-REG is 7.6 times higher than for EBP. To sum up, the prediction accuracy measures presented in section 2 give us more detailed information on prediction accuracy, which is not limited to the average values (as in the case of the MSE). What is more, using $QAPE$ we obtain more stable results of accuracy comparisons, especially in the case of model misspecification.

5. Simulation study – artificial data

In the previous section two problems were considered – the comparison of the accuracy and the choice of the appropriate measures of accuracy. One of the conclusions was the small difference in the accuracy (less than 10% in terms of the MSE for all considered distributions) between the proposed empirical best predictor for longitudinal surveys and the empirical best predictor proposed by Molina and Rao (2012) for surveys conducted in one period. To identify the reasons, we compare some results from the previous section (the column "real data" in Table 2) with two additional simulation scenarios, all of them under normality of random effects and random components for data after transformation (11) and assuming model (13) or its special case.

Table 2. Maximum values of ratios of accuracy measures for different simulation scenarios over D=16 domains under normality of random effects and random components

	Simulation scenario		
ratio of accuracy measures	real data	independent values of x	without x
MSE(EBP-MR)/MSE(EBP)	1.058	1.257	1.123
QAPE0.50(EBP-MR)/QAPE0.50(EBP)	1.046	1.119	1.040
QAPE0.75(EBP-MR)/QAPE0.75(EBP)	1.023	1.134	1.028
QAPE0.90(EBP-MR)/QAPE0.90(EBP)	1.029	1.117	1.042
QAPE0.95(EBP-MR)/QAPE0.95(EBP)	1.047	1.154	1.047

In the first scenario (results in Table 2 in the column "independent values of x"), we generate values of the variable of interest based on model (13) with values of all model parameters obtained for the real data (as in the previous

section), but where the real auxiliary variable is replaced by the artificial one. Values of the auxiliary variable were generated independently from shifted gamma distribution assuming real values of the mean, variance and the third standardized moment for each year. In this case the maximum gain in accuracy of our EBP measured by MSE is 25.7% and measured by *QAPE* is higher than 10%. It means that in the case of longitudinal surveys we should use auxiliary variable, which is weakly autocorrelated but even in this case the gain in accuracy will not be very large.

In the second scenario (results in Table 2 in the column "without x") we do not use the uxiliary variable both in the model and at the estimation stage. Hence, we compare prediction accuracy of empirical best predictors only under random parts of models (13) and (12). The accuracy measured by MSE of our EBP is higher by 12.3% compared with EBP-MR (by less than 5% in terms of *QAPE*). The reason is that model (13) chosen based on AIC and BIC for real longitudinal data is quite similar to the model (12) assumed by Molina and Rao (2010). In both models we have domain-specific random effects, although in the case of (13) it additionally implies non-zero covariances between observations within domains in different periods. The main difference between the models is the profile (element)-specific random effect in model (13), but results in the last column of Table 2 show that it does not imply a large gain in prediction accuracy. It means that the larger gain in accuracy can be obtained when the longitudinal model explains the variability of the variable of interest considerably better than the model assumed for one period.

To sum up, in this section based on the Monte Carlo analysis we have identified two reasons of the relatively small gain in accuracy, which was presented in the previous section, comparing our predictor with the predictor proposed by Molina and Rao (2010). Firstly, it has been autocorrelation in time of the auxiliary variable. Secondly, we have presented similarity of the proposed longitudinal model and the model studied by Molina and Rao (2010). Moreover, we have shown that in the studied cases the maximum gain in accuracy comparing these two predictors can be even higher than 25% in terms of MSE.

6. Real data application

In this section we consider values of the same predictors and estimators, the same data and the same sample as discussed in section 4. However, in this case their values are computed once based on the real data (they are not generated as in the simulation studies presented in section 4). Because the whole population data are available, we are able to compare estimates with real values of D=16 domains totals (see Figure 5). The largest differences between estimates and real values for the considered sample are observed for SYNT-REG and EBLUP, whereas the values of EBP-MR and the proposed EBP are very similar.

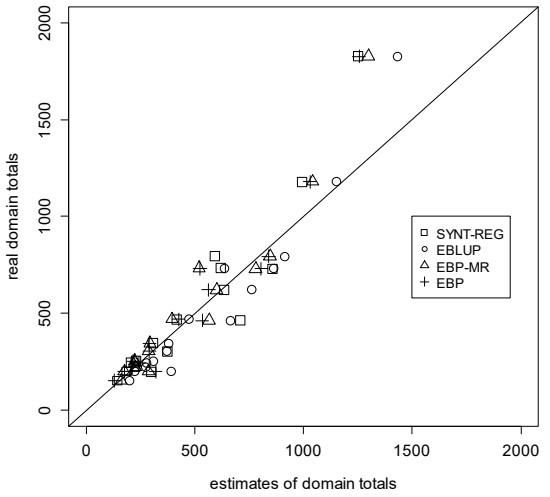


Figure 5. Values of estimates and real domain totals

7. Conclusions

In the paper the problem of assessing and comparing the prediction accuracy is studied. Because of strong positive asymmetry of absolute prediction error, it is shown that prediction accuracy measures alternative to the MSE should be used. These measures allow us to assess the prediction accuracy not limited to the average values and to obtain more stable results of accuracy comparisons, especially in the case of the model misspecification. In the accuracy comparisons based on the Monte Carlo simulation studies our proposal for the empirical best predictor is taken into account. Although its prediction accuracy was only slightly better for the considered data compared with the empirical best predictor proposed by Molina and Rao (2012), we present how to obtain a substantial gain in accuracy. The considerations are also supported by real data application.

REFERENCES

BRACHA, CZ., (1996). Teoretyczne podstawy metody reprezentacyjnej, PWN, Warszawa.

DATTA, G. S., LAHIRI, P., (2000). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems, Statistica Sinica, Vol. 10, pp. 613–627.

FABRIZI, E., TRIVISANO, C., (2010). Robust linear mixed models for Small Area Estimation, Journal of Statistical Planning and Inference, Vol. 140, pp. 433–443.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M.J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, Computational Statistics and Data Analysis, Vol. 51, pp. 2720–2733.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2008). Bootstrap mean squared error of small-area EBLUP, Journal of Statistical Computation and Simulation, Vol. 78, 443–462.

JACQMIN-GADDA, H., SIBILLOT, S., PROUST, C., MOLINA J.-M., THIÉBAUT, R., (2007). Robustness of the Linear Mixed Model to Misspecified Error Distribution, Computational Statistics & Data Analysis, Vol. 51, pp. 5142–5154.

JIANG, J., (1996). REML Estimation: Asymptotic Behavior and Related Topics, The Annals of Statistics, Vol. 24, pp. 255–286.

KRZCIUK M., ŻĄDŁO T., (2014a). On some tests of variance components for linear mixed models, Studia Ekonomiczne, Vol. 189, pp. 77–85.

KRZCIUK M., ŻĄDŁO T., (2014b). On some tests of fixed effects for linear mixed models, Studia Ekonomiczne, Vol. 189, pp. 49–57.

MOLINA, I., RAO, J. N. K., (2010). Small Area Estimation of Poverty Indicators, The Canadian Journal of Statistics, Vol. 38, pp. 369–385.

NISSINEN, K., (2009). Small Area Estimation With Linear Mixed Models For Unit-Level Panel and Rotating Panel Data, University of Jyväskylä Printing House, Jyväskylä.

R CORE TEAM, (2016). R: A Language and Environment For Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

ROYALL, R. M., (1976). The Linear Least Squares Prediction Approach to Two-Stage Sampling, Journal of the American Statistical Association, Vol. 71, pp. 657–473.

STUKEL, D. M, RAO, J. N. K., (1999). On Small-Area Estimation Under Two-Fold Nested Error Regression Models, Journal of Statistical Planning and Inference, Vol. 78, pp. 131–147.

ŻĄDŁO, T., (2013). On Parametric Bootstrap and Alternatives of MSE, Proceedings of 31st International Conference Mathematical Methods in Economics 2013, College of Polytechnics Jihlava, pp. 1081–1086.

ŻĄDŁO, T., (2015). Statystyka małych obszarów w badaniach ekonomicznych. Podejście modelowe i mieszane, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

APPENDIX

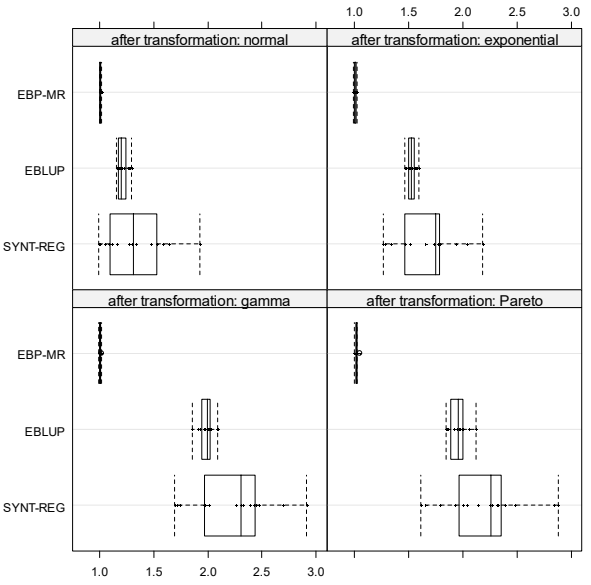


Figure 6. Values of QAPE0.75(.)/QAPE0.75(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for D=16 domains)

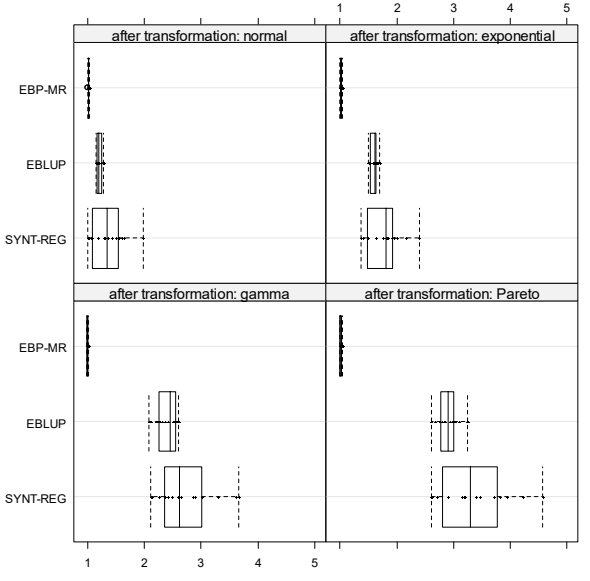


Figure 7. Values of QAPE0.90(.)/QAPE0.90(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for D=16 domains)

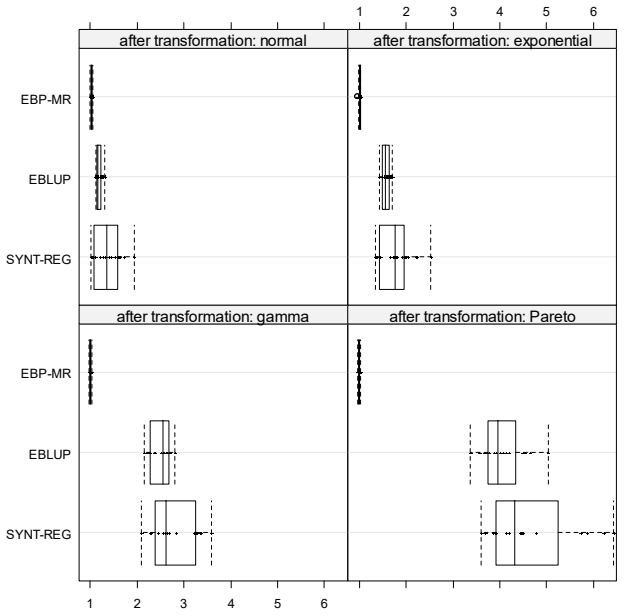


Figure 8. Values of QAPE0.95(.)/QAPE0.95(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for D=16 domains)

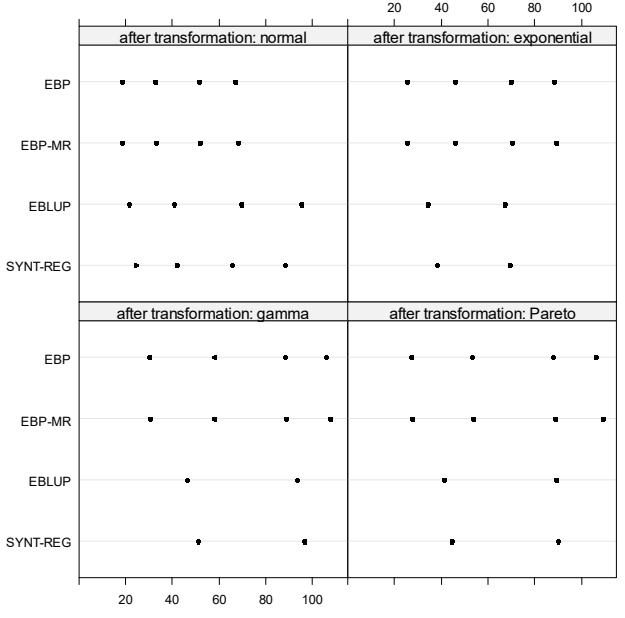


Figure 9. Values of rMQAPE(p) for p=(0.5, 0.75, 0.9, 0.95), different predictors and different distributions of random effects and random components – selected results

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 433–442, DOI 10. 21307

AN APPLICATION OF FUNCTIONAL MULTIVARIATE REGRESSION MODEL TO MULTICLASS CLASSIFICATION

Mirosław Krzyśko[1], **Łukasz Smaga**[2]

ABSTRACT

In this paper, the scale response functional multivariate regression model is considered. By using the basis functions representation of functional predictors and regression coefficients, this model is rewritten as a multivariate regression model. This representation of the functional multivariate regression model is used for multiclass classification for multivariate functional data. Computational experiments performed on real labelled data sets demonstrate the effectiveness of the proposed method for classification for functional data.

Key words: functional data analysis, multi-label classification problem, multivariate functional data, regression model.

1. Introduction

In recent decades, the analysis of data given as functions or curves has become a very popular branch of statistics. In the literature, such data are called functional data and have a broad perspective of applications, for example, in economics and medicine. The aim of functional data analysis (FDA) is to develop methods for analysing functional data. For instance, the books Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012) and Zhang (2013), and the references therein, offer a broad perspective of such methods.

Methods for analysing multivariate functional data (e.g. vectors of functions) are of particular interest. Some solutions of such problems as analysis of variance, canonical correlation analysis, classification, cluster analysis, linear regression and prediction, or principal component analysis are known in the literature. For example, we refer to the following papers by Górecki and Smaga (2017), Górecki et al. (2016), Górecki et al. (2015), Jacques and Preda (2014), Collazos et al. (2016) and Berrendero et al. (2011), respectively, and the references therein.

[1]Inter-Faculty Department of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland. E-mail: mkrzysko@amu.edu.pl.

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

This paper discusses the multiclass classification problem for multivariate functional data. The classifiers are constructed based on the scale response functional multivariate regression model and basis functions representation of functional predictors and coefficients. The presented results may be seen as extensions of those given in Górecki et al. (2015) from binary to multi-label case.

The rest of the paper is organized as follows. We first (Section 2) construct and rewrite (using the basis functions representation of predictors and coefficients) the scale response functional multivariate regression model. We consider two versions of this model, i.e. with and without intercepts. In Section 3, we apply these results to the multi-label classification problem for multivariate functional data. Section 4 contains the description of computational experiments for comparison of the proposed classifiers and a discussion of their results. We conclude in Section 5 with discussion of possible improvement of performance of the proposed method.

2. Functional multivariate regression model

In this Section, we consider the scalar response functional multivariate regression model, which can be seen as an extension of the one-dimensional model studied, for example, in Horváth and Kokoszka (2012).

Let $L_2(T)$ denote the Hilbert space of square integrable functions over $T=[a,b]$. Assume that we have measured p (scalar) responses $Y_1,\ldots,Y_p$ and the same set of k (functional) predictors $x_1(t),\ldots,x_k(t)$ belonging to $L_2(T)$ on each sample unit. Moreover, suppose that the responses follow the scalar regression models, i.e.

$$Y_j=\sum_{i=1}^{k}\int_T x_i(t)\xi_{ji}(t)dt+e_j,\qquad j=1,\ldots,p,$$

where $\xi_{ji}\in L_2(T)$ are the unknown functional coefficients and e_j are the random errors such that $\mathbf{e}^\top=[e_1,\ldots,e_p]$ has zero expectation and covariance matrix $\mathbf{\Sigma}$. When we have a sample of N independent observations $\mathbf{Y}_1,\ldots,\mathbf{Y}_N$ of the vector $[Y_1,\ldots,Y_p]^\top$, the scalar response functional multivariate regression model is formulated as follows:

$$\mathbf{Y}=\int_T \mathbf{X}(t)\mathbf{\Xi}(t)dt+\mathbf{E}, \tag{1}$$

where

$$\mathbf{Y}=\begin{bmatrix}\mathbf{Y}_1^\top\\ \vdots\\ \mathbf{Y}_N^\top\end{bmatrix},\ \mathbf{X}(t)=\begin{bmatrix}\mathbf{x}_1^\top(t)\\ \vdots\\ \mathbf{x}_N^\top(t)\end{bmatrix},\ \mathbf{\Xi}(t)=\left[\boldsymbol{\xi}_1(t),\ldots,\boldsymbol{\xi}_p(t)\right],\ \mathbf{E}=\begin{bmatrix}\mathbf{e}_1^\top\\ \vdots\\ \mathbf{e}_N^\top\end{bmatrix}, \tag{2}$$

and $\mathbf{x}_i^\top(t) = [x_{i1}(t), \ldots, x_{ik}(t)]$, $i = 1, \ldots, N$, $\boldsymbol{\xi}_j^\top(t) = \left[\xi_{j1}(t), \ldots, \xi_{jk}(t)\right]$, $j = 1, \ldots, p$.

To handle the model (1), we assume that the predictors and functional coefficients can be represented by a finite number of orthonormal basis functions $(\varphi_{mn}(t))_{n=0}^{\infty}$, $m = 1, \ldots, k$ in $L_2(T)$, i.e. for $i = 1, \ldots, N$ and $j = 1, \ldots, p$

$$x_{im}(t) = \sum_{n=0}^{B_m} c_{imn}\varphi_{mn}(t), \quad \xi_{jm}(t) = \sum_{n=0}^{B_m} d_{jmn}\varphi_{mn}(t), \tag{3}$$

where c_{imn} and d_{jmn} are the unknown coefficients. More precisely, c_{imn} are the random variables with finite variance (see Ramsay and Silverman, 2005). To estimate the coefficients c_{imn} (for each predictor separately), the least squares method can be used (see, for instance, Krzyśko and Waszak, 2013). The selection method of the values B_m may depend on the aim of the research. For example, when we want to obtain the best fit, the Bayesian information criterion should perhaps be used (see Shmueli, 2010). Different bases can be used for different predictors.

For easier presentation of our results, we represent the equations (3) in matrix notation. Let

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}_1^\top(t) & \mathbf{0}_{B_2+1}^\top & \cdots & \mathbf{0}_{B_k+1}^\top \\ \mathbf{0}_{B_1+1}^\top & \boldsymbol{\varphi}_2^\top(t) & \cdots & \mathbf{0}_{B_k+1}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{B_1+1}^\top & \mathbf{0}_{B_2+1}^\top & \cdots & \boldsymbol{\varphi}_k^\top(t) \end{bmatrix},$$

where $\boldsymbol{\varphi}_l^\top(t) = [\varphi_{l0}(t), \ldots, \varphi_{lB_l}(t)]$ for $l = 1, \ldots, k$ and $\mathbf{0}_n$ is an $n \times 1$ vector of zeros. Then, the equations given in (3) can be rewritten as follows:

$$\mathbf{x}_i(t) = \boldsymbol{\Phi}(t)\mathbf{c}_i, \qquad \boldsymbol{\xi}_j(t) = \boldsymbol{\Phi}(t)\mathbf{d}_j \tag{4}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, p$, where $\mathbf{c}_i^\top = [c_{i10}, \ldots, c_{i1B_1}, \ldots, c_{ik0}, \ldots, c_{ikB_k}]$ and $\mathbf{d}_j^\top = [d_{j10}, \ldots, d_{j1B_1}, \ldots, d_{jk0}, \ldots, d_{jkB_k}]$.

By (4), for $i = 1, \ldots, N$ and $j = 1, \ldots, p$, we have

$$\begin{aligned} \int_T \mathbf{x}_i^\top(t)\boldsymbol{\xi}_j(t)dt &= \int_T \mathbf{c}_i^\top \boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t)\mathbf{d}_j dt \\ &= \mathbf{c}_i^\top \int_T \boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t)dt\ \mathbf{d}_j \\ &= \mathbf{c}_i^\top \mathbf{d}_j, \end{aligned} \tag{5}$$

since the bases $(\varphi_{mn}(t))_{n=0}^{\infty}$, $m = 1, \ldots, k$, are orthonormal, i.e. $\int_T \boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t)dt$ is the identity matrix of size $\sum_{l=1}^{k} B_l + k$. From (2), it follows that

$$\int_T \mathbf{X}(t)\boldsymbol{\Xi}(t)dt = \begin{bmatrix} \int_T \mathbf{x}_1^\top(t)\boldsymbol{\xi}_1(t)dt & \dots & \int_T \mathbf{x}_1^\top(t)\boldsymbol{\xi}_p(t)dt \\ \vdots & \ddots & \vdots \\ \int_T \mathbf{x}_N^\top(t)\boldsymbol{\xi}_1(t)dt & \dots & \int_T \mathbf{x}_N^\top(t)\boldsymbol{\xi}_p(t)dt \end{bmatrix}.$$

Thus, by (5), we obtain

$$\begin{aligned} \int_T \mathbf{X}(t)\boldsymbol{\Xi}(t)dt &= \begin{bmatrix} \mathbf{c}_1^\top\mathbf{d}_1 & \dots & \mathbf{c}_1^\top\mathbf{d}_p \\ \vdots & \ddots & \vdots \\ \mathbf{c}_N^\top\mathbf{d}_1 & \dots & \mathbf{c}_N^\top\mathbf{d}_p \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{c}_1^\top \\ \vdots \\ \mathbf{c}_N^\top \end{bmatrix} [\mathbf{d}_1, \dots, \mathbf{d}_p] \\ &= \mathbf{CD}. \end{aligned}$$

Hence, the model (1) can be rewritten as

$$\mathbf{Y} = \mathbf{CD} + \mathbf{E}, \tag{6}$$

which is the multivariate regression model with the parameter matrix $\mathbf{D}$. Therefore, the problems connected with the functional multivariate regression model (1) (e.g. estimation of $\boldsymbol{\Xi}(t)$) can be replaced by the ones in the multivariate regression model (6). In the next Section, this relation is used for multiclass classification for multivariate functional data. Other results of such type and their usage are presented, for instance, in Kayano and Konishi (2009), Matsui and Konishi (2011), Matsui (2014), Górecki et al. (2015) and Collazos et al. (2016).

In the model (1), the intercepts were not considered. However, adding them to the model may improve the classification procedure based on it as we will see in Section 4. Thus, we extend the above results to the functional multivariate regression model with intercepts. Now, the scalar responses Y_j are modelled by the following regression models

$$Y_j = \xi_{j0} + \sum_{i=1}^{k} \int_T x_i(t)\xi_{ji}(t)dt + e_j, \qquad j = 1, \dots, p,$$

where ξ_{j0} are the (unknown) intercepts, and further the model (1) is replaced by

$$\mathbf{Y} = \boldsymbol{\Xi}_0 + \int_T \mathbf{X}(t)\boldsymbol{\Xi}(t)dt + \mathbf{E}, \tag{7}$$

where $\boldsymbol{\Xi}_0 = [\xi_{10}\mathbf{1}_N, \ldots, \xi_{p0}\mathbf{1}_N]$ and $\mathbf{1}_N$ is the $N \times 1$ vector of ones. Using the basis functions representation of predictors and functional coefficients given in (4), the model (7) can be rewritten as

$$\mathbf{Y} = [\mathbf{1}_N, \mathbf{C}] \begin{bmatrix} \boldsymbol{\xi}_0^\top \\ \mathbf{D} \end{bmatrix} + \mathbf{E} = \mathbf{C}_* \mathbf{D}_* + \mathbf{E}, \tag{8}$$

where $\boldsymbol{\xi}_0^\top = [\xi_{10}, \ldots, \xi_{p0}]$. Thus, the parameter matrix has one row more than in the earlier model.

3. Multiclass classification for functional data

In this Section, we investigate the multi-label classification problem for multivariate functional data by using the functional multivariate regression model considered in Section 2. More general information and results on classification problems based on regression models can be found in Krzyśko et al. (2008).

Assume that there are $K \geq 2$ populations and the objects are characterized by k features, which are given as functions in the space $L_2(T)$. Let

$$\mathbf{x}_i^\top(t) = [x_{i1}(t), \ldots, x_{ik}(t)], \quad i = 1, \ldots, N$$

be a sample from these populations. Each vector of functions $\mathbf{x}_i(t)$ is accompanied by the group label given by the $K \times 1$ vector

$$\mathbf{Y}_i^\top = [0, \ldots, 0, 1, 0, \ldots, 0]$$

with 1 in the lth place when the ith observation belongs to lth population.

In a classification problem, one wants to determine a procedure by which a given object can be assigned to one of K populations. For this purpose, the relation between vectors $\mathbf{x}_i(t)$ and $\mathbf{Y}_i$, $i = 1, \ldots, N$ is described by the scalar response functional multivariate regression model (1) or (7). Here we use the rewritten form (6) or (8) of it. The parameter matrices $\mathbf{D}$ and $\mathbf{D}_*$ in the models (6) and (8) can be estimated by the least squares method. The obtained estimators are of the form

$$\hat{\mathbf{D}} = (\mathbf{C}^\top \mathbf{C})^+ \mathbf{C}^\top \mathbf{Y}, \quad \hat{\mathbf{D}}_* = (\mathbf{C}_*^\top \mathbf{C}_*)^+ \mathbf{C}_*^\top \mathbf{Y},$$

where $\mathbf{M}^+$ is the Moore-Penrose pseudoinverse of the matrix $\mathbf{M}$. Then, the predicted

matrix is given by the formula

$$\hat{\mathbf{Y}} = \begin{cases} \mathbf{C}\hat{\mathbf{D}} = \mathbf{C}(\mathbf{C}^\top\mathbf{C})^+\mathbf{C}^\top\mathbf{Y}, & \text{for model (1),} \\ \mathbf{C}_*\hat{\mathbf{D}}_* = \mathbf{C}_*(\mathbf{C}_*^\top\mathbf{C}_*)^+\mathbf{C}_*^\top\mathbf{Y}, & \text{for model (7).} \end{cases}$$

To obtain the prediction for a new observation $\mathbf{x}_{\text{new}}(t)$, first its components have to be represented by a finite number of orthonormal basis functions, as it was described in Section 2, i.e.

$$\mathbf{x}_{\text{new}}(t) = \mathbf{\Phi}(t)\mathbf{c}_{\text{new}}.$$

Hence, the predicted vector $\hat{\mathbf{Y}}(\mathbf{x}_{\text{new}})$ for the new observation is of the form

$$\hat{\mathbf{Y}}(\mathbf{x}_{\text{new}})^\top = [\hat{Y}_1(\mathbf{x}_{\text{new}}), \ldots, \hat{Y}_K(\mathbf{x}_{\text{new}})] = \begin{cases} \mathbf{c}_{\text{new}}^\top\hat{\mathbf{D}}, & \text{for model (1),} \\ \left[1, \mathbf{c}_{\text{new}}^\top\right]\hat{\mathbf{D}}_*, & \text{for model (7).} \end{cases}$$

The lth component of the vector $\hat{\mathbf{Y}}(\mathbf{x}_{\text{new}})$ is the estimated value of the posterior probability of belonging to the lth population. Unfortunately, the components of this vector may not belong to the interval $[0,1]$. However, this may not matter if we get good predictions. Moreover, it can be shown that the sum of the components of $\hat{\mathbf{Y}}(\mathbf{x}_{\text{new}})$ is equal to one (see, for example, Krzyśko et al., 2008). Therefore, the classifier is given by the following formula

$$\hat{d}(\mathbf{x}_{\text{new}}) = \arg\max_{l=1,\ldots,K} \hat{Y}_l(\mathbf{x}_{\text{new}}). \tag{9}$$

In practice, the performance of this simple classifier may be satisfactory, as indicated by the real data examples of the next Section.

4. Computational experiments

In this Section, the accuracy of the proposed classifiers is examined using six real labelled data sets. All computational experiments were performed with R environment (R Development Core Team, 2015), and the codes are available from the authors.

The experiments were carried out on the following data sets: Arabic digits, Australian language, Character trajectories, Japanese vowels, ECG and Wafer. Table 1 shows the information on them. The first four data sets originate from Bache and Lichman (2013), and the remaining ones from Olszewski (2001). The discrete functional samples in each data set are of different lengths (see Table 1). For this reason, all discrete functional variables in a given data set were extended to the same length of the longest one by the method described and used, for example, in Górecki et al.

(2015) (see also Rodriguez et al., 2005).

Table 1. Summary of data sets

Data sets	k	N	K	Max length	Min length
Arabic digits	13	8800	10	93	4
Australian language	22	2565	95	136	45
Character trajectories	3	2858	20	205	109
ECG	2	200	2	152	39
Japanese vowels	12	640	9	29	7
Wafer	6	1194	2	198	104

To obtain the basis functions representation (3) of the observations, the orthonormal Fourier basis and the least squares method of estimating the coefficients were used (see Krzyśko and Waszak, 2013). As we noted in Section 2, the quantities B_m, $m = 1,\ldots,k$ in (3) can be chosen depending on the problem at hand. In our classification problem, we choose these quantities which minimize the classification error. In computational experiments, since we used the Fourier basis, we took into account $B_1 = \cdots = B_k = B$ and $B \in \{3,5,\ldots,I\}$, where I is the greatest odd number less than or equal to the number of design time points of a given data set, i.e. points on which functions are observed in practice.

The classifiers (9) based on models (1) and (7) were used for the classification process. The classification error rates are calculated by 10-fold cross-validation method. Figure 1 and Table 2 present the results. Observe that both classifiers give very good classification results for the data sets Arabic digits, Character trajectories, Japanese vowels and Wafer. However, the classification error rates are not so satisfactory for the data sets Australian language and ECG. This suggests that they are difficult to recognize.

Table 2. The smallest 10-fold cross-validation error rates (as percentages) and B's for which they are achieved by using classifiers (9) based on models (1) and (7)

	Model (1)		Model (7)	
Data sets	10CV error	B	10CV error	B
Arabic digits	4.35	15	4.01	27 or 33
Australian language	13.1	11	13.3	11
Character trajectories	1.23	175	1.19	127 or 171
ECG	11.5	31	11.5	31
Japanese vowels	1.88	5	1.41	5
Wafer	0.50	25 or 27 or 39	0.50	25 or 27 or 39

It seems that the classifier based on model (7) with intercepts performs at least as good as or even better than that based on model (1) without intercepts in most

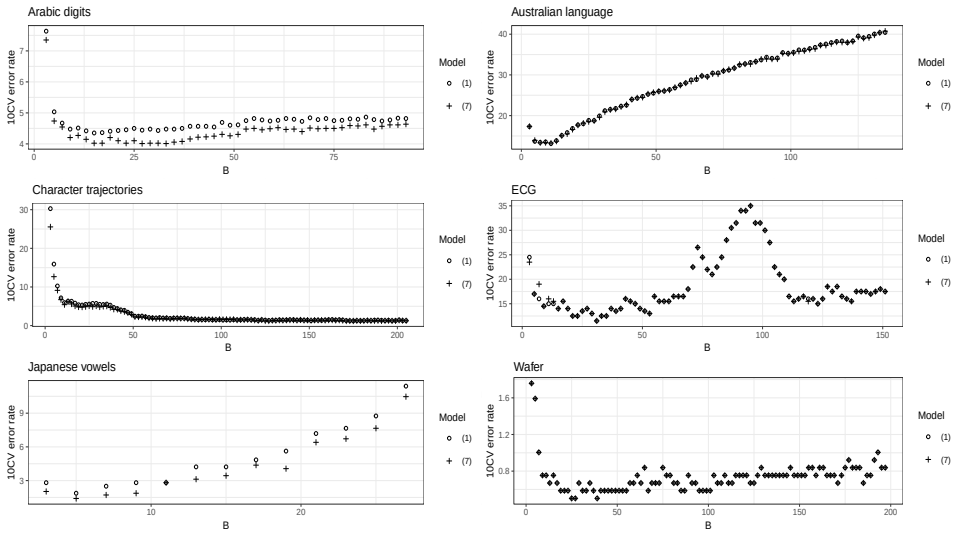


Figure 1: 10-fold cross-validation error rates (as percentages) for different values of B by using classifiers (9) based on models (1) and (7)

situations. However, for the data set Australian language, the smallest classification error rate of the method based on model (1) is slightly smaller than these of the second one (see Table 2). Therefore, for a given practical problem, both models may be examined, and we choose the one which minimizes the classification error.

From Figure 1 and Table 2, we see that the 10-fold cross-validation error rates behave differently for different values of B. In some cases, the best classification results are obtained for small values of B (e.g. for Japanese vowels) while in others for greater ones (e.g. for Character trajectories). Moreover, the values of B, for which the smallest classification error rates were achieved, may not be the same for classifiers based on models (1) and (7).

5. Conclusions

This paper discusses the construction of the scale response functional multivariate regression model and its application to multiclass classification problem for multivariate functional data. The computational experiments based on real labelled data sets suggest good performance of the proposed classification methods. From models with and without intercepts, the first one seems to be preferable.

For simplicity, in our real data examples, we used the orthonormal Fourier basis and equal lengths of basis functions representation of the observations, i.e. equal B_m's in (3). However, in practice, the performance of the considered classifiers may

be improved by using more appropriate orthonormal bases to different features and more varied values of B_m in (3).

REFERENCES

BACHE, K., LICHMAN, M., (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (http://archive.ics.uci.edu/ml).

BERRENDERO, J. R., JUSTEL, A., SVARC, M., (2011). Principal Components for Multivariate Functional Data. Computational Statistics & Data Analysis, 55, 2619–2634.

COLLAZOS, J. A. A., DIAS, R., ZAMBOM, A. Z., (2016). Consistent Variable Selection for Functional Regression Models. Journal of Multivariate Analysis, 146, 63–71.

FERRATY, F., VIEU, P., (2006). Nonparametric Functional Data Analysis: Theory and Practice, New York: Springer.

GÓRECKI, T., KRZYŚKO, M., WASZAK, Ł., WOŁYŃSKI, W., (2016). Selected Statistical Methods of Data Analysis for Multivariate Functional Data. Statistical Papers (Accepted) doi:10.1007/s00362-016-0757-8.

GÓRECKI, T., KRZYŚKO, M., WOŁYŃSKI, W., (2015). Classification Problem Based on Regression Models for Multidimensional Functional Data. Statistics in Transition new series, 16, 97–110.

GÓRECKI, T., SMAGA, Ł., (2017). Multivariate Analysis of Variance for Functional Data. Journal of Applied Statistics, 44, 2172–2189.

HORVÁTH, L., KOKOSZKA, P., (2012). Inference for Functional Data with Applications, New York: Springer.

JACQUES, J., PREDA, C., (2014). Model-Based Clustering for Multivariate Functional Data. Computational Statistics & Data Analysis, 71, 92–106.

KAYANO, M., KONISHI, S., (2009). Functional Principal Component Analysis via Regularized Gaussian Basis Expansions and its Application to Unbalanced Data. Journal of Statistical Planning and Inference, 139, 2388–2398.

KRZYŚKO, M., WASZAK, Ł., (2013). Canonical Correlation Analysis for Functional Data. Biometrical Letters, 50, 95–105.

KRZYŚKO, M., WOŁYŃSKI, W., GÓRECKI, T., SKORZYBUT, M., (2008). Learning Systems, Warsaw: WNT (in Polish).

MATSUI, H., (2014). Variable and Boundary Selection for Functional Data via Multiclass Logistic Regression Modeling. Computational Statistics & Data Analysis, 78, 176–185.

MATSUI, H., KONISHI, S., (2011). Variable Selection for Functional Regression Models via the L_1 Regularization. Computational Statistics & Data Analysis, 55, 3304–3310.

OLSZEWSKI, R. T., (2001). Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA (http://www.cs.cmu.edu/bobski).

RAMSAY, J. O., SILVERMAN, B. W., (2005). Functional Data Analysis, Second Edition, New York: Springer.

R DEVELOPMENT CORE TEAM, (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (https://www.R-project.org/).

RODRIGUEZ, J. J., ALONSO, C. J., MAESTRO, J. A., (2005). Support Vector Machines of Interval Based Features for Time Series Classification. Knowledge-Based Systems, 18, 171–178.

SHMUELI, G., (2010). To Explain or to Predict? Statistical Science, 25, 289–310.

ZHANG, J.-T., (2013). Analysis of Variance for Functional Data, London: Chapman & Hall.

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 443–458, DOI 10. 21307

STACKED REGRESSION WITH A GENERALIZATION OF THE MOORE-PENROSE PSEUDOINVERSE

Tomasz Górecki[1], **Maciej Łuczak**[2]

ABSTRACT

In practice, it often happens that there are a number of classification methods. We are not able to clearly determine which method is optimal. We propose a combined method that allows us to consolidate information from multiple sources in a better classifier. Stacked regression (SR) is a method for forming linear combinations of different classifiers to give improved classification accuracy. The Moore-Penrose (MP) pseudoinverse is a general way to find the solution to a system of linear equations.

This paper presents the use of a generalization of the MP pseudoinverse of a matrix in SR. However, for data sets with a greater number of features our exact method is computationally too slow to achieve good results: we propose a genetic approach to solve the problem. Experimental results on various real data sets demonstrate that the improvements are efficient and that this approach outperforms the classical SR method, providing a significant reduction in the mean classification error rate.

Key words: stacked regression, genetic algorithm, Moore-Penrose pseudoinverse.

1. Introduction

Suppose that a training sample has been collected by sampling from a population P consisting of K subpopulations or classes $G_1, \ldots, G_K$. The ith observation is a pair denoted by $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is a d-dimensional feature vector and y_i is the label for recording class membership. The corresponding pair for an unclassified observation is denoted by $(\mathbf{x}, y)$. In this case $\mathbf{x}$ is observed, but the class label y is unobserved. The goal of classification is to construct a classification rule for predicting the membership of an unclassified feature vector $\mathbf{x} \in P$. An automated classifier can be viewed as a method of estimating the posterior probability of membership of G_j. The classification rule assigns $\mathbf{x}$ to the group with the largest posterior probability estimate. We denote the posterior probability of membership of G_k by

$$p_k(\mathbf{x}) = P(y = k|\mathbf{x}). \tag{1}$$

[1]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Umultowska 87, 61-614 Poznań, Poland. E-mail: tomasz.gorecki@amu.edu.pl.

[2]Faculty of Civil Engineering, Environmental and Geodetic Sciences, Koszalin University of Technology, Śniadeckich 2, 75-453 Koszalin, Poland. E-mail: mluczak@wilsig.tu.koszalin.pl.

In practice, it is not clear how one should choose a classifier. An even more practical difficulty in choosing a classifier is that different classifiers have different merits and, as a result, in a given situation one classifier may perform better than another. Consider the following typical situation (Mojirsheibani (2002)). Suppose that there are 3 classes, two of which are approximately multivariate normal distributions, while the third class is non-normal. Then linear or quadratic classification function might work best for separating the first two classes (the normal distributions), while the nearest neighbor rule is perhaps more appropriate in the non-normal case. This example suggests that perhaps one should consider methods that somehow combine the best features of different individual classifiers. Some possible benefits of such combined methods are as follows:

1. Lowering the risk of choosing the wrong classifier.

2. Obtaining more stable prediction performance, since in combining different methods certain biases inherited from particular models could be offset.

3. Producing a better prediction of the classification of new observations, since the combined method gives decision-makers additional information from different sources.

The purpose of ensemble learning is to construct a learning rule which combines a number of base methods, so that the final classifier gives better performance than any individual classifier (Rokach (2010)). Three groups of combining methods could be distinguished as follows (Duin and Tax (2000)):

- Parallel combining of classifiers computed for different feature sets. Parallel classifiers are often of the same type.

- Stacked combining of different classifiers computed for the same feature space. Stacked classifiers may be of a different nature, e.g. the combination of a neural network, a nearest neighbor classifier and a parametric decision rule.

- Combining weak classifiers. In this case, large sets of simple classifiers are trained on modified versions of the original data set.

For all cases, the question arises how the classifiers should be combined. The most intuitive approach is a simple majority vote (Kuncheva (2004)), whereby every classifier computes a class label and the label that receives the most votes is the output of the ensemble. In addition, one may also train a classifier using, e.g. the BKS method (Huang and Suen (1995)), Wernecke's method (Wernecke (1992)) or the fuzzy integral (Cho and Kim (1995)). Currently, the most interesting ensemble

methods are bagging (Breiman (1996b)) and boosting (Schapire (1990)), random forests (Breiman (2001)), and finally SR, introduced by Wolpert (1992). In SR, the posterior probability estimates are combined by weighted sums, where the weights are obtained by classical least squares regression. Stacking is still used in practice (Sehgal et al. (2005), Doumpos and Zopounidis (2007), Marqués et al. (2012)). Although SR is applied to real-world problems less frequently than other ensemble methods, such as bagging or boosting, the exponential growth of data as well as the diversity of these data continue to make SR an interesting alternative (Sesmero et al. (2015)). There are also some new papers which propose extensions to SR (Džeroski and Ženko (2004), Rooney et al. (2004a), Rooney et al. (2004b), Xu et al. (2007), Ozay and Vural (2008), Ni et al. (2009)), Ledezma et al. (2010), Shunmugapriya and Kanmani (2013). An informative overview of SR methods can be found in Sesmero et al. (2015).

Sigletos et al. (2005) pointed out that stacking using probabilities performs comparably or significantly better than voting. This result has inspired us to consider some extension of SR. The classical stacked regression method uses the MP inverse of a matrix to solve a set of normal equations, whereas we try to find a specific generalization of the MP inverse. We construct a parametric family of generalized MP inverses and use it in the SR model. Then we choose models with the lowest cross-validation (leave-one-out) error rate and combine them by a mean rule (Kuncheva (2004)).

However, for most datasets there are too many models to compute the cross-validation (CV) error for all of them. The problem is too complex to find an exact solution (or if done, it takes too long to calculate the solution exactly). The most feasible approach, then, is to use a meta-heuristic method (Michalewicz, Fogel (2004)). A genetic algorithm (GA) is meta-heuristic, which means it estimates a solution. Therefore, we propose GA to solve our problem. GA has a number of advantages. It can quickly scan a vast solution set. Bad proposals do not negatively affect the end solution, as they are simply discarded. It can solve every optimization problem which can be described with the chromosome encoding. It solves problems with multiple solutions. Since the genetic algorithm execution technique is not dependent on the error surface, we can solve multi-dimensional, non-differential, non-continuous, and even non-parametric problems. It is a method which is very easy to understand and it demands practically no mathematical knowledge.

In this paper, we first present the main ideas of SR (Section 2). In the same section we describe generalized inverses of matrices. At the end of this section we explain our concept for extended SR and we precisely describe the genetic approach to our extension. In the paper the performances of the methods are compared and

the bootstrap error of classification is considered. A total of 15 real data sets are used. The methods and data sets used are described in Section 3. Section 4 contains the results of our experiments on the described real data sets. The results of the research are explained, the differences between the classifiers being shown accurately. The same section contains a statistical comparison of the described methods. Final conclusions are given in Section 5.

2. Methods

2.1. Stacked regression

Wolpert (1992) presented an interesting idea for the combining of classifiers, known as stacked generalization. He was not searching for the best classifier in the set of all c classifiers, but for a linear combination of them. Since each single one has some advantages, combining them is reasonable. Wolpert's proposal was translated into the language of statistics by Breiman (1996a). He called it SR. Then, Leblanc and Tibshirani (1996) took advantage of it to construct a combined classifier in discriminant analysis. Stacking was shown by them theoretically to be a bias-reducing technique. A combined classifier is a linear combination of estimated posterior probabilities. An estimate of $p_k(\mathbf{x})$ obtained by the jth classifier is denoted by

$$\hat{p}_k^j(\mathbf{x}); \quad k = 1,2,\ldots,K; \quad j = 1,2,\ldots,c. \tag{2}$$

We have c classifiers and K classes, so we have $K \cdot c$ estimates, which are arranged in the vector:

$$\hat{\mathbf{p}}(\mathbf{x}) = (\hat{p}_1^1(\mathbf{x}),\ldots,\hat{p}_K^1(\mathbf{x}),\ldots,\hat{p}_1^c(\mathbf{x}),\ldots,\hat{p}_K^c(\mathbf{x}))'. \tag{3}$$

These estimates are arranged in the stack as rows of the matrix $\mathbf{P}$. Let $\mathbf{u}_k$ be a vector having a 1 in the ith position if the observation belongs to class k and 0 otherwise, so

$$u_{i,k} = \begin{cases} 1, & \text{if } y_i = k, \\ 0, & \text{if } y_i \neq k. \end{cases} \tag{4}$$

The SR model has the form:

$$\mathbf{u}_k = \mathbf{P}\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k, \tag{5}$$

where $\boldsymbol{\beta}_k$ is a $K \cdot c \times 1$ vector of unknown SR coefficients and $\boldsymbol{\varepsilon}_k$ a vector of errors with zero mean. A least-squares estimate of $\hat{\boldsymbol{\beta}}_k$ can be obtained by solving the

following equation:

$$\mathbf{P}'\mathbf{P}\boldsymbol{\beta}_k = \mathbf{P}'\mathbf{u}_k \tag{6}$$

with respect to $\boldsymbol{\beta}_k$.

The estimates of posterior probability obtained from the classifiers sum to one, so

$$\sum_{k=1}^{K} \hat{p}_k^j = 1; \quad j = 1,2,\ldots,c. \tag{7}$$

Hence, the columns of matrix $\mathbf{P}$ are subject to c linear constraints, $\mathbf{P}$ is not full column rank and $\mathbf{P}'\mathbf{P}$ is a singular matrix. We can use the MP generalized inverse of the matrix $\mathbf{P}'\mathbf{P}$ (Breiman (1996)), denoted by $(\mathbf{P}'\mathbf{P})^+$, and

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{P}'\mathbf{P})^+\mathbf{P}'\mathbf{u}_k. \tag{8}$$

Given the estimates $\hat{\boldsymbol{\beta}}_1,\ldots,\hat{\boldsymbol{\beta}}_K$, we classify $\mathbf{x}$ using the dot product:

$$\hat{u}_{0,k} = \hat{\mathbf{p}}'(\mathbf{x})\hat{\boldsymbol{\beta}}_k. \tag{9}$$

We select the class with the largest values of $\hat{u}_{0,k}$. These scalar products are called discriminant indices.

2.2. Algorithm

In SR, the MP generalized inverse $\mathbf{A}^+$ is used to compute the coefficients $\hat{\boldsymbol{\beta}}_k$ (see Equation (8)). The main idea of this paper is to use another generalized inverse. The MP pseudoinverse is a general way to find the solution to a system of linear equations (eg. Ben-Israel and Greville (2003), Kyrchei (2015)).

We consider a general (real) matrix $\boldsymbol{A}$ of order $m \times n$ and rank which may be less than $\min(m,n)$. If $\boldsymbol{M}$,$\boldsymbol{N}$ are positive definite matrices, and there exist factorizations $\hat{\boldsymbol{N}}'\hat{\boldsymbol{N}} = \boldsymbol{N}$, $\hat{\boldsymbol{M}}'\hat{\boldsymbol{M}} = \boldsymbol{M}$, then

$$\boldsymbol{A}^+_{\boldsymbol{MN}} = \hat{\boldsymbol{N}}^{-1}(\hat{\boldsymbol{M}}\boldsymbol{A}\hat{\boldsymbol{N}}^{-1})^+\hat{\boldsymbol{M}}, \tag{10}$$

satisfies the condition

$$\begin{aligned} &\|\boldsymbol{A}^+_{\boldsymbol{MN}}\boldsymbol{y}\|_n \le \|\boldsymbol{x}\|_n \\ &\quad \forall \boldsymbol{x} \in \{\boldsymbol{x}\colon \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|_m \le \|\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}\|_m \forall \boldsymbol{z} \in \mathbb{R}^n\}, \end{aligned} \tag{11}$$

where $\|\boldsymbol{x}\|_n = \sqrt{\boldsymbol{x}'\boldsymbol{N}\boldsymbol{x}}$ and $\|\boldsymbol{y}\|_m = \sqrt{\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}}$ are norms in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. $\boldsymbol{A}^+_{\boldsymbol{MN}}$ is referred to as the minimum $\boldsymbol{N}$-norm $\boldsymbol{M}$-least-squares g-inverse of $\boldsymbol{A}$. When $\boldsymbol{M}$ and $\boldsymbol{N}$ are identity matrices, we use the notation $\boldsymbol{A}^+$ and call it the MP inverse (pseudoinverse). The matrix $\boldsymbol{A}^+_{\boldsymbol{MN}}$ is also called the weighted Moore-Penrose inverse of $\boldsymbol{A}$. The weighted MP inverse of a matrix has many important applications eg. in statistics, prediction theory and curve fitting. For a wider survey and more details we refer readers to Rao and Mitra (1971).

If $\boldsymbol{M}$ is positive semi-definite, then $\|\boldsymbol{y}\|_m$ is a seminorm and the right side of Equation (10) does not need to be a g-inverse. We denote it by $\boldsymbol{A}^*_{\boldsymbol{MN}}$ and $\boldsymbol{A}^*_{\boldsymbol{M}}$ if $\boldsymbol{N} = \boldsymbol{I}$.

In our method we use $\mathbf{A}^*_{\mathbf{M}}$ with a special form of matrix $\mathbf{M}$ instead of $\mathbf{A}^+$. Precisely, we use Equation (10) with the assumptions

$$\hat{\mathbf{N}} = \mathbf{N} = \mathbf{I}, \qquad \hat{\mathbf{M}} = \mathbf{M} = \begin{bmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_m \end{bmatrix} \tag{12}$$

where $a_i = 0$ or 1 for $i = 1, \dots, m$ ($m = K \cdot c$). This leads to the seminorm

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{M}\mathbf{x}} = \sqrt{x_{i_1}^2 + x_{i_2}^2 + \dots + x_{i_k}^2}, \qquad 1 \le k \le m \tag{13}$$

for $\mathbf{x} = (x_1, x_2, \dots, x_m) \in \mathbf{R}^m$. Then Equation (10) has the form

$$\mathbf{A}^*_{\mathbf{M}} = (\mathbf{M}\mathbf{A})^+\mathbf{M}. \tag{14}$$

Thus, we can calculate SR coefficients $\hat{\boldsymbol{\beta}}_k$ by the formula

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{P}'\mathbf{P})^*_{\mathbf{M}}\mathbf{P}'\mathbf{u}_k = (\mathbf{M}\mathbf{P}'\mathbf{P})^+\mathbf{M}\mathbf{P}'\mathbf{u}_k. \tag{15}$$

In the algorithm the matrix $\mathbf{N}$ corresponds to the norm $\|\cdot\|_n$ in Equation (11). In the case of SR the norm operates on the space of probabilities, so it seems that the simplest choice is to take the Euclidean norm, i.e. $\mathbf{N} = \mathbf{I}$.

We only take ones and zeros in the diagonal of the matrix $\boldsymbol{M}$ because it has been proven (Górecki, Łuczak (2013)) that the value of $\boldsymbol{A}^*_{\boldsymbol{M}}$ depends only on whether the coefficients a_i are zeros or not. Each zero in the diagonal trims a part (but not all) of the information about one pair consisting of a class and a classifier.

The number of models (diagonals) which have to be tested by the CV process at the learning phase is equal to $2^{K \cdot c}$, where c and K depends neither on the number of elements of the learning data set nor on the number of features of the data.

In our algorithm we choose the best combinations of ones and zeros in the diagonal of matrix **M** using the genetic algorithm, and form the SR model with the lowest CV error rate. If there is more than one best model the mean classifier is performed for them; the classification index of our method is the mean of indices for the best models (in the sense of CV) that joins all the information from them. We will call this method generalized stacked regression (GSR).

2.3. Genetic algorithm

The sequence of steps in a basic GA is shown in Fig. 1. The population consists of individuals (genotypes) which are diagonal of matrix $\boldsymbol{M}$. Each individual is a binary vector (genes) that corresponds to numbers (ones or zeros) in the diagonal of $\boldsymbol{M}$. All populations in the algorithm have a constant number n of individuals.

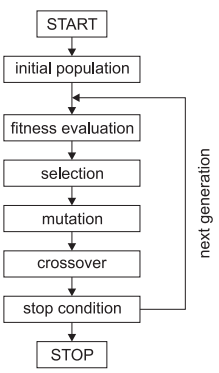


Figure 1: Genetic algorithm

Initial population: This is generated randomly. We construct n individuals where each position in the vector (diagonal) may be 0 or 1 with probability of 0.5. *Fitness evaluation*: The fitness function value is computed by the leave-one-out CV method. The CV error rate is the fitness value of any individual. The smaller the value, the better fitness an individual has. *Selection*: We use tournament selection. Two individuals are chosen from the population at random. The one with higher fitness is selected for mutation and crossover. This is repeated n times to make a new population. *Mutation*: We use standard one-point mutation. For each individual each position in the vector has the same probability of mutation p_m. The mutation is negation of the number (0 or 1) in the position (Fig. 2). It is repeated an appropriate number of times to make a new population of size n. *Crossover*: We use a standard one-point crossover operation. Each individual can be chosen to crossover with constant probability p_c. For every pair of chosen individuals, the point of crossing is fixed at random. Then the positions to the right of the point

are exchanged with one another (Fig. 2). The operation is repeated an appropriate number of times to make a new population of size n.

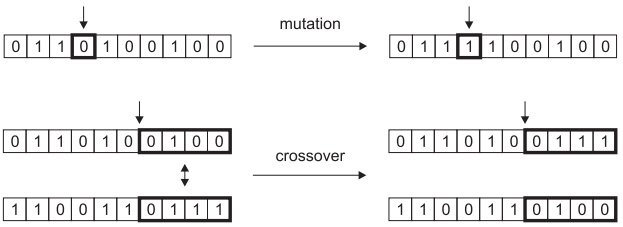


Figure 2: Reproduction

Stop condition: We do not use a fixed number of generations in GA. For so many different data sets, the algorithm needs different numbers of steps to reach a satisfactory result. The process is repeated until a stop condition has been reached. The stop condition depends on the behaviour of the mean fitness value in the populations over k steps of the algorithm. If during k steps the mean does not become smaller than the smallest value of the mean up to the current generation, the algorithm is terminated. We shall call the number k the stop condition number.

3. Computational experiments

Data sets

We performed experiments on 15 real data sets. The description of the data sets used is presented in Table 1.

The data set *beetles* comes from Seber (1984), *chemistry* and *irradiation* come from Morrison (1976), and *football* is from Gleim (1984). The other data sets originate from the UCI Machine Learning Repository (Frank and Asuncion (2010)).

Experimental setup

The classification errors were estimated by the leave-one-out and bootstrap methods. Leave-one-out was used to find the best diagonals (those with the smallest error rates) of matrix $\mathbf{M}$. The method was used to compute the value of the fitness function in the GA. The number of individuals per population was fixed at a constant value of $n = 20$. We chose probabilities of mutation $p_m = 0.01$ and crossover $p_c = 0.8$. As a selection method we use tournament selection. Different stop condition numbers were tried, $k = 0, \ldots, 10$. For the final result of our method we assumed the best case $k = 10$. For each data set we repeated the algorithm 10 times.

Table 1: The description of the data sets used.

Name	Number of features	Number of classes	Number of instances
beetles	2	3	64
breast tissue	9	6	106
chemistry	3	4	45
flags	28	6	194
football	6	3	90
glass	9	6	214
heart_c	13	5	297
heart_h	10	5	261
heart_s	10	5	105
iris	4	3	150
irradiation	3	4	45
libras	90	15	360
sonar	60	2	208
wine	13	3	178
zoo	16	7	100

In the next step, the mean classifier was performed for models with each of these diagonals. We calculated the bootstrap classification error rate (1000 repetitions). We finally fixed the mean of these bootstrap error rates as the error rate of our method.

The success of stacked generalization depends on the methods that are combined. Obviously, if all the methods provide the same class assignments, then a combined model will not provide any improvement in classification accuracy. The classification performance of the methods is of rather limited interest in this context, i.e. one is not interested in combining highly accurate methods, but in combining methods that are able to consider different aspects of the problem and the data used. Of course, it is rather difficult to find which methods meet this requirement. However, it is expected that consideration of different types of methods (e.g. methods which are not simple variations of one another) should be beneficial in stacking (Wolpert (1992)). We performed computations for three basic classifiers:

1. Nearest neighbors classifier with 5 neighbors (5NN). Objects are assigned based on a majority vote among the classes of the 5 nearest training points. The 5NN variant of the nearest neighbor classifier was chosen on the one hand to avoid an excess of zero posterior probabilities, and on the other hand because too large a number of neighbors leads to an excessive number of ties, whose resolution can be problematic (Górecki, (2005)). Too many neighbors may also be problematic for small data sets and for data sets with small classes.

2. Naive Bayes classifier (NB). We assume that the value of a particular feature is independent of the value of any other feature, given the class variable. This reduces the problem to d one-dimensional density estimation problems, within each of the K groups. We adopted a typical assumption that the continuous values associated with each class are distributed according to a Gaussian distribution. NB classifier works quite well in many complex real-world situations. In addition, Zhang (2004) investigated the optimality of NB under the Gaussian distribution, and presented the explicit sufficient condition under which NB is optimal, even though the independence assumption is violated.

3. Binary decision tree classifier (TREE). The algorithm computes a binary decision tree on a multi-class data set. Thresholds are set such that the Gini impurity is minimized in each step. Early pruning is used in order to avoid overtraining (Breiman et al. (2005)).

We focus on methods with a fast implementation (at the same time popular and relatively efficient), because GA itself is very time-consuming. The methods should be also significantly diversified in order for the ensemble method to yield better results (Kuncheva and Whitaker (2003)). Noteworthy is also Table 2 in Sesmero et al. (2015), where one can find information about base classifiers used in SR. The methods we selected are commonly used and meet the criteria of fast implementation and efficiency. More details about the methods we use can be found in Webb (2002).

In the computational process we used the program PRTools 4.2.1 (`http://www.prtools.org`). This is a Matlab (version R2011a) based toolbox for pattern recognition (van der Heijden et al. (2004)). In each procedure we used the default parameters.

Results

Graphs of example runs of our algorithm are shown in Fig. 3. We can observe rather standard behaviour of GA. We use tournament selection, which is not an elitist selection, so we can observe that the minimum of the fitness function does not decrease monotonically. The mean tends to a minimum and the algorithm is terminated if the stop condition is reached, i.e. if the mean does not decrease for a number of generations.

The results of the research are presented below in tabular form. Bootstrap error rates are presented in Table 2. From left to right the columns show the errors made by individual methods, SR, and our GSR. 5NN performed clearly the best on 2 of the data sets, NB on 3 and GSR on 10.

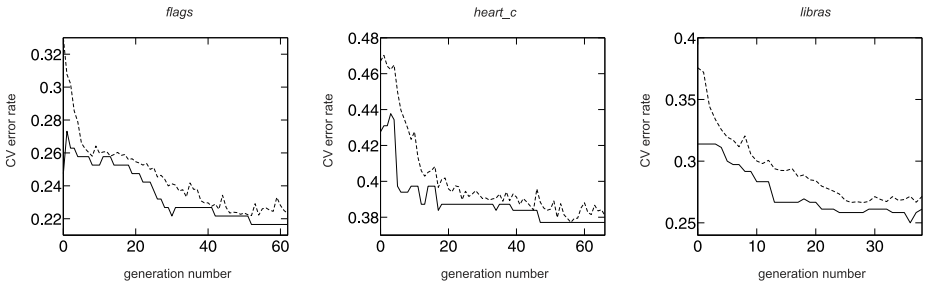


Figure 3: Runs of GA for example data sets. Fitness function value (mean $(\cdots)$ and minimum (—) of CV error rate) depending on the generation number. From left: *flags*, *heart_c*, *libras* data set.

Table 2: Bootstrap error rates (in %). Clearly the best results are marked with the symbol •.

Data set	5NN	NB	TREE	SR	GSR
beetles	6.08	6.84	5.34	4.95	• 4.34
breast tissue	48.72	38.49	43.29	41.17	• 37.47
chemistry	65.38	70.15	66.93	66.92	• 64.67
flags	66.39	35.77	43.50	40.64	• 33.47
football	40.49	• 32.60	40.51	38.99	32.65
glass	34.40	39.67	37.34	34.98	• 33.65
heart_c	57.36	41.52	52.13	51.41	• 41.48
heart_h	50.19	37.09	53.98	54.00	• 35.39
heart_s	64.86	63.76	66.53	66.55	• 63.20
iris	• 4.41	5.98	9.77	7.38	5.88
irradiation	70.62	72.06	71.76	71.79	• 70.21
libras	• 27.92	40.97	56.68	35.09	35.81
sonar	• 23.34	25.71	32.29	32.29	23.39
wine	30.67	• 3.35	11.72	5.14	3.69
zoo	10.34	8.36	10.90	9.82	• 5.54

In Table 3 we present relative differences of bootstrap error rates between SR and other methods (a positive value means that SR is better in that case). We may use the mean ratio of error rates across data sets as a measure of relative performance (Bauer and Kohavi (1998)).

Table 3: Average relative bootstrap error rates (in%) on all data sets.

	$\frac{5NN-SR}{SR}$	$\frac{NB-SR}{SR}$	$\frac{TREE-SR}{SR}$	$\frac{GSR-SR}{SR}$
MEAN	34.51	-7.02	17.65	-16.08

A direct comparison of SR with our revised version strongly favors the revised method. A graphical comparison of GSR and SR is presented in Fig. 4. We see that the new method, GSR, is clearly superior to SR on most of the examined data sets (with a 16.08% average relative error reduction for all data sets). The error rate of our method is slightly greater than for standard stacked regression in only one case (*libras*). One of the models is the standard SR (for $\mathbf{M} = \mathbf{I}$), so if it is the best model then it should be chosen. It sometimes fails because of the procedure for finding parameters. If we tried another, more sophisticated, method of finding the best model instead of CV, we would have better results.

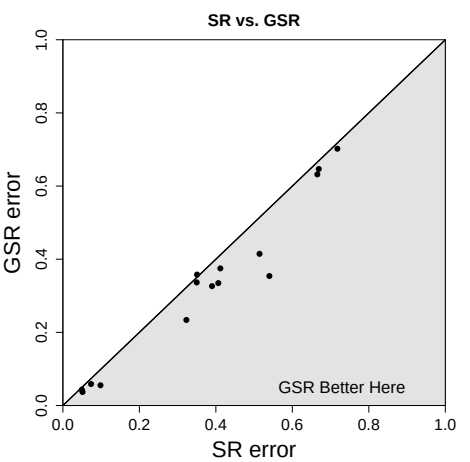


Figure 4: Comparison of test errors.

To distinguish between the methods, we performed a statistical comparison. We tested the hypothesis that there are no differences between the classifiers. We

used Iman and Davenport's (1980) rank test, which is a less conservative variant of Friedman's ANOVA test. We compare the mean ranks of classifiers. The p-value from this test is equal to 5.29E-6. We can therefore proceed with the post hoc tests to detect which classifiers are significantly different from each other. Garcia and Herrera (2008) showed that the dynamic procedure of Bergmann and Hommel (1988) is the most powerful post hoc test. The results of multiple comparisons are given in Table 4 and Table 5. We finally obtained, at the significance level $\alpha = 0.05$, two homogeneous groups of classifiers: GSR and the rest of classifiers. Hence, GSR is significantly better than the other examined classifiers.

Table 4: p-values in the Bergmann–Hommel post hoc test.

i	Hypothesis	p-value
1	TREE vs. GSR	1.65×10^{-5}
2	SR vs. GSR	0.004
3	5NN vs. GSR	0.011
4	NB vs. GSR	0.032
5	NB vs. TREE	0.196
6	5NN vs. TREE	0.220
7	TREE vs. SR	0.332
8	NB vs. SR	1.000
9	5NN vs. SR	1.000
10	5NN vs. NB	1.000

Table 5: Results of the Bergmann–Hommel post hoc test.

Procedure	Ranks mean		
GSR	4.60	a	
NB	3.07		b
5NN	2.87		b
SR	2.63		b
TREE	1.83		b

4. Conclusions

Our research has shown that the use of a generalization of the MP pseudoinverse of a matrix in the SR model of object classification gives good results. In the gen-

eral case our method seems to outperform SR and often even the best individual classifier. Owing to the parametric approach and the genetic optimization method, the proposed method enables one to choose an appropriate model for any data set and any individual classifiers. On the other hand, our method seems to prevent overfitting. Due to the high nonlinearity, the method does not easily lead to a rigorous theoretical analysis. However, the experiments that we have conducted provide evidence of the power and usefulness of our method.

REFERENCES

BAUER, E., KOHAVI, R., (1999). An experimental comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning, 36, 105–139.

BEN-ISRAEL, A., GREVILLE, T.N.E. (2003). Generalized inverses. Theory and applications. Springer.

BERGMANN, G., HOMMEL, G., (1988). Improvements of general multiple test procedures for redundant systems of hypotheses. In Multiple Hypotheses Testing. P. Bauer, G. Hommel and E. Sonnemann (eds.) Springer, 110–115.

BREIMAN, L., (1996a). Stacked regression. Machine Learning, 24, 49–64.

BREIMAN, L., (1996b). Bagging predictors. Machine Learning, 24, 123–140.

BREIMAN, L., (2001). Random forests. Machine Learning, 45, 5–32.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., STONE, C.J., (1984). Classification and regression trees, Wadsworth, California.

CHO, S.B., KIM, J.H., (1995). Multiple network fusion using fuzzy logic. IEEE Transactions on Systems, Man, and Cybernetics, 6, 497–501.

DOUMPOS, M., ZOPOUNIDIS, C., (2007). Model combination for credit risk assessment: A stacked generalization approach. Annals of Operations Research, 151, 289–306.

DUIN, R., TAX, D., (2000). Experiments with classifier combining rules. Lecture Notes in Computer Science, 1857, 16–29.

DŽEROSKI, S., ŽENKO, B., (2004). Is combining classifiers with stacking better than selecting the best one? Machine Learning, 54, 255–273.

FRANK, A., ASUNCION, A., (2010). UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml` Irvine, CA: University of California, School of Information and Computer Science.

GARCIA, S., HERRERA, F., (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research, 9, 2677–2694.

GLEIM, G., (1984). The profiling of professional football players. Clinical Sport Medicine, 3(1), 185–97.

GÓRECKI, T., (2005). Effect of choice of dissimilarity measure on classification efficiency with nearest neighbor method. Discussiones Mathematicae Probability and Statistics, 25(2), 217–239.

GÓRECKI, T., ŁUCZAK, M., (2013). Linear discriminant analysis with a generalization of Moore-Penrose pseudoinverse. International Journal of Applied Mathematics and Computer Science, 26(2), 463–471.

HUANG, Y.S., SUEN, C.Y., (1995). A method of combining multiple experts for the recognition of unconstrained handwritten numerals. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17, 90–93.

IMAN, R., DAVENPORT, J., (1980). Approximations of the critical region of the Friedman statistic. Communications in Statistics - Theory and Methods, 9(6), 571–595.

KUNCHEVA, L.I., (2004). Combining Pattern Classifiers: Methods and Algorithms. Wiley.

KUNCHEVA, L., WHITAKER, C., (2003). Measures of diversity in classifier ensembles. Machine Learning, 51, 181–207.

KYRCHEI, I., (2015). Cramer's rule for generalized inverse solutions. In Advances in Linear Algebra Research I. Kyrchei (ed.) Nova Science Publishers, 79–132.

LEBLANC, M., TIBSHIRANI, R., (1996). Combining estimates in regression and classification. Journal of the American Statistical Association, 91, 1641–1650.

LEDEZMA, A., ALER, R., SANCHIS, A., BORRAJO, D., (2010). GA-stacking: evolutionary stacked generalization. Intelligent Data Analysis, 14, 89–119.

MARQUÉS, A., GARCÍA, V., SÁNCHEZ, J., (2012). Exploring the behavior of base classifiers in credit scoring ensembles. Expert Systems with Applications, 39(11), 10244–10250.

MICHALEWICZ, Z., FOGEL, D.B., (2004). How To Solve It: Modern Heuristics. Springer.

MOJIRSHEIBANI, M., (2002). A comparison study of some combined classifiers. Communications in Statistics - Simulation and Computation, 31(2), 245–260.

MORRISON, D.F., (1976). Multivariate Statistical Methods. McGraw-Hill.

NI, W., BROWN, S., MAN, R., (2009). Stacked partial least squares regression analysis for spectral calibration and prediction. Journal of Chemometrics, 23, 505–517.

OZAY, M., VURAL, F.T.Y., (2008). On the performance of stacked generalization classifiers. Lecture Notes in Computer Science, 5112, 445–454.

RAO, C.R., MITRA, S.K., (1971). Generalized Inverse of Matrices and its Applications. Wiley.

ROKACH, L., (2010). Ensemble-based classifiers. Artificial Intelligence Review, 33, 1–39.

ROONEY, N., PATTERSON, D., ANAND, S., TSYMBAL, A., (2004). Dynamic integration of regression models. Lecture Notes in Computer Science, 3077, 64–173.

ROONEY, N., PATTERSON, D., NUGENT, C., (2004). Reduced ensemble size stacking. Tools with Artificial Intelligence. ICTAI 6th IEEE International Conference, 266–271.

SCHAPIRE, R.E., (1990). The strength of weak learnability. Machine Learning, 5, 197–227.

SEBER, G.A.F., (1984). Multivariate Observations. New York: Wiley.

SEHGAL, M.S.B., GONDAL, I., DOOLEY, L., (2005). Stacked regression ensemble for cancer class prediction. Industrial Informatics INDIN 3rd IEEE International Conference, 831–835.

SESMERO, M., LEDEZMA, A., SANCHIS, A., (2015). Generating ensembles of heterogeneous classifiers using Stacked Generalization. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(1), 21–34.

SHUNMUGAPRIYA, P., KANMANI, S., (2013). Optimization of stacking ensemble configurations through artificial bee colony algorithm. Swarm and Evolutionary Computation, 12, 24–32.

SIGLETOS, G., PALIOURAS, G., SPYROPOULOS, C.D., HATZOPOULOS, M., (2005). Combining information extraction systems using voting and stacked generalization. Journal of Machine Learning Research, 6, 1751–1782.

WERNECKE, K., (1992). A coupling procedure for discrimination of mixed data. Biometrics, 48, 497–506.

WOLPERT, D., (1992). Stacked generalization. Neural Networks, 5, 241–259.

VAN DER HEIJDEN, F., DUIN, R.P.W., DE RIDDER, D., TAX. D.M.J., (2004). Classification, Parameter Estimation and State Estimation: An Engineering Approach Using Matlab, New York: Wiley.

WEBB, A., (2002). Statistical Pattern Recognition. New York: Wiley.

XU L., JIANG J.H., ZHOU Y.P., WU H.L., SHEN G.L., YU R.Q., (2007). MCCV stacked regression for model combination and fast spectral interval selection in multivariate calibration. Chemometrics and Intelligent Laboratory Systems, 87(2), 226–230.

ZHANG, H., (2004). The Optimality of Naive Bayes. 17. FLAIRS Conference 2004: Miami Beach, Florida, USA 562–567.

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 459–479, DOI 10. 21307

AN ADDITIVE RISKS REGRESSION MODEL FOR MIDDLE-CENSORED LIFETIME DATA

P. G. Sankaran[1], **S. Prasad**[2]

Abstract

Middle-censoring refers to data arising in situations where the exact lifetime of study subjects becomes unobservable if it happens to fall in a random censoring interval. In the present paper we propose a semiparametric additive risks regression model for analysing middle-censored lifetime data arising from an unknown population. We estimate the regression parameters and the unknown baseline survival function by two different methods. The first method uses the martingale-based theory and the second method is an iterative method. We report simulation studies to assess the finite sample behaviour of the estimators. Then, we illustrate the utility of the model with a real life data set. The paper ends with a conclusion.

Key words: additive risks model, counting process, martingales, middle-censoring.

1. Introduction

Middle-censoring introduced by Jammalamadaka & Mangalam (2003) occurs in situations where a data point becomes unobservable if it falls inside a random censoring interval. In such situations, the exact values are available for some individuals and for others, random censoring intervals are observed. To be more precise, let T be the random variable representing the lifetime of interest and let (U,V) be a bivariate random variable, representing the censoring interval, such that $P(U < V) = 1$. Under the middle-censored set-up, the exact lifetime T becomes unobservable if $T \in (U,V)$, and in such instances we only observe the censoring interval (U,V). Otherwise we observe T. We may find several such situations in survival studies and reliability applications. For example, in a prognostic study, the patients under observation may be withdrawn from the study for a short period of time for some unforeseen reasons and may return to the study with a changed status of event of interest. In reliability applications, it may happen that a failure of equipment occurs during a period of time when we accidentally fail to observe the study subjects. In

[1]Department of Statistics, Cochin University of Science and Technology, Kerala, India. E-mail: sankaran.p.g@gmail.com

[2]Department of Statistics, Cochin University of Science and Technology, Kerala, India. E-mail: hariprasadtvpm@gmail.com(Corresponding Author).

such contexts we only observe a censorship indicator and the interval of censorship.

As was pointed out by Jammalamadaka & Mangalam (2003), one can observe that the left censored data and right censored data are in fact special cases of this more general censoring scheme, by suitable choices of the interval, and also that such a censoring scheme is not complementary to the usual double censoring discussed in Klein & Moeschberger (2005) and Sun (2006).

Jammalamadaka & Mangalam (2003) pointed out various applications of middle-censoring scheme and developed a nonparametric maximum likelihood estimator (NPMLE) of the distribution function of the random variable. They proved that the NPMLE is always a self-consistent estimator (SCE) (Tarpey & Flury, 1996). Some rigorous treatments of this censoring scheme are found in Jammalamadaka & Iyer (2004), Iyer et al. (2008), Mangalam et al. (2008), Jammalamadaka & Mangalam (2009), Shen (2010, 2011), Davarzani & Parsian (2011) and Davarzani et al. (2015).

In survival studies, covariates or explanatory variables are usually used to represent heterogeneity in a population. The main objective in such situations is to understand and exploit the relationship between the lifetime and covariates. To this end we generally employ regression models. In the presence of covariates, Sankaran & Prasad (2014) discussed a parametric proportional hazards regression model for the analysis of middle-censored lifetime data. Jammalamadaka & Leong (2015) analysed discrete middle-censored data in the presence of covariates with an accelerated failure time regression model. Recently, Jammalamadaka et al. (2016) developed an iterative algorithm for analysing a semiparametric proportional hazards regression model under middle-censoring scheme, while Bennett et al. (2017) considered a parametric accelerated failure time regression model under this censoring scheme.

One extensively used semiparametric regression model is the well-known proportional hazards (PH) model by Cox (1972). It is a multiplicative hazards model in the sense that if T has a baseline hazard function $h_0(t)$ and if $\mathbf{z}$ is a $p \times 1$ vector of the recorded covariates then the hazard function of T conditional on $\mathbf{z}$ is modelled as

$$h(t|\mathbf{z}) = h_0(t)\exp(\mathbf{z}^\top \theta),$$

where $\theta = (\theta_1, \theta_2, ..., \theta_p)^\top$ is the vector of regression coefficients and $h_0(t)$ is left arbitrary. Here, $a^\top$ represents the transpose of vector a . In this model the effect of the covariates is acting multiplicatively on the baseline hazard function. But it is well known that in many occasions the PH model does not fit a given lifetime data well. One important alternative to the PH model is the additive risks (AR) model introduced by Aalen (1989) and later studied by Lin & Ying (1994). The model

associates the conditional hazard function with the covariates by

$$h(t|\mathbf{z}) = h_0(t) + \mathbf{z}^\top \theta. \tag{1}$$

In contrast to the PH model, the AR model given in (1) specifies that the hazard rate associated with a given set of covariates is the sum of the baseline hazard function and the regression function of covariates. This kind of model assumption is particularly useful in tumorigenicity experiments that investigate the dose effect on tumor risk, since the excess risk is often the quantity of interest (Breslow & Day, 1987). For a comprehensive review on properties and inference procedures of model (1), one may refer to Aranda-Ordaz (1983), Cox & Oakes (1984), Thomas (1986), Breslow & Day (1980), and Lin & Ying (1994). For a nonparametric treatment of model (1) one may refer to Aalen (1980, 1989). Model (1) is further explored in the context of left truncated current status data by Wang et al. (2015).

In the present work, we aim at estimating the unknown baseline survival function $S_0(t)$ of a continuous type lifetime variate T, which is subject to middle-censoring, and estimation of the unknown regression coefficients under model (1). We propose two different inference methods in Section 2. Simulation studies to assess the performance of the estimators under both methods for practical sample sizes are carried out and the results are compared in Section 3. The utility of the methods are illustrated with the help of a real life example in Section 4. Finally, some important conclusions are provided in Section 5.

2. Inference Procedure

Let the lifetime variate T admit an absolutely continuous cumulative distribution function (cdf) $F_0(t)$. Assume that T is middle-censored by the random censoring interval (U,V) having bivariate cdf given by $G(u,v) = P(U \leq u, V \leq v)$. Let us further assume that under model (1), T is independent of (U,V), given the covariate $\mathbf{z}$. Thus, one can observe the vector $(X,\delta,\mathbf{z})$, where

$$X = \begin{cases} T & \text{if } \delta = 1 \\ (U,V) & \text{if } \delta = 0, \end{cases}$$

and $\delta = I(X = T)$ is the uncensoring indicator. Now, we state an important assumption regarding the identifiability of the cdf $F_0(t)$. Let $[a,b]$, $a \leq b$ be any arbitrary interval in the support of T. Define, for $r \in [a,b]$,

$$A_0(r) = G(r-,\infty) - G(r-,r) = P(U < r < V). \tag{2}$$

Now, consider a situation where $A_0(r) = 1$ for $r \in [a,b]$ for which $F_0(b) > F_0(a-)$. *i.e.* censoring occurs with probability 1 on this interval where F_0 has a positive mass. Consequently, there will not be any exact observation in this interval, making it impossible to distinguish two distributions which are identical outside $[a,b]$ but differ only on $[a,b]$. To overcome this issue we make the following assumption.
A1: The probability defined in (2) is strictly less than one.

In the following we describe two different estimation methods: one makes use of the classic martingale theory and the other by using of an iterative method.

2.1. Martingale Method

Here we provide an inference procedure based on the martingale feature associated with the observed data. First a partial likelihood function is developed under model (1), similar to the one for the Cox PH model (Kalbfleisch & Prentice, 2011). Then, the stochastic integral representation of the score function derived from the partial likelihood function is used to infer about the unknown regression coefficient.

The observed data consists of n independent and identically distributed replicates $(X_i, \mathbf{z}_i,\ \delta_i)$ of $(X, \mathbf{z}, \delta), 1 \leq i \leq n$. When the lifetime is subject to middle-censoring, we shall define the counting process corresponding to the i'th individual as $N_i(t) = I(X_i \leq t, \delta_i = 1), t \geq 0$, which indicates whether the event occurred at time t, for $i = 1,2,...,n$. The at-risk process may be similarly defined as $R_i(t) = I(X_i \geq t, \delta_i = 1) + I(U_i \geq t, \delta_i = 0)$ which is a 0-1 predictable process, where the value 1 indicates whether the i'th individual is at risk at time t, for $i = 1,2,...,n$, *i.e.*, whether it is uncensored and waiting for a possible event at the epoch t. Denote the filtration $\sigma\{N_i(u), R_i(u+), \mathbf{z}_i : i = 1,2,...,n; 0 \leq u \leq t\}$ by $\mathscr{F}_t$. Under model (1) the conditional cumulative hazard rate for the i'th individual is given by $H(t|\mathbf{z}_i) = H_0(t) + \mathbf{z}_i^\top \theta t$, where $H_0(t) = \int_0^t h_0(a)\mathrm{d}a$ is the baseline cumulative hazard function. Model (1) assumes that

$$E[N_i(t)|\mathscr{F}_{t-}] = (h_0(t) + \theta^\top \mathbf{z}_i) R_i(t) dt,$$

and the intensity function corresponding to the counting process $N_i(t)$ can thus be written as $R_i(t)\mathrm{d}H(t|\mathbf{z}_i) = R_i(t)\{dH_0(t) + \mathbf{z}_i^\top \theta \mathrm{d}t\}$. With this, the counting process can be uniquely decomposed so that for every i and t,

$$N_i(t) = M_i(t) + \int_0^t R_i(a)\, \mathrm{d}H(a|\mathbf{z}_i), \tag{3}$$

where $M_i(\cdot)$ is a local square integrable martingale (Andersen & Gill, 1982). From (3), we have

$$dN_i(t) = dM_i(t) + R_i(t)\mathrm{d}H(t|\mathbf{z}_i), \tag{4}$$

so that

$$\sum_{i=1}^{n} dM_i(t) = \sum_{i=1}^{n} [dN_i(t) - R_i(t)(dH_0(t) + \theta^\top \mathbf{z}_i dt)] = 0. \tag{5}$$

To estimate θ, let us now consider the partial likelihood function suggested by Cox (1972) and further discussed in Cox (1975). It is defined as

$$L(\theta) = \prod_{i=1}^{k} \frac{h(t_{(i)}|\mathbf{z}_{(i)})}{\sum_{l=1}^{n} R_l(t_{(i)}) h(t_{(i)}|\mathbf{z}_l)}, \tag{6}$$

where $t_{(1)}, t_{(2)}, ..., t_{(k)}$ are the k observed exact lifetimes which are arranged in increasing order of magnitude. The motivation for (6) is that when we have the information that an event occurs at time point t and that the at-risk set is $R(t)$, the right-hand side of (6) is precisely the probability that it is individual $i \in R(t)$, who registered the event. Since T is assumed to be of continuous type, the possibility of ties is ruled out. However, (6) is not a usual likelihood, as it is not obtained from the probability of some observable events. A detailed discussion on this is available in Lawless (2011). Under the model assumption (1), we can rewrite (6) as

$$L(\theta) = \prod_{i=1}^{k} \frac{h_0(t_{(i)}) + \mathbf{z}_{(i)}^\top \theta}{\sum_{l=1}^{n} R_l(t_{(i)}) \left((h_0(t_{(i)}) + \mathbf{z}_l^\top \theta \right)}. \tag{7}$$

The value of θ that maximizes (7) can be obtained by maximizing

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{k} \Big[\log\Big(h_0(t_{(i)}) + \mathbf{z}_{(i)}^\top \theta \Big) - \log\Big(\sum_{l=1}^{n} R_l(t_{(i)}) \Big((h_0(t_{(i)}) + \mathbf{z}_l^\top \theta \Big) \Big) \Big]. \tag{8}$$

In terms of the counting process defined earlier, we can rewrite (8) as

$$C(\theta) = \sum_{i=1}^{n} \int_0^\infty \log\Big(h_0(s) + \mathbf{z}_i^\top \theta \Big) \mathrm{d}N_i(s) - \int_0^\infty \log\Big(\sum_{l=1}^{n} R_l(s) \big(h_0(s) + \mathbf{z}_l^\top \theta \big) \Big) \mathrm{d}\bar{N}(s), \tag{9}$$

where $\bar{N}(s) = \sum_{i=1}^{n} N_i(s)$. The score function is simply the derivative of (9) with respect to θ, and is given by

$$U(\theta) = \sum_{i=1}^{n} \int_0^{\infty} \left(h_0(s) + \mathbf{z}_i^{\top} \theta \right)^{-1} \mathbf{z}_i \mathrm{d}N_i(s) - \int_0^{\infty} \left(\sum_{l=1}^{n} R_l(s) \left((h_0(s) + \mathbf{z}_l^{\top} \theta \right) \right)^{-1} \left(\sum_{l=1}^{n} R_l(s) \mathbf{z}_l \right) \mathrm{d}\bar{N}(s). \quad (10)$$

Using the idea of Lin & Ying (1994), we propose to estimate the true regression coefficient θ_0 from the following estimating equation, which is obtained by an algebraic simplification of (10).

$$U(\theta) = \sum_{i=1}^{n} \int_0^{\infty} \mathbf{z}_i \{ \mathrm{d}N_i(t) - R_i(t) d\hat{H}_0(\theta, t) - R_i(t) \mathbf{z}_i^{\top} \theta \mathrm{d}t \},$$

which is equivalent to

$$U(\theta) = \sum_{i=1}^{n} \int_0^{\infty} \{ \mathbf{z}_i - \bar{\mathbf{z}} \} \{ \mathrm{d}N_i(t) - R_i(t) \mathbf{z}_i^{\top} \theta \mathrm{d}t \}, \quad (11)$$

where $\bar{\mathbf{z}} = \sum_{i=1}^{n} \mathbf{z}_i R_i(t) / \sum_{i=1}^{n} R_i(t)$, with the convention that $0/0 = 0$. The identity (11) is based on a simple fact that when θ_0 is the true parameter value, $U(\theta_0)$ is a martingale integral and therefore has mean zero. Note that (11) is linear in θ and the resulting estimator takes an explicit form given by

$$\hat{\theta} = \Big[\sum_{i=1}^{n} \int_0^{\infty} [\mathbf{z}_i - \bar{\mathbf{z}}]^{\otimes 2} R_i(t) \mathrm{d}t \Big]^{-1} \sum_{i=1}^{n} \int_0^{\infty} [\mathbf{z}_i - \bar{\mathbf{z}}] \mathrm{d}N_i(t), \quad (12)$$

where $a^{\otimes 2} = aa^{\top}$. Since $M_i(t)$ is a martingale, we have $\sum_{i=1}^{n} \mathrm{d}M_i(t) = 0$. Thus, from the representation given in (3), a Breslow type estimator (Breslow, 1972) for the cumulative hazard function $H_0(t)$ can be obtained as

$$\hat{H}_o(\hat{\theta}, t) = \int_0^{t} \frac{\sum_{i=1}^{n} \{ \mathrm{d}N_i(a) - R_i(a) \mathbf{z}_i^{\top} \hat{\theta} \mathrm{d}a \}}{\sum_{i=1}^{n} R_i(a)}. \quad (13)$$

This naturally leads to the following estimator of conditional survival function $S(t|\mathbf{z})$.

$$\hat{S}(t|\mathbf{z}) = \exp\{ -\hat{H}_0(\hat{\theta}, t) - \mathbf{z}^{\top} \hat{\theta} t \}. \quad (14)$$

An algebraic manipulation of (4) yields

$$U(\theta) = \sum_{i=1}^{n} \int_0^{\infty} (\mathbf{z}_i - \bar{\mathbf{z}}) dM_i(t), \tag{15}$$

which is a martingale integral. It follows from standard counting process theory that $n^{-1/2}U(\theta_0)$ converges weakly to a p-variate normal with mean zero and a covariance matrix that can be estimated consistently by

$$A = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} (\mathbf{z}_i - \bar{\mathbf{z}})^{\otimes 2} dN_i(t). \tag{16}$$

Also, the random vector $n^{1/2}(\hat{\theta} - \theta_0)$ converges weakly to a p-variate normal distribution with mean zero and a covariance matrix that can be consistently estimated by $B^{-1}AB^{-1}$, where

$$B = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} R_i(t)(\mathbf{z}_i - \bar{\mathbf{z}})^{\otimes 2} dt. \tag{17}$$

Specifically, $(B^{-1}AB^{-1})^{-\frac{1}{2}}(\hat{\theta} - \theta_0)$ converges in distribution to $N(0, I)$. It can be observed that neither A nor B involves the regression parameters. The estimator (13) provides the basis for estimating survival probabilities. Using standard counting process techniques, it follows that the process $\sqrt{n}(\hat{H}_0(\hat{\theta}, t) - H_0(t))$ converges weakly to a zero mean Gaussian process, whose covariance function at $(t, s), t \geq s$ can be estimated consistently by

$$\int_0^s \frac{n \sum_{i=1}^{n} dN_i(a)}{(\sum_1^n R_i(a))^2} + C'(t)B^{-1}AB^{-1}C(s) - C'(t)B^{-1}D(s) - C'(s)B^{-1}D(t),$$

where $C(t) = \bar{\mathbf{z}}t$ and $D(t) = \int_0^t \frac{\sum_1^n (\mathbf{z}_i - \bar{\mathbf{z}}) dN_i(a)}{\sum_1^n R_i(a)}$ with $k'(a) = \mathrm{d}k(a)/\mathrm{d}a$.
Using functional delta method (Andersen et al., 2012), it follows that the process $\sqrt{n}(\hat{S}(t|\mathbf{z}) - S(t|\mathbf{z}))$ converges weakly to a zero-mean Gaussian process, whose covariance function at $(t, s), t \geq s$ can be estimated consistently by

$$\hat{S}(t|\mathbf{z})\hat{S}(s|\mathbf{z}) \Big(\int_0^s \frac{n \sum_{i=1}^{n} dN_i(a)}{(\sum_1^n R_i(a))^2} + W'(t, \mathbf{z})B^{-1}AB^{-1}W(s, \mathbf{z}) + W'(t, \mathbf{z})B^{-1}D(s) + W'(s, \mathbf{z})B^{-1}D(t) \Big),$$

where $W(t, \mathbf{z}) = (\mathbf{z} - \bar{\mathbf{z}})t$.

2.2. The Iterative Method

In this section, an iterative method is proposed for estimating the unknown baseline survival function $S_0(t)$ of the lifetime variate T and the regression coefficient vector θ under model (1). Assume that T is middle-censored by the random censoring interval (U,V) such that, given the covariate $\mathbf{z}$, T and (U,V) are independently distributed. Let the observed data be as before. For convenience let us assume that the first n_1 observations are exact lifetimes, and the remaining n_2 are censored intervals, with $n_1+n_2=n$. Now, the likelihood corresponding to the observed data, excluding the normalizing constant, can be written as

$$L(\theta)=\prod_{i=1}^{n_1} f(t_i|\mathbf{z}_i)\cdot\prod_{i=n_1+1}^{n_1+n_2}\left(S(u_i|\mathbf{z}_i)-S(v_i|\mathbf{z}_i)\right). \tag{18}$$

Under the model assumption given in (1), the conditional survival function is obtained as

$$S(t|z)=S_0(t)\exp(-\theta^\top \mathbf{z}t), \tag{19}$$

where $S_0(t)=\exp(-H_0(t))$. Thus, the density function of T given $\mathbf{z}$ is given by

$$f(t|\mathbf{z})=\exp(-\theta^\top \mathbf{z}t)\big(\theta^\top \mathbf{z}S_0(t)-S_0'(t)\big). \tag{20}$$

Therefore, (18) becomes

$$L(\theta)=\prod_{i=1}^{n_1}\exp(-\theta^\top \mathbf{z}_i t_i)\big(\theta^\top \mathbf{z}_i S_0(t_i)-S_0'(t_i)\big)\times \prod_{i=n_1+1}^{n_1+n_2}\Big(S_0(u_i)\exp(-\theta^\top \mathbf{z}_i u_i)-S_0(v_i)\exp(-\theta^\top \mathbf{z}_i v_i)\Big). \tag{21}$$

The log-likelihood is given by

$$l(\theta)=\log L(\theta)=\sum_{i=1}^{n_1}\big(-\theta^\top \mathbf{z}_i t_i+\log(\theta^\top \mathbf{z}_i S_0(t_i)-S_0'(t_i))\big)+ \sum_{i=n_1+1}^{n_1+n_2}\log\big(S_0(u_i)\exp(-\theta^\top \mathbf{z}_i u_i)-S_0(v_i)\exp(-\theta^\top \mathbf{z}_i v_i)\big), \tag{22}$$

and its partial derivative with respect to θ_r, for $r=1,2,...,p$, is given by

$$\frac{\partial l(\theta)}{\partial \theta_r} = \sum_{i=1}^{n_1} z_{ir}(t_i + (\theta^\top \mathbf{z}_i S_0(t_i) - S_0'(t_i))^{-1} S_0(t_i)) + \sum_{i=n_1+1}^{n_1+n_2} z_{ir}\Big(S_0(u_i)\exp(-\theta^\top \mathbf{z}_i u_i) - S_0(v_i)\exp(-\theta^\top \mathbf{z}_i v_i)\Big)^{-1}\Big(v_i S_0(v_i)\exp(-\theta^\top \mathbf{z}_i v_i) - u_i S_0(u_i)\exp(-\theta^\top \mathbf{z}_i u_i)\Big), \tag{23}$$

where z_{ir} is the r'th component in the covariate vector corresponding to i'th individual. Note that (23) involves both unknown quantities θ and $S_0(t)$ and explicit solution for θ cannot be obtained directly from it. We provide an iterative algorithm to estimate the maximum likelihood estimates of these two quantities, where at each iteration a better update is obtained. To begin with the algorithm we consider the SCE of the baseline survival function as an initial approximation.

In the case of middle-censored data, Jammalamadaka & Mangalam (2003) showed that the NPMLE of $S_0(t)$ is always an SCE, which takes the form

$$\hat{S}_0(t) = 1 - \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i I(T_i \le t) + (1-\delta_i) I(V_i \le t) + (1-\delta_i) I(t \in (U_i, V_i)) \frac{\hat{F}_0(t) - \hat{F}_0(U_i)}{\hat{F}_0(V_i-) - \hat{F}_0(U_i)}\right\}. \tag{24}$$

Now, we give the algorithm in the following few steps.

Step 1. Set the vector $\theta = 0$.

Step 2. At the first iteration, find the SCE $S_0^{(1)}(t)$ of $S_0(t)$ using (24) and substitute this in (23) and solve $\partial l(\theta)/\partial \theta_r = 0, r = 1,2,...,p$ to get the estimator $\theta^{(1)}$ of θ.

Step 3. Find $\tilde{t}_i^{(1)} = S_0^{(1)^{-1}}\left(S_0^{(1)}(t_i)\exp(-\theta^{(1)^\top}\mathbf{z}_i t_i)\right)$ and similarly find $\tilde{u}_i^{(1)}$ and $\tilde{v}_i^{(1)}$ as our updated observations at the first iteration.

Step 4. At the j'th iteration ($j > 1$), use $\tilde{t}_i^{(j-1)}, i = 1,2,...,n_1$ and $(\tilde{u}_i^{(j-1)}, \tilde{v}_i^{(j-1)}), i = n_1+1,...,n$ as our data points in (24) and obtain $S_0^{(j)}(t)$. Substitute $S_0^{(j)}(t)$ in (23) and solve $\partial l(\theta)/\partial \theta_r = 0, r = 1,2,...,p$ to obtain the j'th iterated update $\theta^{(j)}$ of θ.

Step 5. Repeat Step 4 until convergence is met, say when $\|\theta^{(k)} - \theta^{(k+1)}\| < 0.0001$ and $\sup_t\left\{\left|S_0^{(k)}(t) - S_0^{(k+1)}(t)\right|\right\} < 0.001$, for some finite positive integer k.

Note that Step 3 in the algorithm is justified, because if $a_i = S_0^{(1)}(t_i)\exp(-\theta^{(1)^\top}\mathbf{z}_i t_i)$, then the a_i 's have a uniform distribution over $[0,1]$. Therefore, to scale these back to baseline distribution we need to find $\tilde{t}_i = \inf\{t : S_0^{(1)}(t) \le a_i\}$. Thus, the correct choice is $\tilde{t}_i = S_0^{(1)^{-1}}(a_i) = S_0^{(1)^{-1}}\left(S_0^{(1)}(t_i)\exp(-\theta^{(1)^\top}\mathbf{z}_i t_i)\right)$.

We now define our parameter space to be (Θ, Φ), where $\Theta \subseteq \mathbb{R}_p$ contains θ and

$\Phi = \{\phi(t) : [0,\infty] \to [0,1]$ and $\phi(\cdot)$ is absolutely continuous and nonincreasing$\}$ contains $S_0(t)$. Let us name the estimator obtained for θ as $\hat{\theta}_{(n)}$ and that for $S_0(t)$ as $\hat{S}_{0(n)}(t)$. Besides the identifiability condition A1, the following conditions are also assumed to hold for establishing the consistency property.
A2: Conditional on $\mathbf{z}$, T is independent of (U,V).
A3: The joint distribution of $(U,V,\mathbf{z})$ does not depend on the true parameter $(\theta^0, S_0^0(t))$.
A4: The covariate space is bounded. That is, there exist some finite $M > 0$ such that $P\{\|\mathbf{z}\| \le M\} = 1$, where $\|\cdot\|$ is the usual metric on $\mathbb{R}_p$.
A5: Distribution of $\mathbf{z}$ is not concentrated on any proper affine subspace of $\mathbb{R}_p$.

Theorem: Suppose that $\Theta \in \mathbb{R}_p$ is bounded and assumptions (A1) to (A5) hold. Then, the estimator $(\hat{\theta}_{(n)}, \hat{S}_{0(n)}(t))$ is consistent for the true parameter $(\theta^0, S_0^0(t))$ in the sense that if we define a metric $d : \Theta \times \Phi \to \mathbb{R}$ by

$$d\big((\theta_1, S_{01}(t)),(\theta_2, S_{02}(t))\big) = \|\theta_1 - \theta_2\| + \int |S_{01}(t) - S_{02}(t)| dF_0(t) + \left[\int \left((S_{01}(u) - S_{02}(u))^2 + (S_{01}(v) - S_{02}(v))^2\right) dG(u,v)\right]^{\frac{1}{2}}, \quad (25)$$

where $\theta_1, \theta_2 \in \Theta$ and $S_{01}(t), S_{02}(t) \in \Phi$, then $d\left((\hat{\theta}_{(n)}, \hat{S}_{0(n)}(t)), (\theta^0, S_0^0(t))\right) \to 0$ almost surely (a.s.).

Proof:
In the following discussion we denote $Y_i = (X_i, \delta_i)$. Let the probability function of $Y = (X, \delta)$ be given by

$$p(y;\theta, S_0(t)) = \prod_{i=1}^{n} f(t_i|\mathbf{z}_i)^{\delta_i} [S_0(u_i)\exp(-\theta^\top \mathbf{z}_i u_i) - S_0(v_i)\exp(-\theta^\top \mathbf{z}_i v_i)]^{1-\delta_i} \times g(u_i, v_i|\mathbf{z}_i) q(\mathbf{z}_i), \quad (26)$$

where g is the joint density of (U,V), conditional on $\mathbf{z}$ and q is the density of $\mathbf{z}$. Using $(A2)$ and $(A3)$, the log-likelihood function scaled by $1/n$ for the sample $(y_i, \mathbf{z}_i), i = 1,2,...,n$, up to terms not depending on $(\theta^0, S_0^0(t))$ is

$$l(\theta, S_0(t)) = \frac{1}{n}\sum_{i=1}^{n} \big\{\delta_i \log f(t_i|\mathbf{z}_i) + (1-\delta_i)\log[S_0(u_i)\exp(-\theta^\top \mathbf{z}_i u_i) - S_0(v_i)\exp(-\theta^\top \mathbf{z}_i v_i)]\big\}. \quad (27)$$

We write $p_n(y) = p(y; \hat{\theta}_{(n)}, \hat{S}_{0(n)}(t))$ and $p_0(y) = p(y; \theta^0, S_0^0(t))$ where $(\hat{\theta}_{(n)}, \hat{S}_{0(n)}(t))$ is the MLE that maximizes the likelihood function over $\Theta \times \Phi$ and $(\theta^0, S_0^0(t)) \in$

$\Theta \times \Phi$. Therefore,

$$\sum_{i=1}^{n} \log p_n(Y_i) \geq \sum_{i=1}^{n} \log p_0(Y_i)$$

and hence

$$\sum_{i=1}^{n} \log \frac{p_n(Y_i)}{p_0(Y_i)} \geq 0.$$

By the concavity of the function $x \mapsto \log x$, for any $0 < \alpha < 1$,

$$\frac{1}{n} \sum_{i=1}^{n} \log \left((1-\alpha) + \alpha \frac{p_n(Y_i)}{p_0(Y_i)} \right) \geq 0. \tag{28}$$

The left hand side can be written as

$$\int \log \left((1-\alpha) + \alpha \frac{p_n(Y_i)}{p_0(Y_i)} \right) \mathrm{d}(\mathbb{P}_n - \mathbb{P})(Y) + \int \log \left((1-\alpha) + \alpha \frac{p_n(Y_i)}{p_0(Y_i)} \right) \mathrm{d}\mathbb{P}(Y), \tag{29}$$

where $\mathbb{P}_n$ is the empirical measure of Y and $\mathbb{P}$ is the joint probability measure of Y. Let us assume that the sample space Ω consists of all infinite sequences $Y_1, Y_2, ...$, along with the usual sigma field generated by the product topology on $\prod_1^\infty (\mathbb{R}^3 \times \{0,1\})$ and the product measure $\mathbf{P}$. For p defined in (26) let us define a class of functions $\mathscr{P} = \left\{ p(y, \theta, S_0(t)) : (\theta, S_0(t)) \in (\Theta \times \Phi) \right\}$ and a class of functions $\mathscr{H} = \left\{ \log(1 - \alpha + \alpha p/p_0) : p \in \mathscr{P} \right\}$, where $p_0 = p(y, \theta^0, S_0^0(t))$. Then, it follows from Huang & Wellner (1995) that $\mathscr{H}$ is a Donsker class. With this and Glivenko-Cantelli theorem, there exists a set $\Omega_0 \in \Omega$ with $\mathbf{P}(\Omega_0) = 1$ such that for every $\omega \in \Omega_0$, the first term of (29) converges to zero. Now, fix a point $\omega \in \Omega_0$ and write $\hat{\theta}_{(n)} = \hat{\theta}_{(n)}(\omega)$ and $\hat{S}_{0(n)}(\cdot) = \hat{S}_{0(n)}(\cdot, \omega)$. By our assumption Θ is bounded, and hence for any subsequence of $\hat{\theta}_{(n)}$, we can find a subsequence converging to $\theta_* \in \Theta^C$, the closure of Θ. Also, by Helly's selection theorem, for any subsequence of $\hat{S}_{0(n)}(t)$, we can find a further subsequence converging to some nonincreasing function $S_{0*}(t)$. Choose the convergent subsequence of $\hat{\theta}_{(n)}$ and the convergent subsequence of $\hat{S}_{0(n)}(t)$ so that they have the same indices, and without loss of generality, assume that $\hat{\theta}_{(n)}$ converges to θ_* and that $\hat{S}_{0(n)}(t)$ converges to $S_{0*}(t)$. Let $p_*(y) = p(y, \theta_*, S_{0*}(t))$. By the bounded convergence theorem, the second term of (29) converges to

$$\int \log \left((1-\alpha) + \alpha \frac{p_*(y)}{p_0(y)} \right) d\mathbb{P}(y)$$

and by (28) this is nonnegative. But by Jensen's inequality, it must be non-positive.

Therefore, it must be zero and it follows that

$$p_*(y) = p_0(y) \quad \mathbb{P}-\text{almost surely.}$$

This implies

$$S_{0*}(t) = S_0^0(t) \quad F_0-\text{almost surely.}$$

Therefore, by bounded convergence theorem,

$$\int |\hat{S}_{0(n)}(t) - S_0^0(t)| \mathrm{d}F_0(t) \to 0. \tag{30}$$

Also,

$$S_{0*}(u)\exp(-\theta_*^\top \mathbf{z}u) = S_0^0(u)\exp(-\theta^{0^\top}\mathbf{z}u) \quad \mathbb{P}-\text{almost surely}$$

and

$$S_{0*}(v)\exp(-\theta_*^\top \mathbf{z}v) = S_0^0(v)\exp(-\theta^{0^\top}\mathbf{z}v) \quad \mathbb{P}-\text{almost surely.}$$

This together with (A5) imply that there exist $\mathbf{z}_1 \neq \mathbf{z}_2$ such that for some $c > 0$,

$$S_{0*}(c)\exp(-\theta_*^\top \mathbf{z}_1 c) = S_0^0(c)\exp(-\theta^{0^\top}\mathbf{z}_1 c)$$

and

$$S_{0*}(c)\exp(-\theta_*^\top \mathbf{z}_2 c) = S_0^0(c)\exp(-\theta^{0^\top}\mathbf{z}_2 c).$$

Since $S_{0*}(c) > 0$ and $S_0^0(c) > 0$, this implies $(\theta_* - \theta^{0^\top})(\mathbf{z}_1 - \mathbf{z}_2) = 0$. Again, by (A5), the collection of such $\mathbf{z}_1$ and $\mathbf{z}_2$ has positive probability and there exist at least p such pairs that constitute a full rank $p \times p$ matrix. It follows that $\theta_* = \theta^0$. This in turn implies that

$$S_{0*}(u) = S_0^0(u) \quad \text{and} \quad S_{0*}(v) = S_0^0(v) \quad G-\text{almost surely.}$$

Therefore, by bounded convergence theorem,

$$\int \left((\hat{S}_{0(n)}(u) - S_0^0(u))^2 + (\hat{S}_{0(n)}(v) - S_0^0(v))^2\right) \mathrm{d}G(u,v) \to 0. \tag{31}$$

Equations (30) and (31) together with $\theta_* = \theta^0$ hold for all $\omega \in \Omega_0$ with $\mathbf{P}(\Omega_0) = 1$. This completes the proof.

Table 1: Absolute bias, MSE and bootstrap coverage probability (BCP) of the estimator of θ under Method-1 and Method-2 with mild (10%) censoring

			$n=30$			$n=50$			$n=75$		
λ	θ	Method	Bias	MSE	BCP	Bias	MSE	BCP	Bias	MSE	BCP
0.1	0.25	1	0.0033	0.0008	0.903	0.0061	0.0054	0.900	0.0091	0.0067	0.898
		2	0.0347	0.0011	0.940	0.0380	0.0051	0.937	0.0396	0.0073	0.934
1.0	0.5	1	0.0104	0.0009	0.895	0.0134	0.0021	0.893	0.0163	0.0069	0.889
		2	0.0373	0.0018	0.928	0.0405	0.0063	0.924	0.0454	0.0078	0.920
2.5	-0.50	1	0.0077	0.0019	0.921	0.0089	0.0037	0.916	0.0108	0.0073	0.915
		2	0.0247	0.0012	0.926	0.0259	0.0049	0.925	0.0307	0.0067	0.921
4.0	-0.01	1	0.0336	0.0017	0.924	0.0366	0.0029	0.922	0.0410	0.0055	0.918
		2	0.0448	0.0013	0.934	0.0484	0.0062	0.931	0.0507	0.0106	0.929

Table 2: Absolute bias, MSE and bootstrap coverage probability (BCP) of the estimator of θ under Method-1 and Method-2 with moderate (20%) censoring

			$n=30$			$n=50$			$n=75$		
λ	θ	Method	Bias	MSE	BCP	Bias	MSE	BCP	Bias	MSE	BCP
0.1	0.25	1	0.0047	0.0024	0.901	0.0085	0.0095	0.897	0.0109	0.0114	0.893
		2	0.0366	0.0021	0.938	0.0401	0.0063	0.936	0.0424	0.0111	0.930
1.0	0.5	1	0.0121	0.0027	0.894	0.0152	0.0075	0.891	0.0197	0.0088	0.886
		2	0.0385	0.0028	0.927	0.0418	0.0102	0.922	0.0493	0.0126	0.918
2.5	-0.5	1	0.0091	0.0031	0.919	0.0101	0.0052	0.914	0.0139	0.0114	0.910
		2	0.0265	0.0025	0.925	0.0278	0.0078	0.923	0.0344	0.0115	0.917
4.0	-0.01	1	0.0346	0.0036	0.923	0.0402	0.0061	0.918	0.0435	0.0077	0.916
		2	0.0465	0.0032	0.932	0.0512	0.0083	0.927	0.053	0.0118	0.924

Table 3: Absolute bias, MSE and bootstrap coverage probability (BCP) of the estimator of θ under Method-1 and Method-2 with heavy (30%) censoring

			$n=30$			$n=50$			$n=75$		
λ	θ	Method	Bias	MSE	BCP	Bias	MSE	BCP	Bias	MSE	BCP
0.1	0.25	1	0.0057	0.0042	0.900	0.0118	0.0107	0.894	0.0151	0.0129	0.889
		2	0.0384	0.0031	0.937	0.0415	0.0079	0.934	0.0441	0.0147	0.925
1.0	0.5	1	0.0141	0.0041	0.892	0.0163	0.0121	0.889	0.0222	0.0143	0.882
		2	0.0405	0.0044	0.925	0.0429	0.0139	0.919	0.0539	0.0174	0.916
2.5	-0.5	1	0.0104	0.0050	0.918	0.0143	0.0097	0.909	0.0170	0.0151	0.904
		2	0.0283	0.0038	0.923	0.0296	0.0090	0.918	0.0372	0.0156	0.915
4.0	-0.01	1	0.0361	0.0056	0.921	0.0432	0.0088	0.916	0.0462	0.0101	0.913
		2	0.0484	0.0044	0.931	0.0541	0.0096	0.925	0.0561	0.0131	0.921

Remark 2.1

The asymptotic distributions of the estimators $\hat{\theta}_{(n)}$ and $\hat{S}_{0(n)}(t)$ do not seem to be easy to establish under the iterative method. We consider this as a problem for future research.

Remark 2.2

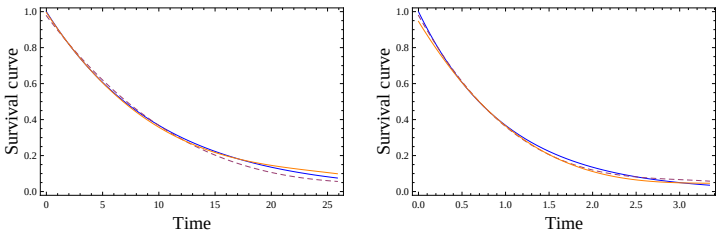
A likelihood ratio test can be carried out to test the significance of regression coefficients. The null hypothesis $H_0 : \theta = 0$ can be tested against $H_1 : \theta \neq 0$, where 0 is the null vector of the same order, with the test statistic $-2\log\frac{L(0)}{L(\hat{\theta})}$, which follows $\chi^2_{(p)}$ distribution. The test results in rejecting the null hypothesis for small P-values.

3. Simulation Studies

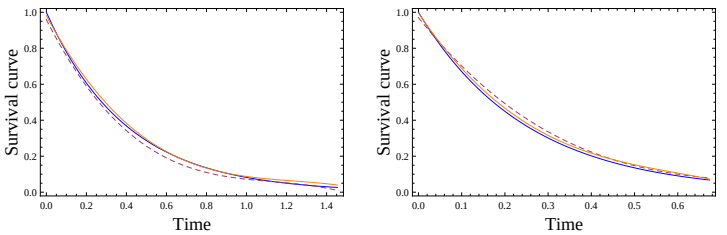
A simulation study is carried out to assess the finite sample properties of the estimators. We consider the exponential distribution with mean λ^{-1} as the distribution of lifetime variable T. Also, we choose independent exponential distributions with fixed means λ_1^{-1} and λ_2^{-1} as the distributions for the censoring random variate U and the interval of censorship $V-U$ respectively, and these two distributions are assumed to be independent of T. We consider a single covariate z in the present study, which is generated from uniform distribution over $[0,10]$ and let θ be the corresponding regression coefficient. Under the AR model in (1), the survival function of T given z may be written as

$$S(t|z) = S_0(t)\exp(-\theta z t), \tag{32}$$

where $S_0(t) = \exp(-\lambda t)$. It can be observed that (32) is the survival function corresponding to an exponential variate with mean $(\lambda+\theta z)^{-1}$. A large number of observations are generated from (32) for fixed values of λ and θ. Now corresponding to each observation on T, a random censoring interval is generated from (U,V), where the distribution parameters are fixed as $\lambda_1^{-1} = 20$ and $\lambda_2^{-1} = 10$. If we find $T \notin (U,V)$ then T is selected in the sample, otherwise we choose the interval as the observation. As we generate large number of observations we can now choose a sample of required size n. We consider three different censoring rates: 10% (mild), 20% (moderate) and 30% (heavy) for our inference. The martingale-based inference procedure, denoted as Method-1, and iterative inference procedure, denoted as Method-2, which are described in Section 2, are employed to obtain the estimates of $S_0(t)$ and θ and using 1000 iterations for various choices of λ and θ. The absolute bias and mean squared error (MSE) are computed and are given in Table 1 to Table 3. Also in each case, a 95% bootstrap confidence interval for regression parameter is computed. The proportion of times the true parameter value lies in such intervals is called bootstrap coverage probabilities (BCP). They are also reported in Table 1 to Table 3. It is evident that both bias and MSE are small in each case and they decrease as the sample size increases. The bootstrap coverage probabilities are found fairly large, close to one. Further, as the censoring rate increases the bias and MSE increase, while the BCP decreases. Also, for each combination of parameter values, and with sample size 75, we shall find out a cubic polynomial estimate of the form $S_0(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3$ with each of its coefficients being the average of corresponding coefficients obtained for all the iterations, for the baseline survival function. These estimated survival curves corresponding to both methods are plotted in Figure 1 to Figure 3, where continuous curve represents the true baseline

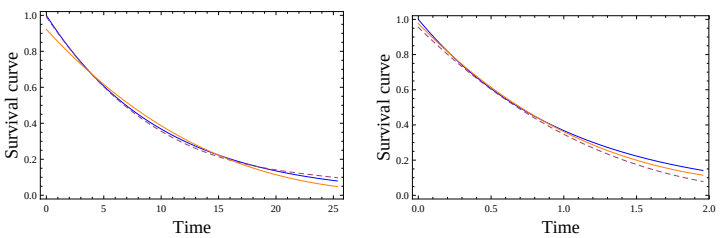


(a) $\lambda = 0.1$ (b) $\lambda = 1.0$

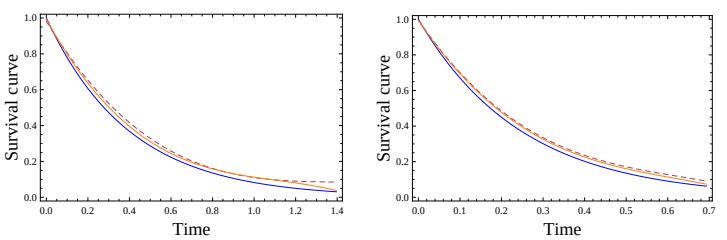


(c) $\lambda = 2.5$ (d) $\lambda = 4.0$

Figure 1: Plots of baseline survival curve and its estimates under Method-1 (dashed curve) and Method-2 (dotted curve) with mild (10%) censoring



(a) $\lambda = 0.1$ (b) $\lambda = 1.0$



(c) $\lambda = 2.5$ (d) $\lambda = 4.0$

Figure 2: Plots of baseline survival curve and its estimates under Method-1 (dashed curve) and Method-2 (dotted curve) with moderate (20%) censoring

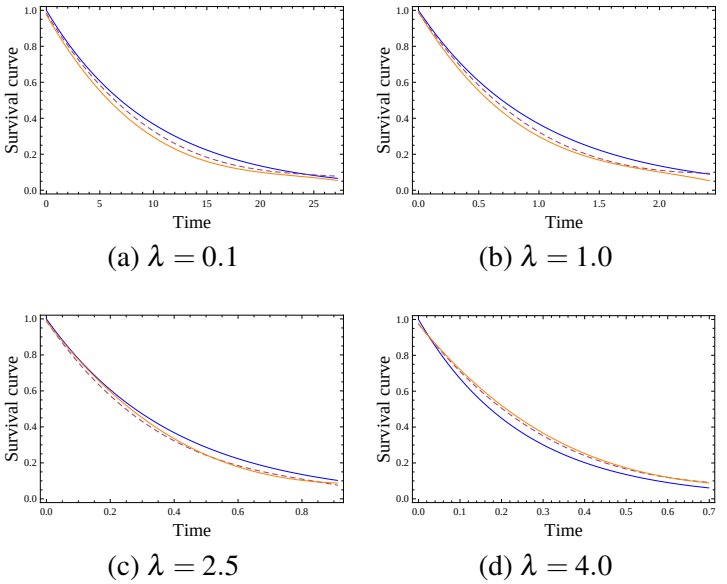


Figure 3: Plots of baseline survival curve and its estimates under Method-1 (dashed curve) and Method-2 (dotted curve) with heavy (30%) censoring

survival function and dashed curve represents the corresponding the estimate under Method-1 and dotted curve represents the corresponding estimate under Method-2. We see that both the estimated curves are close to the true curve.

4. Illustrative Data Analysis

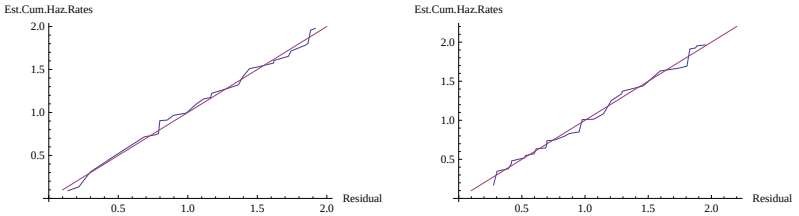
The proposed methods are applied to a real life data studied by Ichida et al. (1993). The data deals with an evaluation of a protocol change in disinfectant practices in a medical center where patients are suffering from burn wounds. The control of infection is the major concern in burn management and the study aims at comparing two different controlling methods: routine bathing care method and body cleansing method. The time (in days) until a patient develops staphylococcus infection is considered as the lifetime variable. Although the original study involves several covariates, for the illustration purpose we consider two of them, namely treatment (z_1), which is coded as 1-for routine bathing and 2-for body cleansing, and percentage of total surface area burned (z_2). Let θ_1 and θ_2 respectively be the unknown regression coefficients. A random censoring interval (U,V), where U and $V-U$ are independent exponential variates with means $\lambda_1^{-1} = 20$ and $\lambda_2^{-1} = 10$ is generated first. Then, an individual from among all exact 48 lifetimes is selected at random and if lifetime of the patient happens to fall in the generated censoring interval, that

Table 4: Estimates of coefficients of survival curve and regression coefficients under Method-1 and Method-2.

	$S_0(t)$				θ	
Method	c_0	c_1	c_2	c_3	θ_1	θ_2
1	0.93665	-0.04872	0.000635	-9.223e-6	0.0112	0.1005
2	0.96574	-0.05991	0.00121	-9.256e-6	0.00895	0.1760

lifetime is assumed to have middle censored and that interval is considered as the corresponding observation. Otherwise the lifetime is maintained. This process is repeated until around 25% of the observations are censored. The data resulted consists of twelve censored observations. We apply the two methods of estimation given in Section 2 and obtained the estimates of the baseline survival function of the form $S_0 = c_0 + c_1 t + c_2 t^2 + c_3 t^3$ and the regression coefficient θ. The estimated values, under both methods, of the coefficients of survival curves as well as regression coefficients are listed in Table 4. To test the significance of the covariate effect under the iterative method, we consider the null hypothesis $H_0 : \theta = 0$, where $\theta = (\theta_1, \theta_2)$ and 0 is null vector of the same order, and we use the likelihood ratio test described in Remark 2.2. The P-value of 0.008 indicates that the covariate effects are significant.

Now, we check the overall fit of the model by using Cox-Snell residuals (Cox & Snell, 1968). Suppose that the AR model given in (1) is fitted to the data. If the model assumption is correct then the probability integral transform of the true death time T assumes a uniform distribution over $[0, 1]$ or equivalently the random variable $H(T_j|\mathbf{z}_j)$, which is the true cumulative hazard function corresponding to (1), has an exponential distribution with hazard rate 1. Then, the Cox-Snell residuals are defined to be the fitted cumulative hazard function values $\hat{r}_j = \hat{H}_0(t_j) + \mathbf{z}_j^\top \hat{\theta}\, t_j$ with the estimated parameters. If the model is reasonable and the estimates of the parameters are close to the true values, then these quantities should look like a censored sample from unit exponential distribution. To check whether the r_j's behave as a sample from the unit exponential distribution we compute the Nelson-Aalen estimator of the cumulative hazard rate of r_j's. If the unit exponential distribution fits the data, then this estimator should be approximately equal to the cumulative hazard rate of the unit exponential distribution. Thus, a plot of r_j's versus their estimated cumulative hazard rates should be a straight line through origin and with a slope of 1. Figure 2 shows the plots so obtained under both the models. The curves are close to the straight line indicating AR assumption is reasonable.



(a) (b)

Figure 4: Plot of r_j's against estimated cumulative hazard rates under (a) Method-1 and (b) Method-2.

5. Conclusion

The present study discussed the semiparametric regression problem for the analysis of middle-censored lifetime data. We considered two different methods of estimation, one making use of martingale-based theory and the other based on an iterative method for which a maximization procedure for finding the NPMLE is developed. Large sample properties including consistency and weak convergence of the estimators were established under the martingale-based method. Consistency of estimators was proved under the iterative method, whereas their weak convergence do not appear to be easy to establish, although one can perhaps extend the ideas used in (Huang & Wellner, 1995). Simulation studies showed that the inference procedures were efficient. The model was applied to a real data set. Although we considered time-fixed covariates in this work, the procedure can easily be extended to the case of time-varying covariates, as in the work of Lin & Ying (1994). The middle-censored data has a connection with mixed interval-censored (MIC) data (Yu et al., 2001). Although both sampling schemes differ in character, the observed data from MIC will reduce to data from middle-censoring, when there are no left censored or right censored observations. For a detailed discussion on this interrelationship one may refer to Shen (2011).

Acknowledgements

We thank the editor and reviewers for their constructive comments.

References

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. *Mathematical Statistics and Probability Theory*, 1–25.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, *8*(8), 907–925.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (2012). *Statistical Models based on Counting Processes*. Springer Science & Business Media.

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100–1120.

Aranda-Ordaz, F. J. (1983). An extension of the proportional-hazards model for grouped data. *Biometrics*, 109–117.

Bennett, N., Iyer, S. K., & Jammalamadaka, S. R. (2017). Analysis of gamma and Weibull lifetime data under a general censoring scheme and in the presence of covariates. *Communications in Statistics - Theory and Methods*, *46*(5), 2277–2289.

Breslow, N. & Day, N. (1980). Conditional logistic regression for matched sets. *Statistical Methods in Cancer Research*, *1*, 248–279.

Breslow, N. E. (1972). Discussion of paper of D. R. Cox. *Journal of Royal Statistical Society Series B*, *34*, 216–7.

Breslow, N. E. & Day, N. E. (1987). *Statistical Methods in Cancer Research*, volume 2. International Agency for Research on Cancer, Lyon.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, *34*, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62(2)*, 269–276.

Cox, D. R. & Oakes, D. (1984). *Analysis of Survival Data*, volume 21. CRC Press.

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 248–275.

Davarzani, N. & Parsian, A. (2011). Statistical inference for discrete middle-censored data. *Journal of Statistical Planning and Inference*, *141*(4), 1455–1462.

Davarzani, N., Parsian, A., & Peeters, R. (2015). Statistical inference on middle-censored data in a dependent setup. *Journal of Statistical Theory and Practice*, *9*(3), 646–657.

Huang, J. & Wellner, J. A. (1995). Efficient estimation for the proportional hazards model with "case 2" interval censoring. *Technical Report No. 290, Department of Statistics, University of Washington, Seattle, USA*.

Ichida, J., Wassell, J., Keller, M., & Ayers, L. (1993). Evaluation of protocol change in burn-care management using the Cox proportional hazards model with time-dependent covariates. *Statistics in Medicine*, *12*(3-4), 301–310.

Iyer, S. K., Jammalamadaka, S. R., & Kundu, D. (2008). Analysis of middle-censored data with exponential lifetime distributions. *Journal of Statistical Planning and Inference*, *138*(11), 3550–3560.

Jammalamadaka, S. R. & Iyer, S. K. (2004). Approximate self consistency for middle-censored data. *Journal of Statistical Planning and Inference*, *124*(1), 75–86.

Jammalamadaka, S. R. & Leong, E. (2015). Analysis of discrete lifetime data under middle-censoring and in the presence of covariates. *Journal of Applied Statistics*, *42*(4), 905–913.

Jammalamadaka, S. R. & Mangalam, V. (2003). Nonparametric estimation for middle-censored data. *Journal of Nonparametric Statistics*, *15*(2), 253–265.

Jammalamadaka, S. R. & Mangalam, V. (2009). A general censoring scheme for circular data. *Statistical Methodology*, *6*(3), 280–289.

Jammalamadaka, S. R., Prasad, S. N., & Sankaran, P. G. (2016). A semi-parametric regression model for analysis of middle censored lifetime data. *Statistica*, *76*(1), 27.

Kalbfleisch, J. D. & Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*, volume 360. John Wiley & Sons.

Klein, J. P. & Moeschberger, M. L. (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media.

Lawless, J. F. (2011). *Statistical Models and Methods for Lifetime Data*, volume 362. John Wiley & Sons.

Lin, D. & Ying, Z. (1994). Semiparametric analysis of the additive risks model. *Biometrika*, *81*(1), 61–71.

Mangalam, V., Nair, G. M., & Zhao, Y. (2008). On computation of NPMLE for middle-censored data. *Statistics & Probability Letters*, *78*(12), 1452–1458.

Sankaran, P. G. & Prasad, S. (2014). Weibull regression model for analysis of middle-censored lifetime data. *Journal of Statistics and Management Systems*, *17*(5-6), 433–443.

Shen, P. (2010). An inverse-probability-weighted approach to the estimation of distribution function with middle-censored data. *Journal of Statistical Planning and Inference*, *140*(7), 1844–1851.

Shen, P. (2011). The nonparametric maximum likelihood estimator for middle-censored data. *Journal of Statistical Planning and Inference*, *141*(7), 2494–2499.

Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*. Springer Science & Business Media.

Tarpey, T. & Flury, B. (1996). Self-consistency: a fundamental concept in statistics. *Statistical Science*, 229–243.

Thomas, D. C. (1986). Use of auxiliary information in fitting nonproportional hazards models. *Modern Statistical Methods in Chronic Disease Epidemiology*, 197–210.

Wang, P., Tong, X., Zhao, S., & Sun, J. (2015). Regression analysis of left-truncated and case I interval-censored data with the additive hazards model. *Communications in Statistics - Theory and Methods*, *44*(8), 1537–1551.

Yu, Q., Wong, G. Y., & Li, L. (2001). Asymptotic properties of self-consistent estimators with mixed interval-censored data. *Annals of the Institute of Statistical Mathematics*, *53*(3), 469–486.

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 481–500, DOI 10. 21307

OPTION FOR PREDICTING THE CZECH REPUBLIC'S FOREIGN TRADE TIME SERIES AS COMPONENTS IN GROSS DOMESTIC PRODUCT

Luboš Marek[1], **Stanislava Hronová**[2], **Richard Hindls**[3]

ABSTRACT

This paper analyses the time series observed for the foreign trade of the Czech Republic (CR) and predictions in such series with the aid of the SARIMA and transfer-function models. Our goal is to find models suitable for describing the time series of the exports and imports of goods and services from/to the CR and to subsequently use these models for predictions in quarterly estimates of the gross domestic product (GDP) component resources and utilization. As a result we get suitable models with a time lag, and predictions in the time series of the CR exports and imports several months ahead.

Key words: transfer-function models, SARIMA models, quarterly estimates of the Gross Domestic Product (GDP), imports and exports of goods and services, exchange rate.

1. Introduction

Imports and exports of goods and services are ranked among the most important economic indices. The data of imports and exports describe the economic relationships from the viewpoint of goods and services circulation among residents and non-residents (in this sense, we consider the so-called national concept to foreign trade, based on the goods circulation as the change in ownership. Another possible approach is that of the so-called cross-border one, i.e. the principle of the goods passing the state borders); they express the extent to which the economy is open (regarding GDP). In this respect, they not only play an indispensable role for assessing economic performance, but are also significant components in the GDP estimate by the expense method.

[1] Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: marek@vse.cz.
[2] Department of Economic Statistics, University of Economics, Prague, Czech Republic. E-mail: hronova@vse.cz.
[3] Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: hindls@vse.cz.

Short-term (quarterly) estimates of GDP are based on two fundamental methods related to the two fundamental data sources: the production method is based on the estimates of the gross value added created in individual industries of the national economy, i.e. the GDP resources; and the expense method stems from the estimated components of GDP, i.e. final consumption, gross capital formation, and net imports. By making the estimates more accurate and better balanced, we get to the sole GDP value informing us about the economy evolution in the past quarter.

As data collection and processing become faster, and the economic evolution more turbulent during the year, the demands put forth by the users of statistical data are naturally growing with respect to the speed and quality of the short term estimates. These two parameters, speed and quality, go in opposite directions from the viewpoint of governmental statistics; if the user requires *quality* (i.e. as accurate as possible) statistical information, it must be derived from rather extensive surveys. Naturally, such surveys need time for preparation, collection and processing of data and, logically, such procedures are summarily used for annual data, or annual national accounts. If the user requires data *as quickly as possible* upon the completion of the period under consideration, each statistical office must necessarily take into account limited data sources and prefer modelling to direct findings. As s result we get quick estimates of the economy evolution during the year, but the user must allow for the estimates to be subsequently made more accurate. The quarterly estimates must be a kind of a trade-off between the speed and quality. If the GDP components are concerned, they must provide, as quickly as possible, reliable information on economic circulation during the year. The criterion of this reliability should be according to the ability to provide the source information for the initial estimates of the GDP annual values that would ensure the smallest possible deviations from the annual values based on extensive annual surveys. Hence, the quality of estimating the GDP components (imports and exports of goods and services, final consumption, and gross capital formation) plays the decisive role in this respect. Demands have been ever-growing for how quickly the initial information on the GDP evolution should be provided. Currently, it is expected that the initial estimates on the GDP growth rate should be available within 30 days from the end of the respective quarter in all EU countries; a more accurate GDP estimate together with the structure of resources and utilization within 60 days from that date; and the complete sector accounts within 90 days from that date.

The goal of the present paper is to provide a new, original methodological support to those quick model estimates – here they are focused on the estimates of imports and exports of goods and services as significant components of the expense method for the GDP estimate. The proposed original methodology will enable us not only to estimate the values for the imports and exports of goods and services, but also to calculate their estimates quicker, as source data for estimating the quarterly GDP. This model has been verified on the data for the imports and exports of goods and services to/from the Czech Republic found in the database

of the Czech Statistical Office (the data is stated in tsd. CZK) and the data of the exchange rate evolution taken from the database of the Czech National Bank. The analysis was carried out in the SCA software and eViews software.

2. Formulation of Problem

The issues connected with the short-term estimates for macroeconomic aggregates are inseparably connected with the effort to provide users of statistical data with reliable information about the evolution of the national economy as quickly as possible. Should such information be consistent, it must necessarily be viewed within a wider context of the system of macroeconomic statistical data, that is, the national accounts. In this sense, the short-term estimates are only relevant for a limited number of macroeconomic aggregates, namely, those entering the relationships in the GDP creation and utilization, based on the production and expense methods for estimating GDP (we are not going to consider the context of completing the quarterly national accounts in this paper; cf. e.g. Eurostat (2013a), Eurostat (2013b) or Marini (2016) for more details).

The GDP quarterly estimate, therefore, is based, on the one hand, on estimated gross value added, created in individual industries of the national economy (the estimates for resources, i.e. the gross value added in individual industries in the Czech Republic, are considered in the paper by Marek *et al.* (2016)) and, on the other hand, on the estimated expense components (imports and exports of goods and services, final consumption and gross capital formation) with subsequent balancing so that the sole value of the quarterly GDP is achieved.

For modelled (indirect) estimates of macroeconomic aggregates, a number of approaches can be utilized, stemming from the methods for time series analysis or regression analysis taking into account the relationships between annual and quarterly values. These methods are:

- without the quarterly or monthly data in the form of a reference index,
- using a reference index.

The estimating methods not utilizing a reference index enable us to get preliminary estimates of quarterly values, exclusively using formal mathematical procedures and criteria, providing smoothed quarterly estimates fulfilling a constraint that the sum of quarterly values over all four quarters equals the respective annual value. In other words, the annual value is disaggregated into quarters on the basis of purely formal criteria, without any knowledge about the evolution of the chosen index (or other indices) during the year. The best-known methods of this type include BFL (cf. e.g. Boot *et al.* (1967) or Wei and Stram (1990), Al-Osh (1989)). The usual models of time series (ARIMA) can also be viewed as members of this group. Such methods should only be used if a suitable reference index cannot be established, and for less important values. A natural utilization of them would also be a correction of an estimate obtained in the first step of desegregation with the aid of a reference index (for example, one variant

of the BFL method is based on minimizing the sum of squares of second differences, which criterion can also be used for corrections of an estimate after the first step).

The methods with the aid of a reference index make use of external quarterly, or even monthly, information about a related index (or several related indices). The main feature of these procedures is the use of a quarterly or monthly established index that is factually tied with the value of an annual aggregate, to facilitate the distribution of the annual value into quarterly ones. Mathematically, a formal (regression) model is used for the relationship between the annual value of the aggregate to be estimated and a quarterly (usually average) value of the reference index. This model makes it possible to get an initial estimate of the aggregate; in the second step this estimate is corrected so that the sum of all four quarterly values equals the annual value of the aggregate (in particular, the INSEE method is used, created by Bournay and Laroque (1979); more information about this method can be found in, e.g. Nasse (1973) or Dureau (1991); approaches making use of a dynamic variable can also be classified into this group, e.g. Moauro and Savio (2005) or Mitchell *et al.* (2005)).

In parallel with this traditional method of disaggregation with subsequent correction, methods of disaggregation without subsequent corrections have been developed and utilized; that is, methods which establish already at the first step such estimates for the aggregate's quarterly values that their sum complies with its annual value (this group of models contains, for example, those described by Chow and Lin (1971), and Kozák *et al.* (2000)).

Indisputably, the core aspect of the estimation quality in this group of methods is finding a suitable reference index. In the range of all indices coming into consideration, i.e. fulfilling the above-stated conditions, we have to seek for the one that best corresponds to the short-term evolution of the aggregate in question. This suitability should be observed on a prolonged time horizon and separately for each aggregate whose values are to be estimated with the aid of the disaggregation method. This stage requires thorough analytical work, which must not be either neglected or underestimated. When applying indirect methods, we need not only to create a formal statistical model but also to build up an entire system of short-term statistical data.

Another category of short-term estimating methods is focused on predicting quarterly values of aggregates (regardless of their relationships to the annual values) with the aid of methods used in time series analysis. Having in mind the nature of the problem, we can use classical decomposition, linear dynamic models or spectral analysis (more details of these methods can be found in, e.g. Green (2008), Pankratz (1991), Wei (2006), Anderson (1976), and Granger and Newbold (1986), Proietti (2011), Bikker (2013)). This paper offers one option for the utilization of the available short-term survey results in deriving estimates of the quarterly values from the underlying model. As a result of this approach, based on the analysis of time series, we obtain a stable and factually relevant model for estimating the quarterly values of imports and exports of goods and

services as components in the expense method for estimating GDP while using a suitable reference index. On a long-term basis, this model can be used for estimating quarterly values of other aggregates concerning the formation and use of GDP provided that the reference index is chosen appropriately.

This model is based on the utilization of the reference index and a time lag, due to which unknown values of imports and exports of goods and services can be estimated even without knowledge of the value of the reference index in the current (i.e. currently estimated) or future periods.

In this article, we focused precisely on a quick estimate of imports and exports of goods and services. This is especially due to the significance of foreign exchange for the Czech economy. Imports of goods and services in the Czech Republic is currently around 83% of GDP and exports of goods and services, about 77% of GDP (it is recalled that exports of goods and services are in FOB prices and imports of goods and services at CIF prices). The development of the values of these indicators is very closely related to the phases of the economic cycle because the basis of exports is mainly the products of the manufacturing industry and the basis of the import of raw material. For this reason, a timely and reliable estimate of the value of imports and exports of goods and services can be used to make a faster estimate of the quarterly GDP.

Imports and exports of goods and services are monitored in the Czech Republic in the so-called national concept. This is in line with the national accounts methodology, i.e. with the concept of other macroeconomic indicators. The national concept of foreign trade statistics follows up the actual trade in goods carried out between Czech and foreign entities, i.e. trade, where there is a change of ownership between residents and non-residents. Thus, the national concept of foreign trade reflects the export and import performance of the Czech economy better than the so-called cross-border concept. Conversely, the cross-border conception of foreign trade only reflects the physical movement of goods across the Czech Republic, irrespective of whether there is trade between Czech and foreign entities. These data, which only refer to the physical movement of goods from and into the territory of the Czech Republic, are surveyed for the purpose of international comparison of the movement of goods and services. However, they are not suitable indicators for monitoring the development of the economy in relation to GDP growth.

3. Methodology and results

In each of the above-mentioned groups classifying the time series analysis, there are many other approaches, and choosing from among them is governed by the character of the underlying data. For the time series we encounter here, we have selected a combination of models from the areas of stochastic methods and linear dynamic models to describe not only the behaviour of each time series separately, but also how values of one time series depend on those of another, and

to reflect this relationship in the model as well. As a prerequisite to the utilization of such a model in efficient short-term predictions, the time lag must be incorporated. When selecting suitable models, we viewed ARIMA and SARIMA ones (cf. Anderson (1976), Box *et al.* (1994), Granger and Newbold (1986), Wei (2006)), as well as transfer-function models (cf. Pankratz (1991), SCA Statistical System (1991)).

The sources of our data were the monthly series of imports and exports of goods and services to/from the Czech Republic (in tsd. CZK and current prices; source: www.czso.cz), and the time series of monthly average values of the CZK/EUR exchange rates (source: www.cnb.cz**)**. These series were available from January 1999 to September 2016 for imports and exports (213 observations), and from January 1999 to October 2016 for exchange rates (214 observations). The complete analysis was carried out in the SCA software. The values of the imports and exports of goods and services are published monthly by the Czech Statistical Office and, due to the character of this data (relationships to the quarterly and annual national accounts) they are reviewed several times. On the other hand, the values of the CZK/EUR exchange rates (monthly averages) are published by the Czech National Bank immediately upon the end of the respective month and are not reviewed any more.

We consider the stochastic models of time series in their general form:

$$\phi_p(B)\Theta_P(B^L)(1-B)^d(1-B)^D Y_t = \theta_q(B)\Theta_Q(B^L)\varepsilon_t \qquad (1)$$

where Y_t is the output series, ε_t is the random variable (white noise), B is the shift operator ($BY_t = Y_{t-1}$), L is the length of season, p is the order of AR process, q is the order of MA process, P is the order of seasonal AR process, Q is the order of seasonal MA process, d is the order of differencing, D is the order of seasonal differencing, ϕ_p is the autoregressive operator of order p, θ_q is the moving average operator of order q, Θ_P is the seasonal autoregressive operator of order P and Θ_Q is the seasonal moving average operator of order Q (cf. e.g. Box, Jenkins, Reinsel (1994)).

For model identification, the values of the autocorrelation and partial autocorrelation functions (ACF, PACF), and also the extended and inverse autocorrelation functions (EACF, IACF) were mainly used. The output is very extensive; hence only the most important aspects are explicitly mentioned. All the time series under analysis were non-stationary and had to be transformed to achieve stationarity (mostly by current and seasonal differentiating). The stationarity was tested by several approaches – the unit root, homoscedasticity, and Dickey-Fuller tests.

The values of the cross-correlation function (CCF) were calculated to prove the linear dependence between the analysed (already transformed, i.e., stationary) time series. These values confirmed the linear dependence between the transformed series of imports and exports of goods and services to/from the

Czech Republic and the transformed series of exchange rates. Afterwards, the general transfer-function model was applied:

$$Y_t = c + \nu_0 X_t + \nu_1 X_{t-1} + \nu_2 X_{t-2} + ... + \nu_K X_{t-K} + \frac{1}{(1-\phi_1(B))(1-\Phi_1(B^L))}\varepsilon_t \qquad (2)$$

where Y_t is the output series of exports, or imports (after the relevant transformations); X_t is the input series of the CZK/EUR exchange rates (again after the relevant transformations); and the last term is the perturbation series, denoted by N_t in the literature. The LTF method – cf. Pankratz (1991) and SCA Statistical System (1991) – was used for the parameter estimates. The resulting model was used for the predictions. The quality of the predictions was evaluated with the aid of the Theil coefficient of inequality.

$$TIE = \frac{\sum_{i=1}^{T}(\hat{Y}_t - Y_t)^2}{\sum_{i=1}^{T}(Y_{t-1} - Y_t)^2} \qquad (3)$$

When predicting, we first shortened the analysed time series and created the so-called *dormant predictions*, i.e. predictions for the periods in which we had already known the actual values of the time series. This approach enabled us to compare the predictions with the actual values and thus to assess the model quality in the most objective way. Only afterwards we predicted for several periods ahead. We cannot aim at a too ambitious horizon for the predictions because the economic conditions under which the time series evolves are quickly changing in time. Moreover, predictions for longer horizons are not necessary with respect to the nature of the problem in question.

3.1. CR Exports

The time series of the CR exports (denoted by *Exports* in our analyses) is seasonal and non-stationary. We applied current and seasonal differentiation to the series to achieve stationarity. The tests carried out confirmed our approach.

First of all, a SARIMA model suitable for this series was established. After a thorough analysis and study of ACF, PACF, EACF, IACF, and unit root, homoscedasticity, and Dickey-Fuller tests, we identified our model as (cf. the SCA output):

$$(1 - 0.3856B^3 + 0.177B^{10})Y_t = (1 - 0.6119B)(1 - 0.5628B^{12})\varepsilon_t$$

where $Y_t = (1-B)(1-B^{12})\,Export_t$, ε_t is the white noise, and B is the classical backward-shift operator ($B^k Y_t = Y_{t-k}$). This model was successful at all stages of verification and was proven as fully adequate; this fact is also indicated by the value of the index of determination, which amounts to 0.983.

Table 1. Model parameter estimates for the output series (SCA software output)

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL Exports
-------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
EXPORTN    RANDOM     ORIGINAL     (1-B  ) (1-B  )
-------------------------------------------------------------------------
PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-   VALUE      STD     T
   LABEL      NAME    DENOM.               TRAINT             ERROR  VALUE
  1    TH    EXPORTN    MA     1      1     NONE     .6119     .0635   9.64
  2   TH12   EXPORTN    MA     2     12     NONE     .5628     .0686   8.21
  3   PHI3   EXPORTN    AR     1      3     NONE     .3856     .0723   5.33
  4  PHI10   EXPORTN    AR     1     10     NONE    -.1770     .0693  -2.55

EFFECTIVE NUMBER OF OBSERVATIONS . .         185
R-SQUARE . . . . . .         .983
RESIDUAL STANDARD ERROR. . . . . .100513E+08
```

Subsequently, we determined the model for the CZK/EUR exchange rate series (denoted by *EUR* below). The corresponding SARIMA model is:

$$(1-B)X_t = (1-0.2186B)\varepsilon_t$$

where $X_t = EUR_t$, with the index of determination at 0.991. It is clear from the model that current differentiating was used to achieve stationarity.

Table 2. Model parameter estimates for the input series (SCA software output)

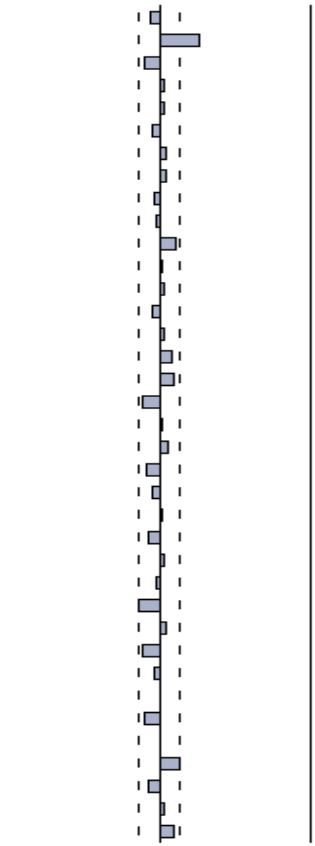
```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL --   EUR
-------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                       1
  EURN     RANDOM     ORIGINAL     (1-B  )
-------------------------------------------------------------------------
PARAMETER   VARIABLE  NUM./  FACTOR  ORDER   CONS-    VALUE      STD    T
   LABEL      NAME    DENOM.               TRAINT             ERROR  VALUE
   1   PHI     EURN     AR      1     1     NONE     .2186     .0633   3.45

EFFECTIVE NUMBER OF OBSERVATIONS . .         206
R-SQUARE . . . . . .         .991
RESIDUAL STANDARD ERROR. . . . . .364515E+00
```

Another output from the SCA software shows the values of the cross-correlation function between (now already stationary) time series $(1-B)(1-B^{12})Y_t$ and $(1-B)X_t$. The Cross-Correlation Function (CCF) values indicate not only the intensity of mutual linear dependence between the differentiated series, but also the direction of that dependence.

Both the values and the curve imply what the significant value of CCF is (95% confidence interval) at time *t*-1. We identified significant linear dependence between the transformed time series of exports at time *t* and the transformed time series of the CZK/EUR exchange rates at time *t*-1.

Correlations are asymptotically consistent approximations



DEXPORT,DRATE(-i)	DEXPORT,DRATE(+i)	i	lag	lead
		0	-0.0654	-0.0654
		1	0.2995	-0.0598
		2	-0.1062	0.0029
		3	0.0392	-0.0497
		4	0.0326	0.0646
		5	-0.0424	0.0219
		6	0.0572	-0.0566
		7	0.0437	0.0583
		8	-0.0389	0.0151
		9	-0.0170	-0.0912
		10	0.1223	0.0299
		11	0.0214	0.0134
		12	0.0261	-0.0842
		13	-0.0544	0.1061
		14	0.0263	-0.0730
		15	0.0873	-0.0410
		16	0.0944	0.0850
		17	-0.1116	-0.0586
		18	0.0228	-0.0176
		19	0.0647	0.0635
		20	-0.0888	-0.0598
		21	-0.0429	-0.0430
		22	0.0177	0.0879
		23	-0.0669	-0.0995
		24	0.0340	0.0978
		25	-0.0241	-0.0667
		26	-0.1385	-0.0766
		27	0.0554	0.0935
		28	-0.1156	0.0036
		29	-0.0259	-0.0293
		30	0.0132	-0.0263
		31	-0.0998	0.0087
		32	0.0008	0.0132
		33	0.1371	-0.0361
		34	-0.0779	-0.0103
		35	0.0317	0.0826
		36	0.0915	0.0179

Figure 1. CCF evolution (95% confidence interval), eViews output

Next we identified the transfer-function model. The LTF method – cf. Pankratz (1991) – was used for this identification. The value of the v_1 weight was

the only one that was significantly different from zero; it means that the values of the output series of exports depend on those of the input series of the CZK/EUR exchange rates with a time lag equal to one. This fact had already been indicated by the CCF values. The remaining weights were identified as insignificant by our testing. After a thorough analysis we had thus established a suitable model and estimated its parameters. The resulting transfer-function model hence is:

$$(1-0.4387B^3+0.1394B^{10})Y_t =$$
$$=\text{-}5,706,000+7,253,000*X_{t\text{-}1}+(1-0.6316B)(1-0.5541B^{12})\varepsilon_t$$

where $Y_t=(1-B)(1-B^{12})\,Export_t$ and $X_t=(1-B)*EUR_t$.

Table 3. Parameter estimates for the TFM model (SCA software output)

```
SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- EXPORT1
-------------------------------------------------------------------------
VARIABLE   TYPE OF    ORIGINAL     DIFFERENCING
          VARIABLE   OR CENTERED
                                        1      12
EXPORTN    RANDOM     ORIGINAL     (1-B  ) (1-B  )
                                        1      12
  EURN     RANDOM     ORIGINAL     (1-B  ) (1-B  )
-------------------------------------------------------------------------
PARAMETER  VARIABLE  NUM./  FACTOR  ORDER   CONS-    VALUE     STD     T
   LABEL       NAME    DENOM.              TRAINT            ERROR  VALUE
  1    V0      EURN    NUM.    1     0     NONE -.5706E+07  .165E+07  -3.45
  2    V1      EURN    NUM.    1     1     NONE  .7253E+07  .166E+07   4.37
  3   PHI    EXPORTN    MA     1     1     NONE       .6316     .0634   9.97
  4  PHI12   EXPORTN    MA     2    12     NONE       .5541     .0698   7.94
  5   PHI3   EXPORTN    AR     1     3     NONE       .4387     .0705   6.22
  6  PHI10   EXPORTN    AR     1    10     NONE      -.1394     .0659  -2.12

EFFECTIVE NUMBER OF OBSERVATIONS . .        185
R-SQUARE . . . . . .       .984
RESIDUAL STANDARD ERROR. . .. . .   .956100E+07
```

This model was successfully verified (with the aid of the above-mentioned procedures and methods) and established as fully adequate. Its quality is, among other things, confirmed by the value of the index of determination, amounting to 0.984. The CCF values between the residuals of the SARIMA and transfer-function models did not significantly differ from zero, which can be seen in the output cited below. In other words, the model we established complied with one of the most important verification criteria.

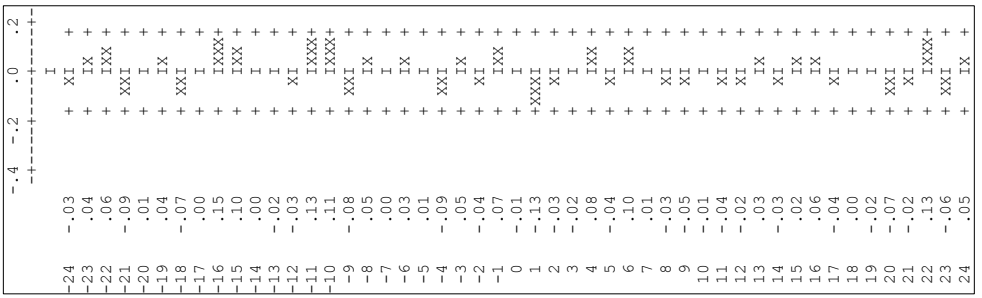


Figure 2. Evolution of CCF residuals (95% confidence interval)

Therefore, we can say that the value of the time series of imports (after the current and seasonal differentiating) at time *t* depends on the past values of the series itself (with time lag values of 3 and 10), the values of the time series of the CZK/EUR exchange rates (after current differentiating) at time *t*–1, and the past values of the random component (with a seasonal parameter).

Let us now have a look at the predictions because of which both of the above-described estimates were predominantly derived. As of the time of writing the present paper, the values of the time series of exports are known up to August 2016, but those of the exchange rates up to October 2016. Therefore, we were able to make use of the time lag, and input into the transfer-function model not predictions (which would be usual) but the actually observed values. Of course, we expected improvement of the predictions from this step. The situation is illustrated in the following Figure:

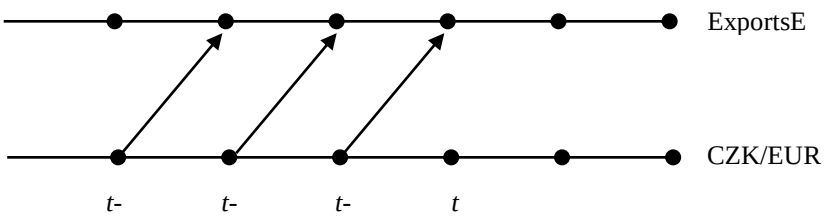


Figure 3. Linear relationship between the exports and CZK/EUR exchange rates

3.2. Predictions for Exports

Table 4 shows the five-month-ahead predictions of the shortened time series of exports as derived with the aid of the SARIMA and the transfer-function (TFM) models. Thanks to our shortening the time series by five values, we can compare the predictions with the actual values.

Table 4. Exports – predictions versus actual values

	Predictions		Actual values
Month	SARIMA	TFM	
May	300,450,000	307,651,000	328,285,676
June	319,340,000	329,600,000	349,080,472
July	287,780,000	287,840,000	277,464,937
August	286,780,000	296,590,000	309,977,136
September	316,340,000	326,160,000	347,200,140

Table 5 sums up the standard deviation values of the predictions. The comparison between the predicted and actual values is satisfactory (cf. Table 6). Somewhat more accurate predictions and smaller values of the standard deviation are in favour of the TFM model.

Table 5. Exports – standard deviations of predictions

	Standard deviations	
Month	SARIMA	TFM
May	9,274,300	9,221,400
June	9,978,900	9,948,600
July	1,063,700	1,063,100
August	1,268,200	1,260,600
September	1,355,400	1,354,800

Table 6. Exports – comparison of predictions versus actual values

	Difference (prediction minus actual value)		Ratio (prediction/actual value)	
Month	SARIMA	TFM	SARIMA	TFM
May	-27,835,676	-20,634,676	0.885	0.937
June	-29,740,472	-19,480,472	0.858	0.944
July	10,315,063	10,375,063	1.037	1.037

August	-23,197,136	-13,387,136	0.861	0.957
September	-30,860,140	-21,040,140	0.882	0.939

Table 7. Theil coefficient of inequality

SARIMA	TFM
0.3985	0.1879

The Theil coefficient of inequality clearly indicates that TFM is better.

After comparing the models with the actual values, we decided for the transfer-function one. Making use of the full available length of the time series, we predict five months ahead. The results are shown in Table 8. Of course, the actual values are unknown to the authors at the time of writing this paper; hence the accuracy can only be measured after five additional months.

Table 8. Exports – predictions

Month	TFM
October	336,240,000
November	360,020,000
December	299,350,000
January	314,050,000
February	332,310,000

3.3 CR Imports

Let us now analyse the time series of imports. The procedure is identical with the one we used for exports. Again, we created the SARIMA model, this time for the time series of imports, and the TFM one – imports depending on the CZK/EUR exchange rates. We made the predictions and compared them.

Here, we only state the particular models and predictions, without detailed reasoning and software output (except for the resulting TFM model).

3.4 Predictions for Imports

The following formula describes the suitable SARIMA model for the time series of imports:

$$(1-0.3368B^3-0.2126B^5+0.2968B^{10})Y_t=(1-0.6029B)(1-0.6622B^{12})\varepsilon_t,$$

where $Y_t = (1-B)(1-B^{12})\,Import_t$, and ε_t is the white noise. This model successfully passed the complete verification stage. The index of determination equals 0.978 – hence we would hardly be able to find a better model.

The SARIMA model for the series of the CZK/EUR exchange rates was already identified when constructing the model for exports. We can directly continue to the construction of the TFM model. According to the SCA output stated below, the TFM formula is:

$$(1-0.3331B^3-0.2036B^5+0.3227B^{10})Y_t = 1{,}070{,}000 * X_{t-1} + (1-0.6289B)(1-0.7207B^{12})\varepsilon_t$$

where $Y_t = (1-B)(1-B^{12}) * Import_t$ and $X_t = (1-B) * EUR_t$

This model also passed successfully all stages of the verification procedure and was found fully adequate, with the value of the index of determination amounting to 0.977. Table 10 shows the prediction values and Table 11 their standard deviation values.

Table 9. Parameter estimates for the TFM model (SCA software output)

SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- IMPORTS1

VARIABLE	TYPE OF VARIABLE	ORIGINAL OR CENTERED	DIFFERENCING
IMPORT	RANDOM	ORIGINAL	$(1-B^1)(1-B^{12})$
EURN	RANDOM	ORIGINAL	$(1-B^1)$

	PARAMETER LABEL	VARIABLE NAME	NUM./ DENOM.	FACTOR	ORDER	CONSTRAINT	VALUE	STD ERROR	T VALUE
1	V1	EURN	NUM.	1	1	NONE	.1070E+07	.834E+06	1.28
2	TH	IMPORT	MA	1	1	NONE	.6289	.0620	10.15
3	TH12	IMPORT	MA	2	12	NONE	.7207	.0528	13.65
4	PHI3	IMPORT	AR	1	3	NONE	.3331	.0703	4.74
5	PHI5	IMPORT	AR	1	5	NONE	.2036	.0677	3.01
6	PHI10	IMPORT	AR	1	10	NONE	-.3227	.0691	-4.67

EFFECTIVE NUMBER OF OBSERVATIONS . . 190
R-SQUARE 977
RESIDUAL STANDARD ERROR. . . .. . .962642E+07

Table 10. Imports – predictions versus actual values

	Predictions		Actual values
Month	SARIMA	TFM	
May	290,450,000	289,300,000	286,759,551
June	299,340,000	299,600,000	302,126,184
July	287,780,000	272,840,000	250,955,952
August	266,780,000	269,590,000	275,437,737
September	306,340,000	305,160,000	298,809,680

At first sight we can see that the SARIMA and TFM predictions do not differ from each other to a great extent. However, the TFM predictions are closer to the actual values – this can also be seen in Table 12. In Table 11, we can see the standard deviations of the predictions, which are again smaller for TFM.

Table 11. Imports – standard deviations of predictions

	SARIMA	TFM
May	9,274,300	9,021,400
June	9,978,900	9,548,600
July	1,062,700	913,100
August	1,262,800	1,005,600
September	1,395,400	1,054,800

Table 12. Imports – comparison of predictions versus actual values

	Difference (prediction minus actual value)		Ratio (prediction/actual value)	
Month	SARIMA	TFM	SARIMA	TFM
May	3,690,449	2,540,449	1.013	1.009
June	-2,786,184	-2,526,184	0.991	0.992
July	36,824,048	21,884,048	1.147	1.087

August	-8,657,737	-5,847,737	0.969	0.979
September	7,530,320	6,350,320	1.025	1.021

Table 13. Theil coefficient of inequality

SARIMA	TFM
0.3772	0.1415

The Theil coefficient of inequality clearly indicates that TFM is better.

For the imports, the transfer-function model was also established as better. We made the five-month-ahead predictions within an approach identical to that for the exports. The results are shown in Table 14.

Table 14. Imports – predictions

Month	TFM
October	962,640,000
November	314,880,000
December	270,950,000
January	278,410,000
February	288,920,000

The above-presented model clearly shows that the theory of stochastic models for time series was used (whether SARIMA or the theoretically more demanding TFM models); this theory by itself is more complex than, for example, the often-used decomposition of time series. Both models provide good results. However, the transfer-function model (TFM) is better with respect to accuracy of predictions, values of the standard deviation of the estimates, as well as to the Theil coefficient of inequality. For these reasons, the authors prefer the TFM model despite the fact that this model is more complex. Our analysis has clearly established that using simpler models is out of the question due to the nature of the data. Of course, the analysis itself is very demanding and laborious for the same reason.

Both SARIMA and TFM are relatively complex models. Those who deal with such models know very well that it would certainly be possible (having in mind the duality between auto-regression and moving averages) to identify more models with a different (or similar) structure of independent variables, and such models could be used for predicting alternatively. However, such models would

hardly be simpler and there is a question whether better predictions would be achieved using them. The authors deal with such models and predictions based on them on a long-term basis. On the basis of empiric experience it can be confirmed that the given models are robust, i.e. relatively stable in time regarding the individual variables (components) in the model. It means that the parameter estimates are changing in time, but the models as such remain stable regarding the structure of the variables.

When the present paper is being published, other data have been published for the foreign-trade time series (the data for several months have been modified and/or added; and later data for approximately the last two years have been changed within the framework of regular reviews of the quarterly and annual national accounts). But the data for the exchange rates will remain unchanged. A question hence arises whether or not our predictions become useless. The answer is, of course, that they do not: the above-described predictions are valid for the given data and model and are important at the time of being calculated, because they can be used for subsequent analyses, source information for decision-making, and future considerations. The validity of predictions and reviews of the original data represent a far more general problem; this problem is valid for most data published by each and every statistical office. What is most important in this context is the model robustness. If the model is robust, it can be used even when the data changes, recalculating just the values of the parameters and predictions.

4. Conclusions

The goal of the present paper is to establish models for the time series of exports and imports of goods and services from/to the Czech Republic suitable for the construction of short-term predictions. Our analysis has proved mutual linear dependence between monthly time series (exports and imports) of the foreign trade of the Czech Republic, expressed in million CZK in current prices, and the time series of the CZK/EUR exchange rates (monthly averages).

Within the framework of our analysis, we created suitable SARIMA models for all the time series concerned. We have also derived transfer-function models for the series of exports and imports, in which the input time series were those of the CZK/EUR exchange rates. When predicting in a TFM model, predictions of the input series are usually used, on the basis of which predictions of the output series are calculated. Thanks to the linear dependence (with a time shift) between the series we have proved within the analysis, as well as the different times of publishing the time series values (the exchange rate values are published several months earlier than those of the foreign trade), we can utilize the actual values of the input series (i.e. the exchange rates) instead of their predictions in the TFM model. This approach, logically, leads to better quality of predictions for the output series (exports or imports for our case). All these facts can be utilized in estimating the evolution of the CR foreign trade on a short time horizon of two to

three months. Nevertheless, it should be noted that (and this comment is even more important at the time of a crisis) predictions of economic time series are purposeful only a few periods ahead because the external influences on their evolution are quickly changing, thus in principle disabling the creation of good-quality long-term predictions. Moreover, we should realize that the time series values of the CR foreign trade published on the website of the Czech Statistical Office are just initial estimates for the most recent periods. Such initial estimates are subsequently reviewed and made more accurate (related to the reviews of the quarterly and annual national accounts). On the other hand, the models described in the present paper enable us, at a relatively high level of quality, to provide estimates of closely watched economic time series that substantially affect quarterly estimates of GDP.

The authors feel that the importance of such an analysis is characterized not so much by publishing the actual values of the predictions but rather by the methods, procedures and models used – those can be instructive for predicting in time series of the foreign trade of the Czech Republic, as important components of quick estimates of GDP.

Acknowledgements

This paper was written with the support of the Czech Science Foundation project No. P402/12/G097 "DYME – Dynamic Models in Economics" and with the support of the Institutional Support to Long-Term Conceptual Development of Research Organization, the Faculty of Informatics and Statistics of the University of Economics, Prague.

REFERENCES

AL-OSH, M., (1989). A Dynamic Linear Approach for Disaggregating Time Series Data, Journal of Forecasting, Vol. 8, pp. 85–96.

ANDERSON, O. D., (1976). Time series analysis and forecasting – Box-Jenkins approach, London, Butterworth.

BIKKER, R., DAALMANS, J., MUSHKUDIANI, N., (2013). Benchmarking large accounting frameworks: a generalized multivariate model, Economic Systems Research, Vol. 25, pp. 390–408.

BOOT, J. C. G., FEIBES, W., LISMAN, J. H. C., (1967). Further Method of Derivation of Quarterly Figures from Annual Data, Applied Statistics, Vol. 16, pp. 65–75.

BOURNAY, J., LAROQUE, G., (1979). Réflexions sur la méthode d´élaboration des comptes trimestriels, Annales de l'INSEE, Vol. 11, pp. 3–18.

BOX, G. E. P., JENKINS, G. M., REINSEL, G. C., (1994). Time series analysis, forecasting and control, Third edition. Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

DENTON, F. T., (1991). Adjustment of monthly and quarterly series to annual tools: An approach based on quadratic minimization, Journal of American Statistical Association, Vol. 66, pp. 99–102.

DUREAU, G., (1991). Les comptes nationaux trimestriels, Paris: INSEE – Méthodes.

FORNI, M., HALLIN, M., LIPPI, M., REICHLIN, L., (2005). The generalized factor model: one-sided estimation and forecasting. Journal of the American Statistical Association, Vol. 100, pp. 830–840.

GRANGER, C. W. J., NEWBOLD, P., (1986). Forecasting Economic Time Series, Academic Press, New York.

GREENE, W. H., (2012). Econometric Analysis, Pearson Education, Prentice Hall.

EUROSTAT, (2013a). European System of Accounts (ESA 2010), Eurostat, Luxembourg.

EUROSTAT, (2013b). Handbook on quarterly national accounts, Eurostat, Luxembourg.

KOZÁK, J., HINDLS, R., HRONOVÁ, S., (2000). Some Remarks to the Methodology of the Allocation of Yearly Observations into Seasons, Statistics in Transition, Vol. 4, pp. 815–826.

MAREK, L., HRONOVÁ, S., HINDLS, R., (2016). Příspěvek k časnějším odhadům hodnot čtvrtletních národních účtů [Contribution to the earlier estimations of quarterly national accounts], Politická ekonomie, Vol. 64, pp. 633–650.

MARINI, M., (2016). Nowcasting Annual National Accounts with Quarterly Indicators: An Assessment of Widely Used Benchmarking Methods. IMF Working Paper 16/71.

MITCHELL, J., SMITH, R. J., WEALE, M. R., WRIGHT, S., SALAZAR, E. L., (2005). An Indicator of Monthly GDP and an Early Estimate of Quarterly GDP Growth. Economic Journal, Vol. 115, pp. 108–129.

MOAURO, F., SAVIO, G., (2005). Temporal Disaggregation Using Multivariate Structural Time Series Models, Econometrics Journal, Vol. 8, pp. 214–234.

NASSE, PH., (1973). Le système des comptes nationaux trimestrieles, Annales de l´INSEE, Vol. 5, pp. 119–161.

PANKRATZ, A., (1991). Forecasting with dynamic regression models, John Wiley & Sons Inc., New York.

PROIETTI, T., (2011). Multivariate temporal disaggregation with cross-sectional constraints, Journal of Applied Statistics, Vol. 38, pp. 1455–1466.

THE SCA STATISTICAL SYSTEM, (1991). Reference manual for general statistical analysis, Scientific Associates Corp. Oak Brook, Illinois, USA.

WEI, W. W. S., (2006). Time series analysis – univariate and multivariate methods, Pearson, Addison Wesley Publishing, New York.

WEI, W. W. S., STRAM, D. O., (1990). Disaggregation of Time Series Models, Journal of Royal Statistical Society, Vol. 52, pp. 453–467.

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 501–520, DOI 10. 21307

SUBJECTIVE APPROACH TO ASSESSING POVERTY IN POLAND – IMPLICATIONS FOR SOCIAL POLICY

Leszek Morawski[1], **Adrian Domitrz**[2]

ABSTRACT

The poverty rates based on the OECD scales are frequently used in public debate. In this scale, large families are usually identified as those most in need of financial support. Poland is an interesting case for applying an alternative, subjective approach to calculating equivalent scales, as Poland has a large mean size for households, and is dependent on means-testing in social policymaking. The overall poverty rates for the two approaches are not distinctly different but they lead to significantly different distributions of poverty, as different types of households are considered in line with the result in Bishop et al. (2014) for the eurozone countries. The subjective approach suggests that one-person households, not large families, should be considered most at risk of material poverty. Futhermore, the relative positions of households in the income distributions also differ considerably. As a consequence, the current shape of social policy in Poland may need to be reconsidered in order to distribute public transfers more accurately.

Key words: subjective poverty, household equivalence scale, social policy.

Introduction

In 2010, households with two adults and three or more children were at the relatively highest risk of poverty in Poland. The at-risk-of-poverty rate calculated for the poverty line set at 60% of median equivalised income was 32.8% in this group. This value for one-person households was 24.5%, for two adults with one child it was 12.3%, while for households classified as "at least three adults with a child" it was 19.5% in the same year. The overall rate in 2010 was 17.7%. The equivalised income applied in those calculations was based on a modified OECD equivalence scale that gave a weight of 1.0 to the first adult in a household, a

[1] Vistula University (Stokłosy 3, 02-787 Warsaw, Poland), Institute of Economics, the Polish Academy of Sciences (Pl. Defilad 1, 00-091 Warsaw, Poland). E-mail: lmorawski15@gmail.com.
[2] University of Warsaw, Department of Economic Sciences, Długa 44/50, 00-241 Warsaw, Poland.

weight of 0.5 to the second one and to each subsequent person aged 14 and over, and a weight of 0.3 to each child aged under 14.[3]

The Eurostat data clearly pointed to "large households" as those units at which social transfers need to be targeted. The poverty statistics published by the Polish Central Statistical Office (CSO) make this conclusion even stronger. The recently published information has revealed the poverty rate among parents with 4 and more children to be equal to 43.7% and among parents with 3 children to reach 25.8%. The overall rate published was 16.7%. These rates were calculated using expenditure data and the original OECD equivalence scale with weights equal to: 1, 0.7, and 0.5 (GUS, 2011a).

Despite growing literature on the non-income factors influencing "subjective well-being" and the multidimensional character of poverty, financial transfers still play a major role among the used solutions. The discussion about the official poverty statistics that are based on the OECD scales may significantly influence the allocation of social financial transfers. For example, in the parliamentary campaigns in Poland in 2007 and 2011, all major parties proposed policies targeted toward large families, which were perceived as needing special assistance on the grounds of the official poverty statistics. Recently, the new Polish government launched a very generous social programme called the "Family 500+". According to this regulation, 500 PLN (117 Euro) per month will be paid unconditionally for the second and each additional child in a family. The income criteria as 800 PLN per month per person (1200 PLN in the case of a disabled child in a family) was introduced for families with one child. It is estimated that about 3.7 children is eligible for that benefit.

Using subjective information on income evaluation is not a new idea and it may be partially attributed to the criticism of the "revealed preferences" concept as an indicator of "true" individual well-being by behavioural welfare economists (Veenhoven, 2002; Schokkaert et al., 2011). This may be attributed to the fact that the equivalence scales derived from the consumer demand data using the basis of the revealed preferences theory suffer from identification problems and, thus, some extra conditions are needed in order to calculate them with such an approach (Pollak and Wales, 1979, Blundell and Lewbel, 1991). Some authors suggested using subjective information from survey declarations about happiness or income satisfactions as a solution to the identification problem (Lewbel and Pendakur, 2008). Apart from that, there are authors developing other empirical methods such as matching estimators or indifference equivalence scales, both based on scrutiny of individual level behaviour.

In practice, simple OECD scales, either the "original" or the "modified" ones are commonly used. Two recent studies show significant differences between the subjective and the OECD scales (Bollinger et al., 2012; de Ree et al., 2013). The study of Bollinger et al. (2012) for England suggests, for example, larger scale

[3] All numbers in the section are from the Eurostat webpage: http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do (last access on 2017.06.05).

economies within couples and substantial diseconomies due to any additional person after considering subjective information on income evaluation. We believe that considering subjective information about income evaluation may lead to interesting results that may be not consistent with those obtained with conventional OECD scales. Also, additional motivation for this paper is the fact that, to our knowledge, subjective poverty in Poland is quite limited despite the fact that the works regarding subjective equivalence scale were initiated already in the 1990s by Podgórski (1990, 1991, 1994). His research showed much flatter equivalence scales implied by the subjective approach than the commonly used OECD scales. More recent works applying a subjective approach for Poland are those of Dudek (2009), Dudek (2012), Dudek and Landmesser (2012), Kalbarczyk-Stęclik (2016). The subjective approach to poverty is discussed in Panek and Czapiński (2015) in research based on data from the Social Diagnosis Program (Diagnoza Społeczna).

The household sector in Poland is dominated by a small size structure (in terms of number of people). A significant share of multi-person households may be of importance to poverty analysis results since the difference between the OECD scales and subjective scale is increasing in household size. This is exactly the case of Poland, which with 2.8 people per household belongs to the group of countries with the largest average household size among EU countries. A similar household size is observed in other less-developed European countries such as Slovakia, Cyprus, Romania, Malta and Bulgaria, while Germany, Netherlands, France, United Kingdom and all the Scandinavian countries are the ones with much smaller average household sizes. At the same time, the structure of the household sector regarding the number of children in a household observed in Poland is very similar to the EU25 average - in the case of households with 4 children the fractions are 2.7% and 2.6%, for those with 3 children 8.6% and 9.0%, while for those with 2 children it is as high as 35.2% and 38.9% (Iacovou and Skew, 2010).

Frequent use of the statistics based on the OECD scales in public discussion, the size structure of the household sector together, as well as differences between the OECD scale and subjective scale, make Poland an interesting case for asking what would happen if politicians used the subjective scale instead of the OECD scale as reference. In this paper we ask whether the conclusion about the need for special treatment of large families is sensitive to a choice of equivalence scales. Although many other approaches to equivalence scales and to poverty analysis as a whole are possible, we take a closer look at comparison of these two methods in detail: OECD (the so-called expert scales) and subjective (known as Leyden Poverty Line) scales. Such an approach allows us to focus on the range and nature of discrepancies between them and to open a broader discussion on avenues of future research on such differences. We restrict our analysis to income poverty keeping in mind the importance of non-monetary measures and multidimensionality of poverty. In the paper we concentrate on the income dimension since we consider it to be most important, as justified by the

Atkinson's argument against that the separation between inequality of outcome and inequality of opportunity. According to his argument, the current inequality of outcome directly determines the future inequality of opportunity (Atkinson, 2017). The aim of the paper – comparison of the subjective equivalence scales with the OECD scales - is very closely related to the paper by Bishop et al. (2013), who made a similar exercise for the eurozone countries using the EU-SILC data.

The structure of the paper is as follows. The first section describes the methodology, namely the Leyden Poverty Line method. The second section contains the results of the estimation of the Leyden Poverty Line for Poland for 2010. The third section compares poverty incidence implied by the subjective approach with the results based on the OECD equivalence scales. The last section summarises our results and contains the final conclusion.

1. Method

In this paper, poverty is defined by the level of welfare that is just sufficient enough for a household to function properly in a society (as in: Van Praag, 1971; Van Praag and Van der Sar, 1988; Van den Bosch, 2002). If we narrow this concept solely to the question of income, we can say that "poverty" begins when a household's material situation (or income) is somehow too low to maintain a basic living standard without serious difficulty. The subjective poverty approach lets every person evaluate his or her income according to his or her feelings or needs. A subjective poverty line can be derived upon these evaluations. This is in significant contrast to the objective poverty approach, in which experts define either absolute or relative poverty lines. The objective approach is straightforward to use in practice but ignores a person's perception of income. On the other hand, the subjective method that takes into account a person's opinion about the actual material needs assumes cardinality of the utility function, which is a disputable issue. However, the subjective approach to empirical research in social science has been getting some popularity because of the recognition that many economic indicators or concepts that had been considered to use *ordinal* utility, de facto assume some sort of cardinality. Such indicators include, among others, the commonly used equity measures as they ascribe certain values to, for example, income inequality in order to say and compare which income distribution is *better* or *worse* (Ferrer-i-Carbonelland and Frijters, 2002; Van Praag and Ferrer-i-Carbonell, 2004; Binder and Coad; 2011).

In this research we return to the approach postulated by the Leyden school based on the Income Evaluation Question, in which a person (presumably the head of the household) declares income amounts corresponding to certain verbal qualifiers. Following the Leyden approach, we assume that 1) households are able to evaluate income in general as well as their own income, also in terms of verbal labels; 2) it is possible to sensibly convert the labels into a numerical evaluation

of welfare on a bounded scale such as an interval [0,1]. These claims are based on an assumption that if a respondent tries to do his best in describing his welfare using a five-label scale, he should respond as if the differences of welfare between all income levels were identical since it maximizes the information value of the respondent's answer. Such claims were criticized by Seidel (1994) and defended in Van Praag and Kapteyn (1994).

The empirical specification used below follows Kapteyn and Van Herwaarden (1981) claims that a log-normal cumulative distribution function fits best the responses from the Income Evaluation Question. That is why we assume the following relation between income and welfare:

$$\Lambda(\mathrm{y_i};\mu_i,\sigma_i) \equiv \Phi\left(\frac{\log(\mathrm{y_i})-\mu_i}{\sigma_i}\right), \tag{1}$$

with $\Phi(.)$ being a standardized cumulative distribution function, μ_i describing *the needs of a household* measured by the income demanded by it to satisfy a certain level of welfare and σ_i which defines the welfare sensitivity of income. This allows us to write the (logarithm of) δ-specific poverty line for a household with income y_i as:

$$\log\big(y_i(\delta)\big) = \mu_i + \sigma_i * \Phi^{-1}(\delta), \tag{2}$$

A parameter μ_i can be estimated by a sample mean of the declared log-incomes for each of welfare points. Estimator of σ_i that reflects how much income a household requires to change its welfare evaluation from one level to another is a sample standard deviation of declared log-incomes. We estimate individual effects μ_i by the Ordinary Least Squares regression, while σ_i is set at the value of sample average as it was found to be difficult to explain. The basic specification for μ_i includes only household size and income:

$$\log\big(y(\delta)\big) = (\beta_0 + \beta_1 * \log(L) + \beta_2 * \log(y)) + \sigma * \Phi^{-1}(\delta) \tag{3}$$

Equation (3) is called a Social Standard Function and it allows us to calculate the income y_δ that is needed for a certain household size to achieve a social standard (welfare) δ (Van Praag and Ferrer-i-Carbonell, 2004). It differs from the individual Welfare Function of Income (eq. (2)) in three points: 1) it concerns social standard income y_δ instead of the current individual income; 2) it takes into account the interaction between current income and household needs, which is a phenomenon called a preference drift; 3) it yields welfare of a social group (defined by household of size L) instead of the individual value. Defining the poverty line as the income y, which brings the welfare δ for a household with the current income equal to y, allows us to write:

$$y(L,\delta) = \exp\left(\frac{\beta_0 + \beta_1 * \log(L) + \bar{\sigma} * \Phi^{-1}(\delta)}{1-\beta_2}\right). \tag{4}$$

Other factors such as the income of a reference group, age of the head of the household, age of children and other socio-economic variables can also be included in the financial needs regression (Van Praag, 1971; Van den Bosch, 2002). This leads to more complex poverty lines in the form of $y(L, X, \delta)$, where X includes other variables explaining financial needs. Van den Bosch (2002) suggested using either the level 0.5 (as poverty risk) or 0.4 (as poverty). Many of the studies using the above approach were conducted by researchers closely related to Van Praag and by Van Praag himself, for example: Van Praag (1971), Van Praag and Kapteyn (1976), Van Praag and Van der Sar (1988), De Vos and Garner (1991) or Van Praag and Ferrer-i-Carbonell (2004). In most cases, the functional form of the household needs or minimum income regression included income and household or family size as the only explanatory variables. Apart from the models with only income and household size as explanatory variables, other variables used were: age of the head of the household, age of children, gender of the head of the household, working status or number of workers in a household, education level and occupation of the head of the household. Generally, there is a lot of diversity in the results of the Leyden poverty incidence, presumably caused by the differences in functional forms of the regression. The comparability of results across countries is difficult due to a multitude of reasons such as differences in methodology of surveys, size of samples, as well as cultural aspects concerning, for example, life aspirations in a society and understanding of terms such as poverty, welfare, or minimum standards. Nonetheless, the direction of explanatory variables influence is quite similar in most studies and generally the equivalence scales implied by the Leyden approach indicate considerable economies of scale within households.

2. Data and results

All calculations in this paper are based on the Polish Household Budget Survey dataset (*orig.* Badanie Budżetów Gospodarstw Domowych, BBGD). The data comes from the 2010 wave of the HBS that includes the IEQ with five levels: "very bad", "hardly sufficient", "sufficient", "good", and "very good". The PHBS is a countrywide survey based on a random sample of households that is conducted every year by the Central Statistical Office (further: CSO; *orig.* Główny Urząd Statystyczny - details on the Polish HBS survey methodology can be found in GUS, 2011b). The monthly rotation of households method is applied, which means that households participate in a survey for only one month. Consequently, all the information reflects a state of the household in the very moment of taking part in the survey, in particular: income obtained and expenditures made are recorded throughout the month of the interview. All these questions are asked at the end of the month of the interview and are recorded at the household level. Asking income evaluation questions after a month of conducting a diary of incomes and expenditures should give more reliable

answers. Work, disability or marital status, age, educational level, etc., are recorded at the beginning of the month, and are updated at the end of the month. Altogether, the HBS provides extraordinarily detailed information on each household and its members. Specifically, there are personal characteristics, labour market activity, incomes from work and outside of work – available at the individual level; as well as housing conditions, expenditures and, above all, subjective evaluation of income – recorded at the household level.

The total sample size of the HBS 2010 exceeds 37 thousand households and corresponds to about 13.3 million households after applying the population weights. Within these households there are altogether almost 108 thousand persons, equivalent to about 37.7 million people in Poland. The most frequent group of households is the one-person household that accounts for almost one-fourth of the population. Only slightly less frequent is the household with two members – over 23% of population. The other household types are in quite similar proportions as without weighing: three- and four-person households account for ca. 20% and 18%, respectively, five-person households for about 8% and the "6+" group for almost 6% of all households (Table A1 in Appendix).

The amounts declared by the households in the IEQ differ considerably for each of the evaluation levels. Declarations of "very bad" income range from 50 PLN to as high as 25 000 PLN per household, reaching its mean at about 1320 PLN, and its median at exactly 1000 PLN. Similar variations apply for the other levels, but the answers seem consistent in that their mean and median values are always higher for each subsequent level. In the whole database there are no records of declarations, for example, stating higher amount of "very bad" income than for "sufficient" one. High variability of income evaluations proves that households' perception of income needs is quite heterogeneous - suggesting that the same amount of money for one household brings different satisfaction (or welfare) for the other one. In fact, it is one of the reasons for utilizing the Leyden approach.

Table 1 presents the estimation results corresponding to the equation (3) for two specifications. The basic form contains two explanatory variables: current income and number of household members, while the extended one includes information about the number of persons aged 14 or over, the number of persons aged 13 or less in a household, education, socio-economic household type (farmers, pensioners, those living on unearned sources), and town size. A dependent variable is declared available income, which refers to the total monthly net household income as defined by the Central Statistical Office. It comprises income from hired work, income from a private farm in agriculture, income from self-employment other than a private farm in agriculture, income from freelance work, income from property, income from rental of a property or land, social insurance benefits and other social benefits and other income. Independent variables explain more than 60.00% of the total variance of μ, although even more important is the fact that standard errors of estimators are low.

Table 1. Comparison of diagnostic results and parameter estimates from basic and extended models

	Basic model	Extended model
No. of observations	37 106	37 106
R-squared	62.06%	64.54%
_constant	4.043 (142.0)	4.614 (122.9)
log(household_size)	**0.151 (43.9)**	x
log(adults)	x	**0.189 (44.8)**
log(children+1)	x	**0.422 (6.1)**
log(income)	**0.449 (117.3)**	**0.390 (81.0)**
log(income)*log(children+1)	x	**-0.431 (-5.04)**
higher_education (d)	x	0.079 (17.4)
Socio-economic groups:		
farmers (d)	x	-0.030 (-2.8)
pensioners (d)	x	-0.060 (-16.5)
unearned_sources (d)	x	-0.128 (-12.6)
Town size:		
town_medium (d)	x	-0.119 (-25.5)
town_rural (d)	x	-0.176 (-34.6)
link test (square of fitted values t-statistic, p-value)	-5.0 (0.000)	-1.0 (0.298)

Source: Own calculations; HBS 2010.

Notes: Incomes lower than 1 PLN dropped out and incomes truncated at 0.1% and 99.9% centile. Robust covariance matrix is applied. For link test there are test statistics values and p-values in parenthesis; for explanatory variables there are parameter estimates and t-statistics in parenthesis. All variables are significant at 1% level; (d) stands for dummy variables; the base level for socio-economic groups contains households of workers and the self-employed; the base level for town size is a large city (above 500 thous. inhabitants).

Two interesting observations follow from these estimates. First, there is a positive relation between the current income and the financial needs as is generally postulated by the literature (Stevenson and Wolfers, 2008). For example, according to a basic model a financial need of a single household with an income of 500 PLN is 928 PLN and of a 4-person household with such income

needs 1141.5 PLN. The needs for the same types of households are much higher if they have 5000 PLN - the respective values are 2610 PLN and 3218 PLN. Such positive preference drift in income valuation means that the *ex-ante* income valuation is higher than the *ex-post* valuation.

Second, the family size elasticity is rather low and equal to 0.27. According to the presented estimates in Table 1, a childless couple needs an income that is higher by 11.75% than a single household, while parents with a child should have an income 15.02% higher than a childless couple to reach the same utility level.

The extended model suggests more complicated relation between the financial need, current income and household size. Still, a positive sign for the estimates on income is still the evidence of positive preference drift, whereas a negative value of the interaction suggests a decreasing drift in the number of children.

The coefficients of categorical variables look sensible, as the highest material needs are obtained for households of employees and the self-employed, living in a large city and with an educated head of household, e.g. a household where the head is highly educated needs about 8% more income to be equally satisfied than a household where the head does not have higher education. Having estimated household needs regression allows us to calculate the poverty lines for all household sizes. A modified version of equation (4) takes the form of:

$$y(\delta) = \exp\left(\frac{\beta_0 + \beta_1 * \log(adults) + \beta_2 * \log(children+1) + \sum_d \beta_d x_d + \bar{\sigma} * \Phi^{-1}(\delta)}{1 - \beta_3 - \beta_4 * \log(children+1)}\right), \quad (5)$$

where $\sum_d \beta_d x_d$ stands for summing up dummy variables coefficients.

In regard to the financial needs, the extended model gives a much wider picture of household diversity than the basic one, which shows that the subjective income evaluation is based also on variables other than the household size.

The results from the models fit well with those published by the Polish official statistics. In 2010 the poverty line for a single household was estimated by the CSO at PLN 1187, and for a couple with two children at PLN 1770 (GUS, 2011a). The poverty line at the average values of all explanatory variables except for the number of adults and children obtained from the extended model is PLN 1212 for a single household and PLN 1797 for a couple with two children. The basic model yields a line of PLN 1182 for a single household and PLN 1725 for a four-person household. The differences between the CSO estimates and our results are rather small and may be attributed to such issues as the treatment of negative incomes or a model specification, as well as the fact that the CSO estimates are only for data from the 4th quarter of the year.

Table 2 compares equivalence scales implied by both models with the modified and original OECD equivalence scales for three selected household types.

Table 2. Equivalence scales implied by basic and extended models compared with OECD scales

	1 adult	2 adults	1 adult+1 child	2 adults+3 children
basic model	1.000	1.208	1.208	1.552
extended model	1.000	1.240	1.134	1.545
modified OECD	1.000	1.500	1.300	2.400
original OECD	1.000	1.700	1.500	3.200

Source: Own calculations. HBS 2010.

Notes*:* All equivalence scales are shown in relation to a one-adult household, where an adult is defined as a person aged 14 or older. In the case of subjective models, the equivalence scale is obtained by dividing the subjective poverty line of a household of certain type by a line of a reference household. For example, if we take as a reference a one-person household, then the equivalence scale for a two-person household will be equal to the ratio of subjective poverty lines of these two types of households. For an extended model for sample average values of education, town size and socio-economic group variables are taken. The original OECD scale (known also as the Oxford scale) assigns a value of 1 to the first household member, of 0.7 to each additional adult and of 0.5 to each child. The modified OECD scale assigns a value of 1 to the first household member, of 0.5 to each additional adult and of 0.3 to each child.

Both subjective scales are much flatter than the OECD which corresponds well to results in the literature (e.g. de Ree et al., 2013; Bollinger et al., 2012; Bishop et al, 2014). In other words, the objective scales underestimate economies of scale within the households relative to subjective perception of income situation. The smallest difference is visible between the basic model and the modified OECD scale for a "1+1" household (1.208 compared to 1.300). In other cases, the differences are high, especially for a couple with three children. The results of the basic and the extended model yield slightly different equivalence scale. The extended model suggests a higher "cost" of the second adult (1.24) than the basic model (1.21). Even more, the "cost" of the first child (1.13) in the extended model is lower than in all other specifications. It means that the extended model better accounts for the households' heterogeneity than the basic one.

Table 3 presents the results for the PHBS 2010 data in respect to a biological type of a household and by the approach to estimation of poverty. [4]

Table 3. Poverty incidence (headcount ratio) 2010 by household biological type

	Total	*1+0*	*1+1*	*2+0*	*2+1*	*2+2*	*2+3*	*2+4+*	*other w.ch.*	*other w/o ch.*
basic model	13.13	30.83	22.74	5.93	6.24	7.16	10.76	11.77	6.09	7.77
extended model	13.49	29.86	22.78	5.60	8.00	8.80	13.64	14.36	7.22	7.75
modified OECD	14.72	16.38	24.88	6.38	10.40	14.96	28.62	45.46	20.65	12.29
original OECD	15.67	9.93	27.35	5.97	11.76	20.23	38.10	61.32	28.52	13.75

Source: Author's calculations, HBS 2010.

Notes: HCR for the OECD scales calculated as 60% of the median equivalent income.

The lowest overall headcount ratio (HCR) occurs in the basic model and amounts to 13.13% of the households. The extended model yields only a slightly higher rate (13.49%). The objective poverty rates are higher, namely the headcount ratio calculated using the modified OECD scale is higher by about 1.2%, and using the original OECD – by 2.2%. The basic model yields the highest HCR (over 30%) for single households. The HCR for the extended model is slightly lower (almost 30%) but for the modified OECD scale the HCR is only about a half (16%) while for the original OECD scale – about one third (10%). An opposite conclusion may be drawn for larger households, e.g. for a couple with two children: the basic model yields HCR of 7.2%, the extended model – about 8.8%, while the traditional poverty lines lead to significantly higher rates: for the modified OECD scale it equals 15% and for the original OECD scale it is as much as 20%.

The results for the subjective models and these implied by the OECD scales are qualitatively different. The first approach suggests that a one-person household and single parents should be targeted by social policy. On the other hand, according to this approach large families are in a significantly better situation that the one postulated by the OCED. Different policy implications are also seen from the results presented in Table 4. It is visible that the basic model

[4] The relative poverty measures can differ due to differences in income distributions and in values of poverty lines when two different equivalence scales are applied. If we are interested only in the impact of the definition of the equivalence scale on the extent of relative poverty, then in both cases the same poverty line should be used. In this paper, following Bishop et al. (2013), we adopt a different approach and we allow for different poverty lines in each method.

classifies households quite similarly as the extended one. However, there are still almost 225 thousand households that are poor in the basic model but not in the extended one and 272 thousand – vice versa. Almost 1.5 million households are treated as poor in both models, thus the ratio of "classified differently" to "classified poor in both models" equals 1:3. In the case of the OECD scales, the differences are significantly larger.

Table 4. Cross-tabulation of households indicated as poor and non-poor, extended model compared with basic model and with the OECD-scales poverty (in thous. households)

	basic model		modified OECD scale		original OECD scale	
extended model:	non-poor	poor	non-poor	poor	non-poor	poor
non-poor	11 091	225	10 674	642	10 346	969
poor	272	1 492	481	1 283	684	1 080

Source: Author's calculations, HBS 2010.

Table 5 presents extra information on the differences in poverty classifications for the two approaches – the modified OECD scale and the extended subjective scale. As one may expect, the biggest differences are observed for the one-person households, for couples with 3 children – "2+3" – and couples with 4 or more children –"2+4+". There are about 440 thous. one-person households that are classified as being poor only when the subjective approach is applied. This accounts for as much as 13.5% of the total number of such units. For single parents the difference in classification results is small. There are 4.6% households that are classified as poor only for the OECD scale and about 2.5% for the subjective approach. Small differences are observed also for couples without a child and those with one or two dependent children. However, a small fraction of "2+2" households that are poor only for the OECD scale – 6.4%, is accompanied by a large absolute number of 93 thous. units.

The relative differences are large for "2+3", "2+4+" and "other household with child". Almost every third of households of parents with 4 and more children ("2+4+") is classified as poor only when the OECD scale is used. Respective fractions for "other household with child" and "couple with 3 children" are 13.7% and 15.0%. In terms of the absolute numbers, a group of "other household with child" is the largest one that is classified as poor only for the OECD approach. There are more than 250 thous. households that are not poor by the subjective standard but when the traditional approach is applied they are regarded as poor.

Table 5. Household classification in subjective approach (extended model) and expert

	Poor in expert approach								**Total**	
	No		Yes		No		Yes			
	Poor in subjective approach									
	No		No		Yes		Yes			
	%	No.	%	No.	%	No.	%	No.	%	No.
1+0	70.1	2 281.3	0.0	0.5	13.5	438.3	16.4	532.9	100.0	3 253.0
1+1	72.6	175.3	4.6	11.1	2.5	6.0	20.3	49.0	100.0	241.4
2+0	93.0	2 172.0	1.4	32.8	0.6	14.4	5.0	116.5	100.0	2 335.7
2+1	89.2	1 269.1	2.9	40.6	0.5	6.4	7.6	107.5	100.0	1 423.6
2+2	84.8	1 226.4	6.4	92.7	0.3	3.6	8.6	123.7	100.0	1 446.4
2+3	71.4	306.6	15.0	64.4	0.0	0.0	13.6	58.6	100.0	429.6
2+4+	54.5	91.0	31.1	51.9	0.0	0.0	14.4	23.9	100.0	166.8
oth w.ch.	79.1	1 469.4	13.7	254.2	0.2	3.0	7.1	131.0	100.0	1 857.6
oth w/o ch.	87.3	1 680.5	5.0	96.2	0.4	8.5	7.3	140.7	100.0	1 925.9
Total	81.6	10 671.6	4.9	644.2	3.7	480.3	9.8	1 283.9	100.0	13 080.0

Source: Author's calculations, HBS 2010 Notes: The category of "other household with child" includes "a couple with a child and other person" "single parent with a child with other person" and "other persons with a child." The category of "other household without child" is a residual one consisting of units not classified elsewhere.

In the Appendix the differences in deciles classifications are compared (Tables A2 a-c in Appendix). It shows that both approaches lead to different conclusions about the relative income situation not only for those who are at risk of poverty but also for those whose situation is relatively good. For example, 70% of one-person households are classified in the second decile by the OECD approach end up in the first decile if the subjective approach is used. An even more striking conclusion may be drawn for the middle part of distribution for the OECD scale. It is observed that 20% of those from the 5th decile are in the 2nd decile according to the alternative approach. Large movements are seen also for higher deciles. Generally, in the case of one-person households, the relative position of the household implied by the subjective approach is worse or at best the same as in the traditional approach. The opposite situation takes place when larger households are considered (Table A2b and Table A2c in Appendix). For instance, among the 2nd decile households with 3 children in the OECD-scale distribution, almost 60% of the households are classified in the 3rd decile and over

25% in the 4th decile when the subjective approach is applied. An even stronger divergence can be seen among multifamily units, where over 20% of households are in the 5th decile using the subjective approach, although they were classified in 2nd decile in the objective approach. The "migrations" from the above deciles seem fully consistent with our results concerning poverty rates within different household types.

3. Discussion on policy implications

The results presented above prove how complex and ambiguous the task to find an appropriate way of targeting social policy is. A seemingly simple question about monetary status of households turns out to be biased from the very beginning because we cannot reliably compare neither material needs nor socio- and psychological traits of different compositions of households. Despite this, the daily routine in policy-making is to take into account equivalised incomes - implicitly assuming the largely simplified OECD scales - without deeper investigation of the consequences of such an approach. Then, the results based on that simplified approach are used in deciding who should be the target group of social transfers. As we show in the paper, this group will be significantly different if we base the identification process on the subjective approach to equivalence scale. This raises the interesting question that is beyond the scope of the paper of whether we shall help people who find *themselves* poor or rather people who are *objectively* poor even if they do not consider themselves as such. Changing the current approach to the equivalence scale would mean that the whole wide range of social tools currently used should be assessed in order to verify who finally receives the transfers.

Our study suggests that at least two changes in social policy should have been considered if the subjective approach to equivalence scale had been taken seriously. First, persons living alone are the most overlooked social group with a much higher poverty risk than has been assumed so far. Simultaneously, we find larger households feeling much better about their current material situation than the objective poverty measures would imply. Joining these two facts together, it is a serious question whether the social budget should be distributed in a different way, so that a part of social tools should be terminated and perhaps a new tool proposed in its place. Second, equivalence scales are important in a discussion about tax and benefit regulations, since they have direct consequence on estimates of relative child costs. According to our results, the subjective equivalence scales suggest lower relative child cost than is embodied in the OECD modified scale. Also, the differences between subjective poverty rates and the expert rates are increasing in the number of children (Tab. 5). The fact that positional rankings of families with more children in the income distribution are better when the subjective scale is used means that we have found support for the conclusion in Bishop et al. (2013) about fixed costs of having children that are not accounted

for by the OECD scale. This has a clear policy implication since the fixed costs have to be taken into account in devising any fertility-enhancing programme.

Also, the subjective approach to equivalence scales can have even broader consequences for macroeconomic and regional policies in general, because it provides completely different income distribution across countries. An analysis of deciles migration between the presented approaches proves that there are substantial differences throughout the whole distribution and not only in its low end. As a consequence, all policy tools that include means-testing or in a different way take into account income of a household can bring a new light on the old issues.

4. Conclusion

Economic thinking on social policy is often based on very advanced models relying on the utility maximization principle and revealed preferences that, at least in theory, lead to complicated equivalence scales. On the other side, solutions used in practice are extremely simple and arguments based on poverty rates calculated with the OECD equivalence scale are often heard in public discussion. It seems that the simple practical solutions based on the OECD approach are located far away from the complex and logically consistent theoretical models.

We believe that a middle ground can be found and that subjective income evaluations give valuable information for public policy judgments, even though the possible measurement errors and the issue of comparing interpersonal satisfaction are involved while using such an approach. Accepting such imperfections does not seem to us to be a worse solution than applying the same three weights (1, 0.5, and 0.3) to all households.

This study used subjective information from surveys in order to compare the results with those based on the OECD. Being aware of the controversial nature of the method, we believe that subjective data can enrich our knowledge from the conventional approach, which may be valuable for policy evaluation. It turns out that although total poverty rates between those two approaches do not differ considerably, there are huge differences for specific sub-groups of households. We found out that the subjective equivalence scales are much flatter in the household size than the OECD ones, which corresponds well to results in the literature (e.g. Bishop et al., 2014; Podgórski, 1994). The range of economies of scale within the households postulated by the subjective approach is wider than the one from the OECD scales. This leads to policy suggestions different from those that are currently discussed. It follows that social groups that are most vulnerable to poverty are totally different in the two approaches. The official statistics based on either the Eurostat or CSO data point to large families as those who are in the relatively worst financial position. Thanks to the availability of the PHBS data, we have shown that more attention should be paid toward small households, and that the large ones are not in as bad situation as it is commonly

thought. In a country like Poland, where there is a relatively big share of large households and where income support policy uses income-testing heavily, such a conclusion might significantly change the allocation of public transfers.

Acknowledgement

This work was supported by the Polish National Science Centre (NCN) under Grant DEC-2013/09/B/HS4/01923.

REFERENCES

ATKINSON, A. B., (2017). Nierówność szans: co da się zrobić, Wydawnictwo Krytyki Politycznej.

BINDER, M., COAD, A., (2011). From Average Joe's happiness to Miserable Jane and Cheerful John: using quantile regression to analyze the full subjective well-being distribution, Journal of Economic Behavior & Organization, 79, 3, pp. 275–290.

BISHOP, J. A., GRODNER, A., LIU, H., AHAMDANECH-ZARCO, I., (2014). Subjective Poverty Equivalence Scales for Euro Zone Countries. The Journal of Economic Inequality, 12 (2), pp. 265–278.

BOLLINGER, C., NICOLETTI, C., PUDNEY, S., (2012). Two can live as cheaply as one…but three's a crowd, ISER Working Paper Series10/2012, Institute for Social and Economic Research, University of Essex.

BLUNDELL, R., LEWBEL, A., (1991). The Information Content of Equivalence Scales, Journal of Econometrics, 50, 3, pp. 49–68.

DE REE, J., ALESSIEZ, R., PRADHANX, M., (2013). The price and utility dependence of equivalence scales: Evidence from Indonesia, Journal of Public Economics, 97, pp. 272–281.

DE VOS, K., GARNER, T., (1991). An Evaluation of Subjective Poverty Definitions: Comparing Results from the U.S. and the Netherlands, Review of Income & Wealth, 37, 3, pp. 267–285.

CHIAPPORI, P.-A., (2016). Equivalence Versus Indifference Scales, Economic Journal, 126, pp. 523–545.

DUDEK, H., (2009). Statystyczna analiza subiektywnej oceny dochodów gospodarstw domowych rolników, Roczniki Nauku Rolniczych, Seria G, 96 (4), Szkoła Główna Gospodarstwa Wiejskiego, Warszawa.

DUDEK, H., (2012). Subiektywne skale ekwiwalentności - Analiza na podstawie danych o satysfakcji z osiąganych dochodów, Research Papers of Wrocław University of Economics Series 242, pp. 153–162.

DUDEK, H., LANDMESSER, J., (2012). Income satisfaction and relative deprivation, Statistics in Transition, 13, 2, pp. 321–334.

FERRER-I-CARBONELL, A., FRIJTERS, P., (2002). How important is Methodology for the Estimates of the Determinants of Happiness? Discussion Paper TI 2002-024/3, Tinbergen Institute, University of Amsterdam.

GUS, (2011). Metodologia badania budżetów gospodarstw domowych, GUS, Warszawa.

GUS, (2011a). Ubóstwo w Polsce w 2011 r. (na podstawie badań budżetów gospodarstw domowych, GUS), Warszawa.

IACOVOU, M., SKEW, A., (2010). Household structure in the EU, Luxemburg: European Commission.

KALBARCZYK, M., MIŚTA, R., MORAWSKI, L., Subjective equivalence scales – cross-country and time differences, International Journal of Social Economics, forthcoming

KAPTEYN, A., VAN HERWAARDEN, F., (1981). Empirical comparison of the shape of welfare functions, European Economic Review 15, 3, pp. 261–286.

KAPTEYN, A., VAN PRAAG, B., (1976). A new approach to the construction of family equivalence scales, European Economic Review, 7, 4, pp. 313–335.

LEWBEL, A., PENDAKUR, K., (2008). Equivalnence scales. In S. Durlauf and L. Blume (ed.), The New Palgrave Dictionary of Economics, Palgrave Macmillan.

PANEK, T., CZAPIŃSKI, J., (2015). Wykluczenie społeczne, Contemporary Economics, Vol. 9, 4, pp. 396–432.

PODGÓRSKI, J., (1990). Zastosowanie metody "Leyden Poverty Line" w warunkach Polski, Wiadomości Statystyczne, 11, pp. 5–9.

PODGÓRSKI, J., (1991). Subiektywne linie ubóstwa - nowe wyniki, Wiadomości Statystyczne, 11, pp. 6–12.

PODGÓRSKI, J., (1994). Metody wyznaczania subiektywnych linii ubóstwa. Wiadomości Statystyczne, 12, pp. 12–19.

POLLAK, R., WALES, T., (1979). Welfare comparisons and equivalence scales, American Economic Review, 69, 2, pp. 216–221.

SCHOKKAERT, E., VAN OOTEGEMY, L., VERHOF, E., (2011). Preferences and Subjective Satisfaction: Measuring Well-being on the Job for Policy Evaluation, CESifo Economic Studies, 57, 4, pp. 683–714.

SEIDEL, B., (1994). How sensible is the Leyden individual welfare function of income?, European Economic Review, 38, 8, pp. 1633–1659.

STEVENSON, B., WOLFERS, J., (2008). Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox, Brookings Papers on Economic Activity (Economic Studies Program), 39, 1, pp. 1–102.

SZULC, A., (2014). Empirical versus policy equivalence scales: matching estimation, Bank i Kredyt, 45, 1, pp. 37–52

VAN DEN BOSCH, K., (2002). Identifying the Poor: Using subjective and consensual measures, Ashgate Publishing Limited.

VAN PRAAG, B., (1971). The welfare function of income in Belgium: An empirical investigation, European Economic Review, 11, 3, pp. 337–369.

VAN PRAAG, B., FERRER-I-CARBONELL, A., (2004). Happiness Quantified, Oxford: Oxford University Press.

VAN PRAAG, B., KAPTEYN, A., (1994). How sensible is the Leyden individual welfare function of income? A Reply, European Economic Review 38, 9, pp. 1817–1825.

VAN PRAAG, B., VAN DER SAR, N., (1988). Household Cost Functions and Equivalance Scales, Journal of Human Resources, 23, 2, pp. 193–210.

VEENHOVEN, R., (2002). Why Social Policy Needs Subjective Indicators? Social Indicators Research, 58, 1, pp. 33–45.

APPENDIX

Table A1. Sample characteristics of Household Budget Survey 2010

	Data set size	
	Sample frequency	Population frequency
No. of households	37 412	13 332 320
No. of persons	107 967	37 726 497

	Household size			
	Sample frequency	Sample percentage	Population frequency	Population percentage
1	6 700	17.91	3 307 035	24.80
2	11 087	29.63	3 097 050	23.23
3	7 838	20.95	2 653 892	19.91
4	6 737	18.01	2 405 045	18.04
5	3 003	8.03	1 085 993	8.15
6+	2 047	5.47	788 003	5.91

Source: Own calculations on HBS 2010.

Table A2a. One-person households (%)

		Subjective approach deciles										
		1	2	3	4	5	6	7	8	9	10	Total
OECD approach deciles	1	**100**	0	0	0	0	0	0	0	0	0	10.67
	2	70.17	**29.72**	0.1	0	0	0	0	0	0	0	12.44
	3	15.4	73.71	**10.89**	0	0	0	0	0	0	0	12.84
	4	6.22	38.56	50.81	**4.28**	0.12	0	0	0	0	0	11.99
	5	0.87	20.57	44.03	31.84	**2.58**	0.12	0	0	0	0	10.83
	6	0	7.34	22.53	44.11	24.62	**1.4**	0	0	0	0	9.35
	7	0	1.55	12.44	29.61	36.99	17.03	**2.39**	0	0	0	8.24
	8	0	0	2.93	16.64	25.54	30.97	22.22	**1.51**	0.18	0	7.06
	9	0	0	0	1.8	12.74	23.98	30.08	25.58	**5.82**	0	7.96
	10	0	0	0	0	0	1.65	10.14	21.11	31.63	**35.47**	8.61
	Total	22.22	20.82	15.61	11.84	8.46	5.79	5.04	3.96	3.2	3.06	3 250 550

Source: Own calculations on HBS 2010.

Table A2b. Parents with three children (%)

OECD approach deciles \ Subjective approach deciles	1	2	3	4	5	6	7	8	9	10	Total
1	**45.39**	44.56	7.83	2.23	0	0	0	0	0	0	20.72
2	0	**11.56**	57.12	26.21	5.1	0	0	0	0	0	14.84
3	0	0	**8.73**	38.78	39.66	11.42	1.41	0	0	0	13.66
4	0	0	0	**3.17**	40.32	45.4	9.26	1.86	0	0	10.89
5	0	0	0	2.39	**6.85**	36.89	42.99	10.87	0	0	9.71
6	0	0	0	0	8.17	**9.69**	29.81	46.53	5.81	0	8.42
7	0	0	0	0	0	1.65	**23.71**	48.74	24.75	1.14	6.45
8	0	0	0	0	0	0	6.07	**30.27**	52.25	11.4	7.45
9	0	0	0	0	0	0	0	2.55	**32.67**	64.78	4.49
10	0	0	0	0	0	0	0	0	6.86	**93.14**	3.37
Total	9.41	10.95	11.29	10.23	11.92	11.01	9.86	10.69	7.68	6.97	429 133

Source: Own calculations on HBS 2010.

Table A2c. Other households with a child (%)

OECD approach deciles \ Subjective approach deciles	1	2	3	4	5	6	7	8	9	10	Total
1	**35.56**	30.84	22.66	8.28	2.14	0.52	0	0	0	0	14.01
2	1	**8.72**	22.53	34.2	22.56	8.74	2.05	0.2	0	0	13.90
3	0	2.74	**7.52**	21.09	33.67	22.68	10.34	1.8	0.17	0	12.05
4	0	0.15	2.36	**9.8**	17.18	31.88	24.22	12.49	1.92	0	11.32
5	0	0	0.19	5.08	**7.9**	20.83	33.7	25.52	6.29	0.49	11.32
6	0	0	0	0.6	3.08	**10.31**	20.05	40.65	24.24	1.06	10.12
7	0	0	0	0.16	1.43	4.45	**12.93**	24.52	46.89	9.61	9.49
8	0	0	0	0	0	0.9	3.32	**12.34**	49.35	34.09	7.84
9	0	0	0	0	0	0	0.51	3.24	**24.1**	72.15	6.41
10	0	0	0	0	0	0	0	0	1.96	**98.04**	3.53
Total	5.12	5.88	7.5	10.22	10.78	11.52	11.64	12.16	13.34	11.84	1 855 766

Source: Own calculations on HBS 2010.

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 521–540, DOI 10. 21307

SELECTING THE OPTIMAL MULTIDIMENSIONAL SCALING PROCEDURE FOR METRIC DATA WITH R ENVIRONMENT

Marek Walesiak[1], **Andrzej Dudek**[2]

ABSTRACT

In multidimensional scaling (MDS) carried out on the basis of a metric data matrix (interval, ratio), the main decision problems relate to the selection of the method of normalization of the values of the variables, the selection of distance measure and the selection of MDS model. The article proposes a solution that allows choosing the optimal multidimensional scaling procedure according to the normalization methods, distance measures and MDS model applied. The study includes 18 normalization methods, 5 distance measures and 3 types of MDS models (ratio, interval and spline). It uses two criteria for selecting the optimal multidimensional scaling procedure: Kruskal's *Stress*-1 fit measure and Hirschman-Herfindahl *HHI* index calculated based on Stress per point values. The results are illustrated by an empirical example.

Key words: multidimensional scaling, normalization of variables, distance measures, *HHI* index, R program.

1. Introduction

Multidimensional scaling is a method that represents (dis)similarity data as distances in a low-dimensional space (typically 2 or 3 dimensional) in order to make these data accessible to visual inspection and exploration (Borg, Groenen, 2005, p. 3). The dimensions are not directly observable. They have the nature of latent variables. MDS allows the similarities and differences between the analyzed objects to be explained.

Multidimensional scaling is a widely used technique in many areas, including psychology (Takane, 2007), sociology (Pinkley, Gelfand, Duan, 2005), linguistics

[1] Wroclaw University of Economics, Department of Econometrics and Computer Science, Jelenia Góra. The project is financed by the Polish National Science Centre, decision DEC-2015/17/B/HS4/00905. E-mail: marek.walesiak@ue.wroc.pl.
[2] Wroclaw University of Economics, Department of Econometrics and Computer Science, Jelenia Góra. E-mail: andrzej.dudek@ue.wroc.pl.

(Embleton, Uritescu, Wheeler, 2013), marketing research (Cooper, 1983), tourism (Marcussen, 2014) and geography (Golledge, Ruhton, 1972).

The starting point of multidimensional scaling is a distance matrix (dissimilarities) between objects in *m*-dimensional space $\boldsymbol{\delta}=[\delta_{ik}]$, where $i,k=1,...,n$ is the number of the object. Methods of determining the distance matrix $\boldsymbol{\delta}=[\delta_{ik}]$ can be divided into direct (typically result from similarity ratings on object pairs, from rankings, or from card-sorting tasks) and indirect (they can be derived from other data) methods (see, e.g. Borg, Groenen, 2005, pp. 111-133).

The article uses an indirect method in which the starting point is a metric data matrix $\mathbf{X}=[x_{ij}]$ (x_{ij} – the value of the *j*-th variable for the *i*-th object, $j=1,...,m$ – the number of metric variable), for which observations are obtained from secondary data sources. It is a typical situation in socio-economic research.

The normalization of variables is carried out when the variables describing the analyzed objects are measured on metric scales (interval or ratio). The characteristics of measurement scales were discussed, e.g. in the study by (Stevens, 1946). The purpose of normalization is to achieve the comparability of variables.

Metric data that requires normalization of variables complicates the problem of choosing a multidimensional scaling procedure. The article proposes a solution that allows the choice of the optimal multidimensional scaling procedure, carried out on the basis of metric data (interval, ratio), according to the normalization methods, distance measures and MDS model applied. The study included 18 normalization methods, 5 distance measures and MDS models (ratio, interval and spline – e.g. polynomial function of second or third degree). For instance, ten normalization methods, five distance measures and four MDS models give 200 multidimensional scaling procedures.

The authors of the monograph (Borg, Groenen, Mair, 2013, chapter 7) pointed out the typical mistakes made by users of multidimensional scaling. A frequent mistake on the part of users of MDS results is to evaluate Stress mechanically (rejecting an MDS solution because its Stress seems "too high"). In their opinion (Borg, Groenen, Mair, 2013, p. 68) "An MDS solution can be robust and replicable, even if its Stress value is high" and "Stress, moreover, is a *summative* index for *all* proximities. It does not inform the user how well a *particular* proximity value is represented in the given MDS space". In addition, we should take into account Stress per point measure (the average of the squared error terms for each point) and acceptability of MDS results (based on "Shepard diagram").

To solve the problem of choosing the optimal multidimensional scaling procedure, two criteria were applied: Kruskal's *Stress*-1 (*Stress* – Standardized residual sum of squares) fit measure and the Hirschman-Herfindahl *HHI* index, calculated based on Stress per point values (*spp*). The article proposes an

algorithm that allows the selection of the optimal multidimensional scaling procedure with implementation in `mdsOpt` package of R program (Walesiak, Dudek, 2017b).

The results are illustrated by an empirical example.

2. Multidimensional scaling based on metric data

A general scheme of multidimensional scaling performed on metric data is as follows:

$$P \rightarrow A \rightarrow X \rightarrow \mathbf{X} \rightarrow \mathbf{Z} \rightarrow \boldsymbol{\delta} \rightarrow S \rightarrow \mathbf{d} \rightarrow \mathbf{V} \rightarrow I\,, \tag{1}$$

where:

P – choice of research problem,

A – selection of objects,

X – selection of variables,

$\mathbf{X}$ – collecting data and construction of data matrix $\mathbf{X}=[x_{ij}]_{nxm}$ for $i,k=1,...,n$ and $j=1,...,m$ (x_{ij} – the value of the *j*-th variable for the *i*-th object),

$\mathbf{Z}$ – choice of variable normalization method and construction of normalized data matrix $\mathbf{Z}=[z_{ij}]_{nxm}$ for $i,k=1,...,n$ and $j=1,...,m$ (z_{ij} – the normalized value of the *j*-th variable for the *i*-th object),

$\boldsymbol{\delta}$ – selection of distance measure (see Table 3) and construction of distance matrix in *m*-dimensional space $\boldsymbol{\delta}=\left[\delta_{ik}(\mathbf{Z})\right]_{nxn}$ for $i,k=1,...,n$,

S – perform multidimensional scaling (MDS): $f:\delta_{ik}(\mathbf{Z})\rightarrow d_{ik}(\mathbf{V})$ for all pairs (i,k) – mapping distances in *m*-dimensional space $\delta_{ik}(\mathbf{Z})$ into corresponding distances $d_{ik}(\mathbf{V})$ in *q*-dimensional space ($q<m$) by a representation function *f*. The distances $d_{ik}(\mathbf{V})$ are always unknown, i.e. MDS must find a configuration $\mathbf{V}$ of predetermined dimensions *q* on which the distances are computed,

$\mathbf{d}$ – Euclidean distance matrix in *q*-dimensional space ($q<m$, typically *q* equals 2 or 3) $\mathbf{d}=\left[d_{ik}(\mathbf{V})\right]_{nxn}$ for $i,k=1,...,n$,

$\mathbf{V}$ – configuration of objects in *q*-dimensional space $\mathbf{V}=[v_{ij}]_{nxq}$,

I – interpretation of multidimensional scaling results in *q*-dimensional space.

In SMACOF (Scaling by Majorizing a Complicated Function) algorithm we minimize Stress (2) over the configuration matrix $\mathbf{V}$ by an iterative procedure (see Borg, Groenen, 2005, pp. 204-205):

1. Set $\mathbf{V} = \mathbf{V}^{[0]}$, where $\mathbf{V}^{[0]}$ is some nonrandom or random start configuration. Starting solution is usually Torgerson-Gower classical scaling (Torgerson, 1952; Gower, 1966). Set iteration counter $k = 0$. Set ε to a small positive constant (convergence criterion), i.e. $\varepsilon = 0.000001$.
2. Find optimal disparities $\hat{d}_{ik}$ for fixed distances $d_{ik}(\mathbf{V}^{[0]})$.
3. Standardize (to avoid degenerated solution) $\hat{d}_{ik}$ so that $\eta^2_{\hat{d}} = n(n-1)/2$.
4. Compute Stress function $\sigma_r^{[0]} = \sigma_r(\hat{\mathbf{d}}, \mathbf{V}^{[0]})$:

$$\begin{aligned}\sigma_r(\hat{\mathbf{d}}, \mathbf{V}) &= \sum_{i<k} w_{ik}(d_{ik}(\mathbf{V}) - \hat{d}_{ik})^2 \\ &= \sum_{i<k} w_{ik}\hat{d}^2_{ik} + \sum_{i<k} w_{ik} d^2_{ik}(\mathbf{V}) - 2\sum_{i<k} w_{ik}\hat{d}_{ik} d_{ik}(\mathbf{V}) \\ &= \eta^2_{\hat{d}} + \eta^2(\mathbf{V}) - 2\rho(\hat{\mathbf{d}}, \mathbf{V}). \end{aligned} \quad (2)$$

where: $\hat{d}_{ik}$ – d-hats, disparities, target distances or pseudo distances (see Borg, Groenen 2005, p. 199). $\hat{d}_{ik} = f(\delta_{ik})$ by defining f in different ways: $\hat{d}_{ik} = b \cdot \delta_{ik}$ – ratio MDS; $\hat{d}_{ik} = a + b \cdot \delta_{ik}$ – interval MDS, $\hat{d}_{ik} = a + b \cdot \delta_{ik} + c \cdot \delta^2_{ik}$ – spline MDS (polynomial function of second degree);

$w_{ik} = 1$ – for object pair i,k a dissimilarity has been observed, $w_{ik} = 0$ – otherwise.

Set $\sigma_r^{[-1]} = \sigma_r^{[0]}$.

5. While $k = 0$ or ($\sigma_r^{[k-1]} - \sigma_r^{[k]} > \varepsilon$ and $k \leq$ maximum iterations) do
6. Increase iteration number *k* by one ($k := k + 1$).
7. Compute Guttman transform $\mathbf{V}^{[k]}$ (see Borg, Groenen, 2005, p. 191; De Leeuw, Mair, 2009, p. 5).
8. Find optimal disparities $\hat{d}_{ik}$ for fixed distances $d_{ik}(\mathbf{V}^{[k]})$.
9. Standardize $\hat{d}_{ik}$ so that $\eta^2_{\hat{d}} = n(n-1)/2$.
10. Compute $\sigma_r^{[k]} = \sigma_r(\hat{\mathbf{d}}, \mathbf{V}^{[k]})$.
11. Set $\mathbf{V} = \mathbf{V}^{[k]}$,
12. End while.

A flowchart of the SMACOF algorithm is given in Figure 1.

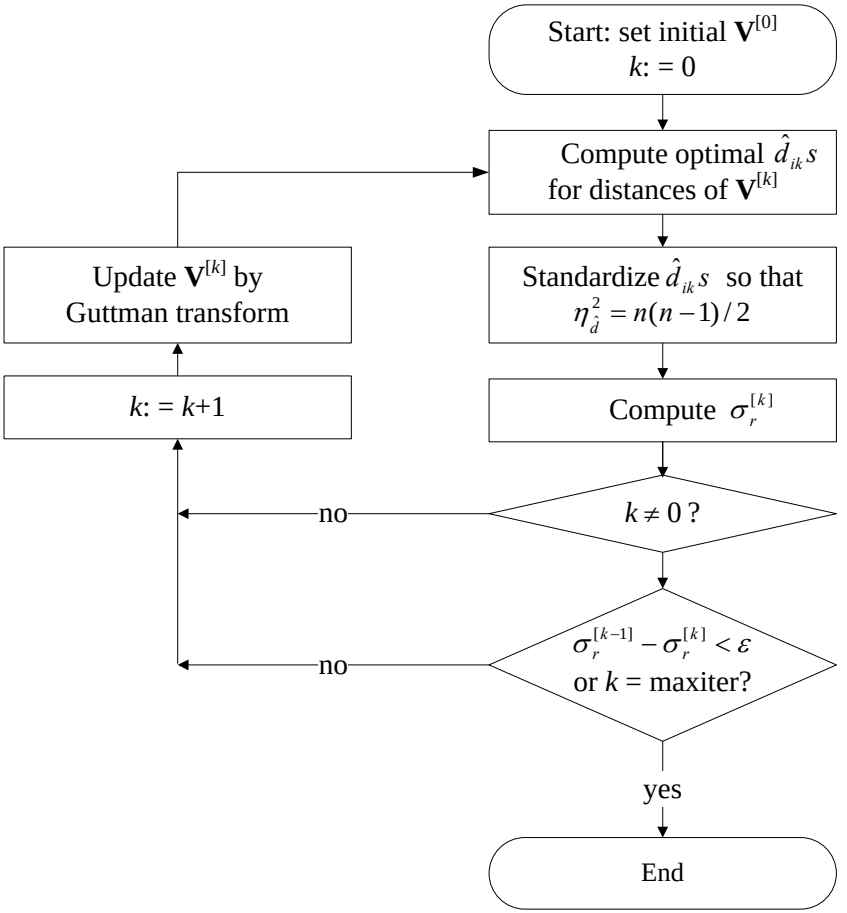


Figure 1. The flowchart of the majorization algorithm (SMACOF)
Source: Borg, Groenen, 2005, p. 205.

In other multidimensional scaling algorithms, different fit measures are applied (see, e.g. Borg, Groenen, 2005, pp. 250-254): Kruskal's *Stress*-1, Kruskal and Carroll *Stress*-2, the Guttman-Lingoes coefficient of alienation, *S-Stress* of Takane, Young and De Leeuw.

3. Criteria for the selection of the optimal multidimensional scaling procedure

The article proposes a solution that allows the optimal multidimensional scaling procedure to be chosen. The study uses the function `smacofSym` of

`smacof` package od R program (R Development Core Team, 2017). In the function `smacofSym` of `smacof` package (Mair et al., 2017) basic decision problems involve the following selection:

- normalization method (the analysis included 18 normalization methods),
- distance measure (the analysis included 5 distance measures),
- MDS model (the analysis included: ratio MDS, interval MDS, spline MDS).

Table 1 presents normalization methods, given by linear formula (3), which were used in the selection of the optimal MDS procedure (see Jajuga, Walesiak, 2000, pp. 106-107; Zeliaś, 2002, p. 792):

$$z_{ij} = b_j x_{ij} + a_j = \frac{x_{ij} - A_j}{B_j} = \frac{1}{B_j} x_{ij} - \frac{A_j}{B_j} \quad (b_j > 0), \tag{3}$$

where: x_{ij} – the value of *j*-th variable for the *i*-th object,

z_{ij} – the normalized value of *j*-th variable for the *i*-th object,

A_j – shift parameter to arbitrary zero for the *j*-th variable,

B_j – scale parameter for the *j*-th variable,

$a_j = -A_j / B_j$, $b_j = 1/B_j$ – parameters for the *j*-th variable presented in Table 1.

Table 1. Normalization methods

Type	Method	Parameter		Scale of variables	
		b_j	a_j	BN	AN
n1	Standardization	$1/s_j$	$-\bar{x}_j / s_j$	ratio or interval	interval
n2	Positional standardization	$1/mad_j$	$-med_j / mad_j$	ratio or interval	interval
n3	Unitization	$1/r_j$	$-\bar{x}_j / r_j$	ratio or interval	interval
n3a	Positional unitization	$1/r_j$	$-med_j / r_j$	ratio or interval	interval
n4	Unitization with zero minimum	$1/r_j$	$-\min_i\{x_{ij}\} / r_j$	ratio or interval	interval
n5	Normalization in range [–1; 1]	$\frac{1}{\max_i \lvert x_{ij} - \bar{x}_j \rvert}$	$\frac{-\bar{x}_j}{\max_i \lvert x_{ij} - \bar{x}_j \rvert}$	ratio or interval	interval
n5a	Positional normalization in range [–1; 1]	$\frac{1}{\max_i \lvert x_{ij} - med_j \rvert}$	$\frac{-med_j}{\max_i \lvert x_{ij} - med_j \rvert}$	ratio or interval	interval

Table 1. Normalization methods (cont.)

Type	Method	Parameter		Scale of variables	
		b_j	a_j	BN	AN
n6	Quotient transformations	$1/s_j$	0	ratio	ratio
n6a		$1/mad_j$	0	ratio	ratio
n7		$1/r_j$	0	ratio	ratio
n8		$1/\max_i\{x_{ij}\}$	0	ratio	ratio
n9		$1/\bar{x}_j$	0	ratio	ratio
n9a		$1/med_j$	0	ratio	ratio
n10		$1/\sum_{i=1}^{n} x_{ij}$	0	ratio	ratio
n11		$1/\sqrt{\sum_{i=1}^{n} x_{ij}^2}$	0	ratio	ratio
n12	Normalization	$\frac{1}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$	$\frac{-\bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$	ratio or interval	interval
n12a	Positional normalization	$\frac{1}{\sqrt{\sum_{i=1}^{n}(x_{ij}-med_j)^2}}$	$\frac{-med_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-med_j)^2}}$	ratio or interval	interval
n13	Normalization with zero being the central point	$\frac{1}{r_j/2}$	$-\frac{m_j}{r_j/2}$	ratio or interval	interval

BN – before normalization, AN – after normalization, $\bar{x}_j$ – mean for the *j*-th variable, s_j – standard deviation for the *j*-th variable, r_j – range for the *j*-th variable, $m_j = \frac{\max_i\{x_{ij}\} + \min_i\{x_{ij}\}}{2}$ – mid-range for the *j*-th variable, $med_j = med_i(x_{ij})$ – median for the *j*-th variable, $mad_j = mad_i(x_{ij})$ – median absolute deviation for the *j*-th variable.

Source: Based on (Jajuga, Walesiak, 2000; Walesiak, Dudek, 2017a).

Column 1 in Table 1 presents the type of normalization method adopted as the function `data.Normalization` of `clusterSim` package (Walesiak, Dudek, 2017a). Similar procedure for data normalization is available as the function `scale` of `base` package. In this function the researcher defines the parameters A_j and B_j.

Due to the fact that the groups of A, B, C and D (see Table 2) normalization methods give identical multidimensional scaling results, further analysis covers

the first methods of the identified groups (n1, n2, n3, n9), as well as the other methods (n5, n5a, n8, n9a, n11, n12a).

Table 2. The groups of normalization methods resulting in identical distance matrices

Groups of normalization methods	Normalization methods	
	GDM1 distance	Minkowski distances, squared Euclidean distance*
A	n1, n6, n12	n1, n6, n12
B	n2, n6a	n2, n6a
C	n3, n3a, n4, n7, n13	n3, n3a, n4, n7, n13
D	n9, n10	n9, n10

* after dividing distances in each distance matrix by the maximum value.

Source: Own presentation.

Table 3 presents selected distance measures for metric data that have been used in the selection of the optimal multidimensional scaling procedure.

Distance GDM1 is available as a function of `dist.GDM` of `clusterSim` package (Walesiak, Dudek, 2017) and the remaining distances in Table 3 are available in the function `dist` of `stats` package (R Development Core Team, 2017).

The initial point of the application of `smacofSym` function is to determine the following values of arguments:

– convergence criterion (eps=1e-06),

– maximum number of iterations (itmax=1000).

These parameters can be changed by the user.

The selection of the optimal procedure for multidimensional scaling takes place in several stages:

1. Set the number of dimensions in MDS to two (ndim=2).
2. Taking into account in the analysis 10 normalization methods, 5 distance measures and 2 MDS models, there are 100 multidimensional scaling procedures. Multidimensional scaling is performed for each procedure separately. It then orders the procedures by increasing *Stress*-1 fit measure (see e.g. Borg, Groenen, Mair, 2013, p. 23):

$$Stress\text{-}1_p = \sqrt{\sum_{i<k}[d_{ik}(\mathbf{V}) - \hat{d}_{ik}]^2 \Big/ \sum_{i<k} d_{ik}^2(\mathbf{V})}\,, \tag{4}$$

where: $p = 1,...,100$ – multidimensional scaling procedure number.

Table 3. Distance measures for metric (interval, ratio) data

Name	Distance δ_{ik}	Range	Allowed normalization
Minkowski ($p \geq 1$)	$\sqrt[p]{\sum_{j=1}^{m} \left\| z_{ij} - z_{kj} \right\|^{p}}$	$[0;\infty)$	n1-n13
Manhattan ($p = 1$)	$\sum_{j=1}^{m} \left\| z_{ij} - z_{kj} \right\|$	$[0;\infty)$	n1-n13
Euclidean ($p = 2$)	$\sqrt{\sum_{j=1}^{m} \left(z_{ij} - z_{kj} \right)^{2}}$	$[0;\infty)$	n1-n13
Chebyshev (maximum) ($p \to \infty$)	$\max_{j} \left\| z_{ij} - z_{kj} \right\|$	$[0;\infty)$	n1-n13
Squared Euclidean	$\sum_{j=1}^{m} \left(z_{ij} - z_{kj} \right)^{2}$	$[0;\infty)$	n1-n13
GDM1	$\frac{1}{2} - \frac{\sum_{j=1}^{m} (z_{ij} - z_{kj})(z_{kj} - z_{ij}) + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} (z_{ij} - z_{lj})(z_{kj} - z_{lj})}{2 \left[\sum_{j=1}^{m} \sum_{l=1}^{n} (z_{ij} - z_{lj})^{2} \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} (z_{kj} - z_{lj})^{2} \right]^{\frac{1}{2}}}$	[0;1]	n1-n13

$i, k, l = 1, \ldots, n$ – object number, m – the number of objects, $j = 1, \ldots, m$ – variable number, m – the number of variables, $z_{ij}(z_{kj}, z_{lj})$ – the normalized value of the j-th variable for the i-th (k-th, l-th) object.

Source: Based on (Everitt et al., 2011, pp. 49-50; Jajuga, Walesiak, Bąk, 2003).

3. Based on Stress per point (*spp*) values (Stress contribution in percentages), the Hirschman-Herfindahl index is calculated (Herfindahl, 1950; Hirschman, 1964):

$$HHI_p = \sum_{i=1}^{n} spp_{pi}^{2}, \tag{5}$$

where: $i = 1, \ldots, n$ – object number.

The HHI_p index takes values in the interval $\left[\frac{10{,}000}{n}; 10{,}000 \right]$. The value $\frac{10{,}000}{n}$ means that the distribution of errors for individual objects is uniform ($\forall_i spp_i = \frac{100}{n}$). The maximal value appears when summary fit measure (*Stress*-1) is the result of loss assigned only to one object. For other objects, loss function will be equal to zero. The optimal situation for a multidimensional scaling procedure is the minimal value of the HHI_p index.

4. The chart with $Stress\text{-}1_p$ fit measure value on *x*-axis and HHI_p index on *y*-axis for *p* procedures of multidimensional scaling is drawn.

5. The maximal acceptable value of $Stress\text{-}1$ is assumed as s. For all multidimensional scaling procedures for which $Stress\text{-}1_p \leq s$, we chose the one for which $\min_p \{HHI_p\}$ occurs.
6. Multidimensional scaling for the selected procedure is performed along with checkout that in the sense of interpretation results are acceptable. Based on the Shepard diagram, the correctness of the model scaling will be evaluated. If the results are acceptable the procedure ends, otherwise it returns to step 1 and multidimensional scaling for three dimensions is performed (ndim=3).

4. Empirical results

The empirical study uses the statistical data presented in the article (Gryszel, Walesiak, 2014) and referring to the attractiveness level of 29 Lower Silesian counties. The evaluation of tourist attractiveness of Lower Silesian counties was performed using 16 metric variables (measured on a ratio scale):

x1 – beds in hotels per 1 km^2 of a county area,
x2 – number of nights spent daily by resident tourists (Poles) per 1,000 inhabitants of a county,
x3 – number of nights spent daily by foreign tourists per 1,000 inhabitants of a county,
x4 – gas pollution emission in tons per 1 km^2 of a county area,
x5 – number of criminal offences and crimes against life and health per 1,000 inhabitants of a county,
x6 – number of property crimes per 1,000 inhabitants of a county,
x7 – number of historical buildings per 100 km^2 of a county area,
x8 – % of a county forest cover,
x9 – % share of legally protected areas within a county area,
x10 – number of events as well as cultural and tourist ventures in a county,
x11 – number of natural monuments calculated per 1 km^2 of a county area,
x12 – number of tourist economy entities per 1,000 inhabitants of a county (natural and legal persons),
x13 – expenditure of municipalities and counties on tourism, culture and national heritage protection as well as physical culture per 1 inhabitant of a county in Polish zlotys (PLN),
x14 – cinema attendance per 1,000 inhabitants of a county,
x15 – museum visitors per 1,000 inhabitants of a county,
x16 – number of construction permits (hotels and accommodation buildings, commercial and service buildings, transport and communication buildings, civil and water engineering constructions) issued in a county in the years 2011-2012, per 1 km^2 of a county area.

The statistical data were collected in 2012 and come from the Local Data Bank of the Central Statistical Office of Poland; the data for x7 variable only were obtained from the regional conservation officer.

Variables (x4, x5 and x6) take the form of destimulants, x9 is a nominant (50% level was adopted as the optimal one). The other variables represent stimulants, whereas x9 nominant was transformed into a stimulant. The definitions of stimulants, destimulants and nominants are available in the study, e.g. (Walesiak, 2016).

A pattern object and an anti-pattern object were added to the set of 29 counties (see Walesiak, 2016). Therefore, the data matrix covers 31 objects described by 16 variables. The coordinates of a pattern object cover the most preferred preference variable (stimulants, destimulants and nominants) values. The coordinates of an anti-pattern object cover the least preferred preference variable values.

The article uses its own script of package `mdsOpt` of R program (Walesiak, Dudek, 2017b) to choose the optimal procedure for multidimensional scaling due to normalization methods, selected distance measures and MDS models (developed in accordance with the methodology described in section 3).

The measurement of variables on a ratio scale accepts all normalization methods (hence the study covered 18 methods). Due to the fact that the groups of A, B, C and D normalization methods give identical multidimensional scaling results (see Table 2), further analysis covers the first methods of the identified groups (n1, n2, n3, n9), as well as the other methods (n5, n5a, n8, n9a, n11, n12a).

Ordering results of 100 multidimensional scaling procedures (10 normalization methods x 5 distance measures x 2 MDS models) according to formula (4) are presented in Table 4. In addition, Table 4 shows values of HHI_p index for each MDS procedure.

Table 4. Ordering results of 100 multidimensional scaling procedures

p	nm	MDS model	Distance measure	*Stress*-1	*HHI*	*p*	nm	MDS model	Distance measure	*Stress*-1	*HHI*
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
1	n9a	interval	euclidean	0.0311	844	51	n2	ratio	seuclidean	0.1391	1328
2	n2	interval	euclidean	0.0369	685	52	n11	ratio	GDM1	0.1391	495
3	n9a	ratio	euclidean	0.0404	715	53	n5a	interval	seuclidean	0.1400	663
4	n9a	interval	maximum	0.0408	1276	54	n5	ratio	seuclidean	0.1402	797
5	n9a	ratio	maximum	0.0441	1230	55	n5a	interval	euclidean	0.1405	508
6	n2	interval	maximum	0.0505	908	56	n11	ratio	manhattan	0.1414	453
7	n2	ratio	euclidean	0.0546	520	57	n5a	ratio	seuclidean	0.1436	791
8	n2	ratio	maximum	0.0576	794	58	n9	ratio	euclidean	0.1473	464
9	n9a	interval	manhattan	0.0627	867	59	n9a	ratio	seuclidean	0.1478	1289
10	n9a	ratio	manhattan	0.0687	645	60	n8	ratio	manhattan	0.1483	428
11	n2	interval	manhattan	0.0704	755	61	n3	ratio	manhattan	0.1502	419
12	n2	interval	GDM1	0.0770	605	62	n1	ratio	manhattan	0.1530	410
13	n9a	interval	GDM1	0.0793	593	63	n5	ratio	manhattan	0.1531	421

Table 4. Ordering results of 100 multidimensional scaling procedures (cont.)

p	nm	MDS model	Distance measure	*Stress*-1	*HHI*	*p*	nm	MDS model	Distance measure	*Stress*-1	*HHI*
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
14	n2	ratio	manhattan	0.0839	521	64	n12a	ratio	manhattan	0.1543	409
15	n2	ratio	GDM1	0.0894	887	65	n5a	ratio	manhattan	0.1548	422
16	n9a	ratio	GDM1	0.0969	924	66	n8	interval	GDM1	0.1598	486
17	n9	interval	manhattan	0.0985	577	67	n8	ratio	GDM1	0.1608	489
18	n9	interval	euclidean	0.1056	580	68	n9	interval	maximum	0.1610	554
19	n9	interval	seuclidean	0.1087	813	69	n3	interval	GDM1	0.1640	473
20	n11	interval	manhattan	0.1092	500	70	n3	ratio	GDM1	0.1653	476
21	n8	interval	manhattan	0.1149	476	71	n1	interval	GDM1	0.1677	431
22	n11	interval	seuclidean	0.1149	739	72	n1	ratio	GDM1	0.1691	435
23	n3	interval	manhattan	0.1155	469	73	n11	ratio	euclidean	0.1698	427
24	n2	interval	seuclidean	0.1161	865	74	n12a	interval	GDM1	0.1718	430
25	n9	ratio	seuclidean	0.1164	1102	75	n12a	ratio	GDM1	0.1732	434
26	n9	interval	GDM1	0.1166	545	76	n5	interval	GDM1	0.1737	494
27	n9	ratio	GDM1	0.1166	545	77	n5	ratio	GDM1	0.1738	494
28	n11	interval	euclidean	0.1168	497	78	n5a	interval	GDM1	0.1774	493
29	n11	ratio	seuclidean	0.1179	922	79	n5a	ratio	GDM1	0.1774	493
30	n1	interval	manhattan	0.1186	457	80	n11	interval	maximum	0.1874	494
31	n12a	interval	manhattan	0.1199	455	81	n9	ratio	maximum	0.1878	489
32	n9a	interval	seuclidean	0.1204	791	82	n8	ratio	euclidean	0.1883	419
33	n5	interval	manhattan	0.1207	479	83	n1	ratio	euclidean	0.1908	399
34	n5a	interval	manhattan	0.1225	479	84	n5	ratio	euclidean	0.1914	420
35	n8	interval	seuclidean	0.1255	688	85	n3	ratio	euclidean	0.1921	411
36	n9	ratio	manhattan	0.1257	486	86	n12a	ratio	euclidean	0.1923	398
37	n3	interval	seuclidean	0.1263	694	87	n5a	ratio	euclidean	0.1925	418
38	n8	ratio	seuclidean	0.1274	803	88	n1	interval	maximum	0.2229	437
39	n3	ratio	seuclidean	0.1279	802	89	n12a	interval	maximum	0.2242	441
40	n1	interval	seuclidean	0.1280	719	90	n11	ratio	maximum	0.2260	442
41	n8	interval	euclidean	0.1292	474	91	n8	interval	maximum	0.2307	460
42	n1	ratio	seuclidean	0.1297	845	92	n5a	interval	maximum	0.2368	424
43	n12a	interval	seuclidean	0.1300	718	93	n3	interval	maximum	0.2398	463
44	n1	interval	euclidean	0.1303	421	94	n5	interval	maximum	0.2442	443
45	n3	interval	euclidean	0.1307	461	95	n1	ratio	maximum	0.2547	396
46	n12a	ratio	seuclidean	0.1318	845	96	n12a	ratio	maximum	0.2557	395
47	n12a	interval	euclidean	0.1322	421	97	n5a	ratio	maximum	0.2606	394
48	n5	interval	seuclidean	0.1369	666	98	n8	ratio	maximum	0.2618	414
49	n11	interval	GDM1	0.1381	493	99	n3	ratio	maximum	0.2652	418
50	n5	interval	euclidean	0.1382	500	100	n5	ratio	maximum	0.2667	405

nm – normalization method; seuclidean – squared Euclidean distance.

Source: Authors' compilation using mdsOpt package and R program.

In the conducted study the maximal acceptable value of $Stress\text{-}1_p$ fit measure has been set to 0.15. Figure 2 presents the chart with $Stress\text{-}1_p$ fit measure value on *x*-axis and HHI_p index on *y*-axis for *p* procedures of multidimensional scaling.

Among acceptable multidimensional scaling procedures, for which $Stress\text{-}1_p \leq 0.15$, we chose the one for each occurs $\min_p \{HHI_p\}$ has been chosen. It is the procedure 47: `n12a` normalization method (positional normalization), interval MDS model, Euclidean distance.

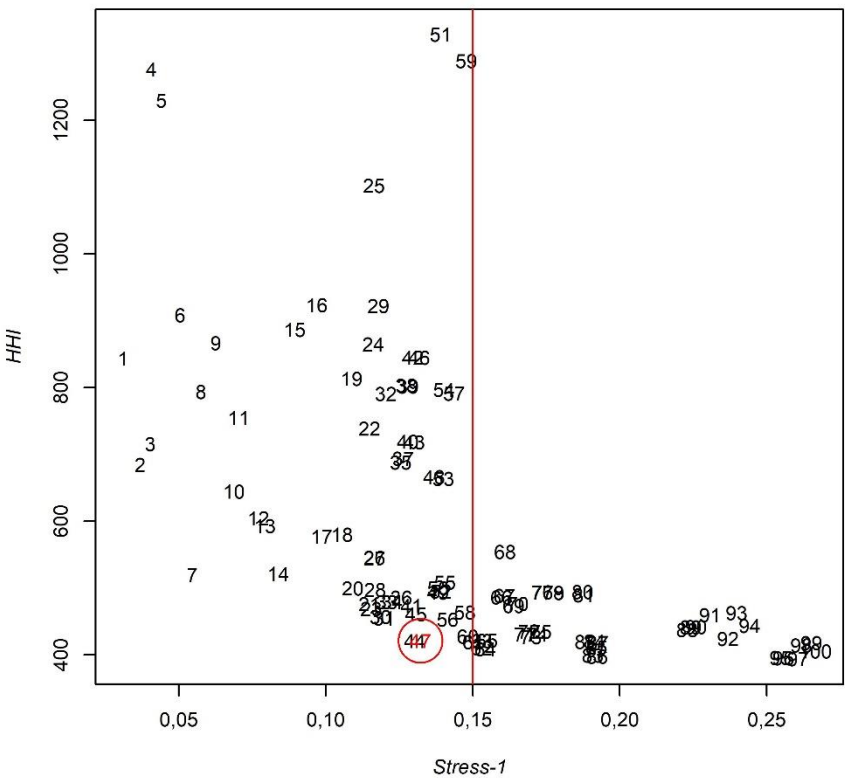
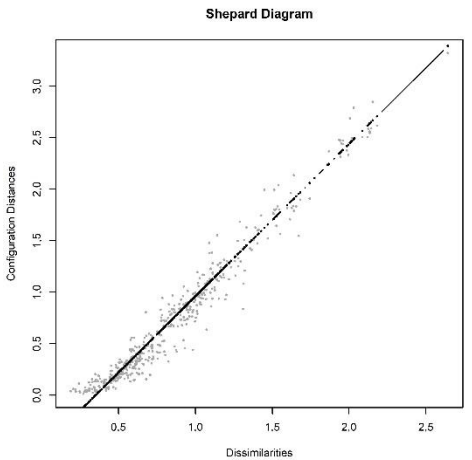


Figure 2. The values of $Stress\text{-}1_p$ fit measure and HHI_p index for *p* multidimensional scaling procedures

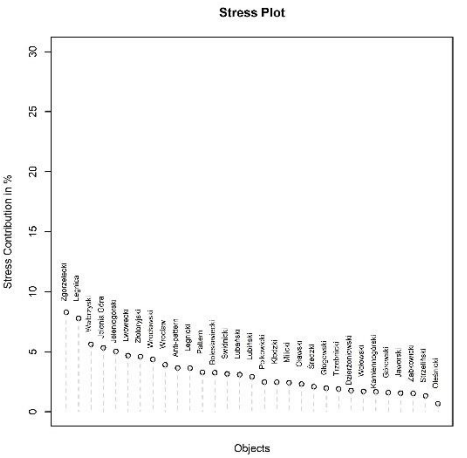
Source: Authors' compilation using `mdsOpt` *package of R program.*

The results of multidimensional scaling (procedure 47) of 31 objects (29 Lover Silesian counties, pattern and anti-pattern object) according to the level of tourist attractiveness are presented on Figure 3.

a)



b)



c)

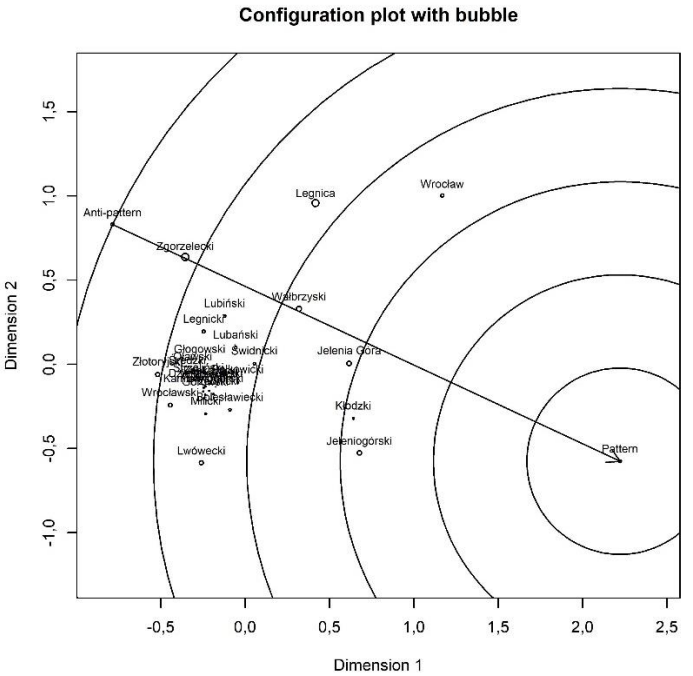


Figure 3. The results of multidimensional scaling (procedure 47) of 31 objects (29 Lover Silesian counties, pattern and anti-pattern) according to the level of tourist attractiveness (d_{ik} – Configuration Distances, δ_{ik} – Dissimilarities)

Source: Authors' compilation using R program.

Figure 3c (Configuration plot with bubble) presents additional quota of each object in total error is shown by the size of radius of the circle around each object. Shepard diagram (Figure 3a) confirms the correctness of the chosen scaling model (Pearson correlation coefficient $r = 0.9794$). Figure 3c (Configuration plot with bubble) shows the axis of the set, which is the shortest connection between the pattern and anti-pattern of development. It indicates the level of development of the tourist attractiveness of counties. Objects that are closer to the pattern of development have higher levels of tourist attractiveness. The isoquants[3] of development (curves of similar development) have been established from the point indicating pattern object. Figure 3c shows six isoquants. The same level of development may be achieved by objects from different locations on the same isoquant of development (due to different configuration of values of variables).

As opposed to the best MDS procedure (47) we show the results for one of the worst procedures (4): `n9a` normalization method, interval MDS model, maximum (Chebyshev) distance. Overall Stress for procedure 4 (0.0408) is significantly better than for procedure 47 (0.1322). The results of multidimensional scaling for procedure 4 according to the level of tourist attractiveness are presented in Figure 4.

Figure 4b (Stress Plot) indicates that objects Jeleniogórski, Anti-pattern and Zgorzelecki contribute most to the overall Stress (55.6%). It also shows (see Shepard diagram – in the lower left-hand corner) that two points (distance between Jeleniogórski county and Anti-pattern object; Jeleniogórski county and Zgorzelecki county) are outliers. These outliers contribute over-proportionally to the total Stress. MDS configuration (Figure 4c) does not represent all proximities equally well. Jeleniogórski county is one of the best of Lover Silesian counties in terms of the level of tourist attractiveness. In Figure 4c (Configuration plot with bubble) this county lies near Anti-pattern object (the worst object). The greater the value of the HHI_p index, the worse is the effect of multidimensional scaling in terms of representing real relationships between objects.

5. Summary and limitations of presented proposal

The article proposes a methodology that allows the selection of the optimum procedure due to the used methods of normalization, distance measures and scaling model of multidimensional scaling carried out on the basis of the metric data matrix. The study includes 18 methods of normalization, 5 distance measures and 3 models of scaling (ratio, interval and spline scaling).

Own package `mdsOpt` of R program to choose the optimal procedure for multidimensional scaling due to the normalization methods of variable values, distance measures and scaling models has been developed. On the basis of the proposed methodology research results are illustrated by an empirical example with the use of the function `smacofSym` of `smacof` package in order to find the

[3] Isoquants were illustrated using `draw.circle` function of `plotrix` package (Lemon et al., 2017).

optimal procedure for multidimensional scaling of set of objects representing 29 counties in Lower Silesia according to the level of tourist attractiveness.

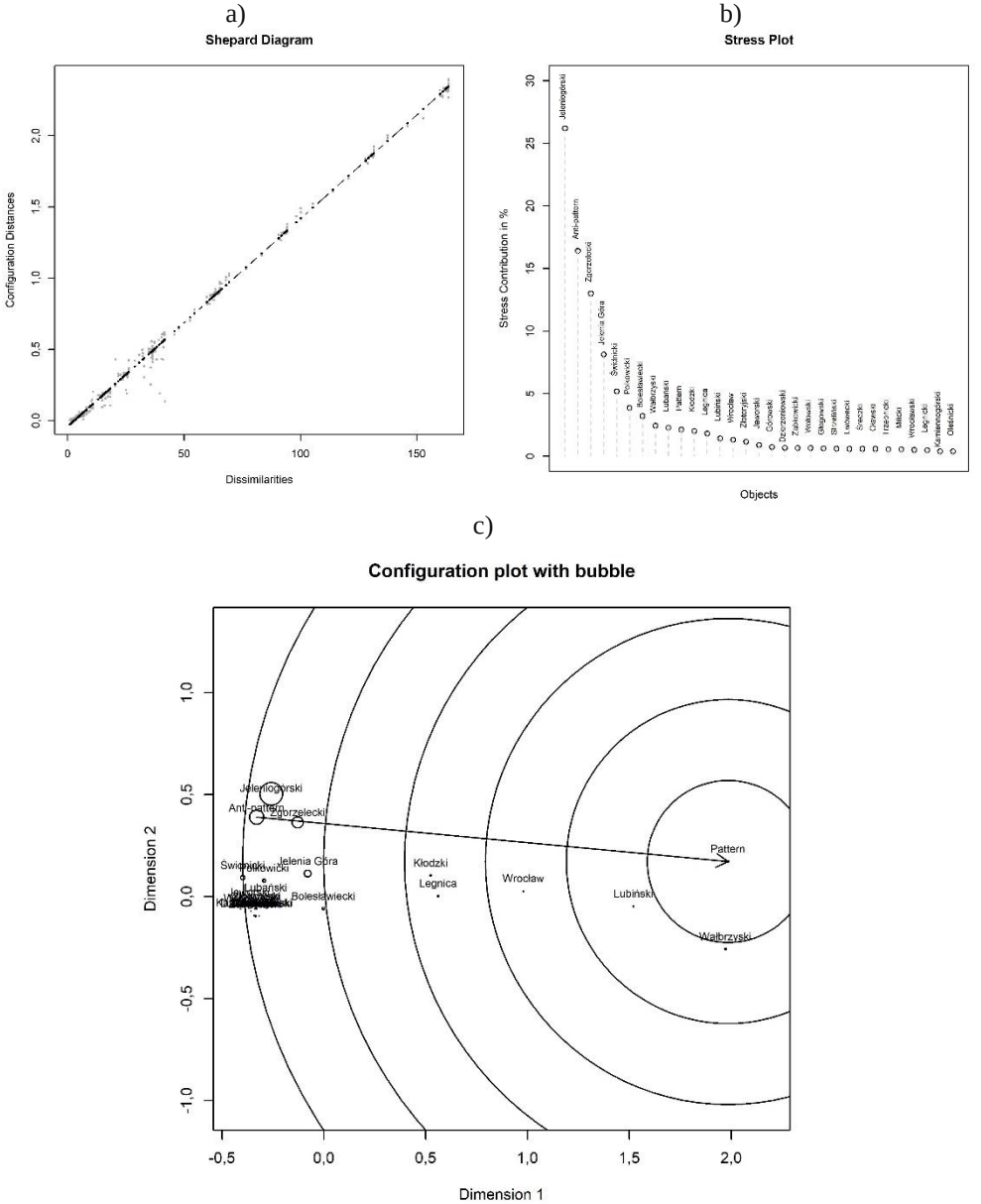


Figure 4. The results of multidimensional scaling (procedure 4) of 31 objects (29 Lover Silesian counties, pattern and anti-pattern) according to the level of tourist attractiveness

Source: Authors' compilation using R program.

The proposed methodology uses two criteria for selecting the optimal procedure for multidimensional scaling: *Stress*-1 loss function and the value of the Hirschman-Herfindahl *HHI* index calculated on the basis of the decomposition *Stress*-1 error by objects.

In step 5 the maximal acceptable value of fit measure $Stress\text{-}1 = s$ has been arbitrary assumed. The extent to which error distribution for each object may deviate from the uniform distribution is not determined. Among the procedures of multidimensional scaling for which $Stress\text{-}1_p \leq s$, the one for which $\min_p \{HHI_p\}$ occurs is selected. This constraint does not essentially limit the presented proposal as the additional criteria for acceptability of the results of multidimensional scaling plots, such as "Shepard diagram" and "Residual plot", make it possible to evaluate the fit quality of the chosen scaling model, and to identify outliers (De Leeuw, Mair, 2015).

REFERENCES

BORG, I., GROENEN, P. J. F., (2005). Modern Multidimensional Scaling. Theory and Applications, 2nd Edition, Springer Science+Business Media, New York. ISBN: 978-0387-25150-9, URL http://www.springeronline.com/0-387-25150-2.

BORG, I., GROENEN, P. J. F., MAIR, P., (2013). Applied Multidimensional Scaling, Springer, Heidelberg, New York, Dordrecht, London, URL http://dx.doi.org/10.1007/978-3-642-31848-1.

COOPER, L. G., (1983). A review of multidimensional scaling in marketing research, Applied Psychological Measurement, Vol. 7, No. 4, pp. 427–450, URL https://doi.org/10.1177/014662168300700404.

DE LEEUW, J., MAIR, P., (2009). Multidimensional Scaling Using Majorization: SMACOF in R. Journal of Statistical Software, 31 (3), pp. 1–30, URL http://dx.doi.org/10.18637/jss.v031.i03.

DE LEEUW, J., MAIR, P., (2015). Shepard Diagram, Wiley StatsRef: Statistics Reference Online, Wiley,
URL http://dx.doi.org/10.1002/9781118445112.stat06268.pub2.

EMBLETON, S., URITESCU, D., WHEELER, E. S., (2013). Defining dialect regions with interpretations: Advancing the multidimensional scaling approach, Literary and Linguistic Computing, Vol. 28, No. 1, pp. 13–22, URL https://doi.org/10.1093/llc/fqs048.

EVERITT, B.S., LANDAU, S., LEESE, M., STAHL, D., (2011). Cluster Analysis. John Wiley & Sons, Chichester. ISBN: 978-0-470-74991-3.

GOLLEDGE, R. G., RUHTON, G., (1972). Multidimensional Scaling: Review and Geographical Applications, Technical Paper No. 10. Association of American Geographers, WASHINGTON D. C.,
URL http://files.eric.ed.gov/fulltext/ED110362.pdf.

GOWER, J. C., (1966). Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika, Vol. 53, Issue 3-4, pp. 325–328, URL https://doi.org/10.1093/biomet/53.3-4.325.

GRYSZEL, P., WALESIAK, M., (2014). Zastosowanie uogólnionej miary odległości GDM w ocenie atrakcyjności turystycznej powiatów Dolnego Śląska [The Application of the General Distance Measure (GDM) in the Evaluation of Lower Silesian Districts' Attractiveness], Folia Turistica, 31, pp. 127–147,
URL http://www.folia-turistica.pl/attachments/article/402/FT_31_2014.pdf.

HERFINDAHL, O. C., (1950). Concentration in the Steel Industry, Doctoral thesis, Columbia University.

HIRSCHMAN, A. O., (1964). The Paternity of an Index, The American Economic Review, Vol. 54, No. 5, pp. 761-762, URL http://www.jstor.org/stable/1818582.

JAJUGA, K., WALESIAK, M., (2000). Standardisation of Data Set under Different Measurement Scales, In: Decker, R., Gaul, W., (Eds.), Classification and Information Processing at the Turn of the Millennium, 105-112. Springer-Verlag, Berlin, Heidelberg, URL http://dx.doi.org/10.1007/978-3-642-57280-7_11.

JAJUGA, K., WALESIAK, M., BĄK, A., (2003). On the General Distance Measure, in Schwaiger, M., Opitz, O., (Eds.), Exploratory Data Analysis in Empirical Research. Berlin, Heidelberg: Springer-Verlag, pp. 104–109, URL http://dx.doi.org/10.1007/978-3-642-55721-7_12.

LEMON, J., at al., (2017). plotrix: Various Plotting Functions. R package version 3.6-5, URL http://CRAN.R-project.org/package=plotrix.

MAIR, P., De LEEUW, J., BORG, I., GROENEN, P. J. F., (2017). smacof: Multidimensional Scaling. R package version 1.9-6, URL http://CRAN.R-project.org/package=smacof.

MARCUSSEN, C., (2014). Multidimensional scaling in tourism literature, Tourism Management Perspectives, Vol. 12, October, pp. 31–40, URL http://dx.doi.org/10.1016/j.tmp.2014.07.003.

PINKLEY, R.L., GELFAND, M.J., DUAN, L., (2005). When, Where and How: The Use of Multidimensional Scaling Methods in the Study of Negotiation and Social Conflict. International Negotiation, Vol. 10, Issue 1, pp 79–96, URL http://dx.doi.org/10.1163/1571806054741056.

R DEVELOPMENT CORE TEAM, (2017). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, URL http://www.R-project.org.

STEVENS, S. S., (1946). On the Theory of Scales of Measurement. Science, Vol. 103, No. 2684, pp. 677–680,
URL http://dx.doi.org/10.1126/science.103.2684.677.

TAKANE, Y., (2007). Applications of multidimensional scaling in psychometrics. In Rao, C.R., Sinharay, S. (Eds.), Handbook of Statistics, Vol. 26, Pyschometrics, Elsevier, Amsterdam, ISBN**:** 9780444521033, pp. 359–400.

TORGERSON, W. S., (1952). Multidimensional scaling: I. Theory and method. Psychometrika, Vol. 17, Issue 4, pp. 401–419,
URL https://link.springer.com/article/10.1007/BF02288916.

WALESIAK, M., (2016). Visualization of Linear Ordering Results for Metric Data with the Application of Multidimensional Scaling, Ekonometria [Econometrics], 2 (52), pp. 9–21,
URL http://dx.doi.org/10.15611/ekt.2016.2.01.

WALESIAK, M., DUDEK, A., (2017a). clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package version 0.45-2, URL http://CRAN.R-project.org/package=clusterSim.

WALESIAK, M., DUDEK, A., (2017b). mdsOpt: Searching for Optimal MDS Procedure for Metric Data. R package version 0.1-4, URL http://CRAN.R-project.org/package=mdsOpt.

ZELIAŚ, A., (2002). Some Notes on the Selection of Normalisation of Diagnostic Variables, Statistics in Transition, 5 (5), pp. 787–802.

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 541–548, DOI 10. 21307

SAMPLE ALLOCATION IN ESTIMATION OF PROPORTION IN A FINITE POPULATION DIVIDED AMONG TWO STRATA

Dominik Sieradzki[1], **Wojciech Zieliński**[2]

ABSTRACT

The problem of estimating a proportion of objects with a particular attribute in a finite population is considered. The classical estimator is compared with the estimator, which uses the information that the population is divided among two strata. Theoretical results are illustrated with a numerical example.

Key words: survey sampling, sample allocation, stratification, estimation, proportion.

1. Introduction

Consider a population $U = \{u_1, u_2, \ldots, u_N\}$ which contains a finite number of N units. In this population we can observe objects which have a given characteristic (property), for example sex, defectiveness, support for a particular candidate in elections, etc. Let M denote an unknown number of units in the population with a given property. We would like to estimate M, or equivalently, a proportion (fraction) $\theta = \frac{M}{N}$. A sample of size n is drawn using simple random sampling without replacement scheme. In the sample the number of objects with a particular attribute is observed. This number is a random variable. To be formal, let ξ be a random variable describing number of units having a certain attribute in the sample. The random variable ξ has hypergeometric distribution (Zieliński 2010) and its statistical model is

$$(\{0,1,\ldots,n\}, \{H(N,\theta N,n), \theta \in \langle 0,1\rangle\}), \tag{1}$$

with probability distribution function

$$P_{\theta,N,n}\{\xi = x\} = \frac{\binom{\theta N}{x}\binom{(1-\theta)N}{n-x}}{\binom{N}{n}}, \tag{2}$$

[1]Department of Econometrics and Statistics, Warsaw University of Life Sciences.
E-mail: dominik_sieradzki@sggw.pl
[2]Department of Econometrics and Statistics, Warsaw University of Life Sciences.
E-mail: wojciech_zielinski@sggw.pl

for integer x from interval $\langle \max\{0, n-(1-\theta)N\}, \min\{n, \theta N\}\rangle$. Unbiased estimator with minimal variance of the parameter θ is $\hat{\theta}_c = \frac{\xi}{n}$ (Bracha 1998). Variance of that estimator equals

$$D^2_\theta \hat{\theta}_c = \frac{1}{n^2} D^2_\theta \xi = \frac{\theta(1-\theta)}{n} \frac{N-n}{N-1} \text{ for all } \theta. \tag{3}$$

It is easy to calculate that variance $D^2_\theta \hat{\theta}_c$ takes on its maximal value at $\theta = \frac{1}{2}$.

2. Stratified estimator

Let contribution of the first strata be w_1, i.e. $w_1 = N_1/N$. Hence, the overall proportion θ equals

$$\theta = w_1\theta_1 + w_2\theta_2, \tag{4}$$

where $w_2 = 1 - w_1$. It seems intuitively obvious to take as our estimate of θ,

$$\hat{\theta}_w = w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2}, \tag{5}$$

where n_1 and n_2 denote sample sizes from the first and the second strata, respectively. Now, we have two random variables describing the number of units with a particular attribute in samples drawn from each strata:

$$\xi_1 \sim H(N_1, \theta_1 N_1, n_1), \quad \xi_2 \sim H(N_2, \theta_2 N_2, n_2). \tag{6}$$

The whole sample size equals $n = n_1 + n_2$. The question now arises: how shall we choose n_1 and n_2 to obtain the best estimate of θ? This problem concerns sample allocation between strata. One of known approaches to this problem is proportional allocation (Armitage 1943, Cochran 1977). Sample sizes n_1 and n_2 are proportional to w_1 and w_2,

$$n_1 = w_1 n \quad \text{and} \quad n_2 = w_2 n. \tag{7}$$

The second approach to sample allocation is Neyman Allocation (Neyman 1934). This method gives values of n_1 and n_2, which minimize the variance of estimator $\hat{\theta}_w$ for given θ_1 and θ_2. The values of n_1 and n_2 are as follows

$$n_i = \frac{w_i \sqrt{\theta_i(1-\theta_i)}}{\sum_i w_i \sqrt{\theta_i(1-\theta_i)}} n, \quad i = 1, 2. \tag{8}$$

Neyman Allocation requires knowledge of the parameters θ_1 and θ_2. Those magnitudes would be known exactly when the population were subjected to exhaustive

sampling. Usually values θ_1 and θ_2 are estimated from a preliminary sample. In some cases fairly good estimates of θ_1 and θ_2 are available from past experience (Armitage 1943).

Since our aim is to estimate θ, hence the parameter θ_1 will be considered as a nuisance one. This parameter will be eliminated by appropriate averaging. Note that for a given $\theta \in [0,1]$, parameter θ_1 is a fraction M_1/N_1 (it is treated as the number, not as the random variable) from the set

$$\mathscr{A} = \left\{ a_\theta, a_\theta + \frac{1}{N_1}, a_\theta + \frac{2}{N_1}, \ldots, b_\theta \right\}, \tag{9}$$

where

$$a_\theta = \max\left\{0, \frac{\theta - w_2}{w_1}\right\} \quad \text{and} \quad b_\theta = \min\left\{1, \frac{\theta}{w_1}\right\} \tag{10}$$

and let L_θ be cardinality of $\mathscr{A}$.

Theorem. Estimator $\hat{\theta}_w$ is an unbiased estimator of θ.

Proof. Note that for a given θ there are L_θ values of θ_1 and θ_2 giving θ. Hence, averaging with respect to θ_1 is made assuming the uniform distribution of θ_1 on the set $\{a_\theta, \ldots, b_\theta\}$. We have

$$\begin{aligned} E_\theta \hat{\theta}_w = E_\theta \left(w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2} \right) &= \frac{1}{L_\theta} \sum_{\theta_1 \in \mathscr{A}} \left(\frac{w_1}{n_1} E_{\theta_1} \xi_1 + \frac{w_2}{n_2} E_{\frac{\theta - w_1\theta_1}{w_2}} \xi_2 \right) \\ &= \frac{1}{L_\theta} \sum_{\theta_1 \in \mathscr{A}} \left(\frac{w_1}{n_1} \frac{\theta_1 N_1 n_1}{N_1} + \frac{w_2}{n_2} \frac{\frac{\theta - w_1\theta_1}{w_2} N_2 n_2}{N_2} \right) \\ &= \theta \end{aligned} \tag{11}$$

for all θ.

Averaged variance of estimator $\hat{\theta}_w$ equals:

$$\begin{aligned} D^2_\theta \hat{\theta}_w = D^2_\theta \left(w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2} \right) &= \\ = \frac{1}{L_\theta} \sum_{\theta_1 \in \mathscr{A}} \left(\left(\frac{w_1}{n_1} \right)^2 D^2_{\theta_1} \xi_1 + \left(\frac{w_2}{n_2} \right)^2 D^2_{\frac{\theta - w_1\theta_1}{w_2}} \xi_2 \right) &= \\ = \frac{1}{L_\theta} \sum_{\theta_1 \in \mathscr{A}} \left[\frac{w_1^2}{n_1} \theta_1 (1 - \theta_1) \frac{N_1 - n_1}{N_1 - 1} + \frac{w_2^2}{n_2} \frac{\theta - w_1\theta_1}{w_2} \left(1 - \frac{\theta - w_1\theta_1}{w_2} \right) \frac{N_2 - n_2}{N_2 - 1} \right]. \end{aligned} \tag{12}$$

Let $f = \frac{n_1}{n}$ denote the contribution of the first strata in the sample. For $0 < \theta < w_1$

variance of $\hat{\theta}_w$ equals $\left(a_\theta = 0 \text{ and } b_\theta = \frac{\theta}{w_1}\right)$:

$$\begin{aligned} &\frac{h(f)}{-6(N_1-1)(N_2-1)Nf(1-f)n}\theta \\ &+\frac{(N_2-1)N_1-(N(n+1)-2(N_1+n))f+(N-2)nf^2}{3(N_1-1)(N_2-1)f(1-f)n}\theta^2, \end{aligned} \tag{13}$$

where

$$\begin{aligned} h(f) &= N_1(N_2-3N_1(N_2-1)-1) \\ &+\left(3N_1^2(N_2-1)+3N_2^2+2n+N_1\left(6N_2n-3N_2^2-4n+1\right)-N_2(4n+1)\right)f \\ &+2\left(N_1(2-3N_2)+2N_2-1\right)nf^2 \end{aligned} \tag{14}$$

For $w_1 \leq \theta \leq 1-w_1$ variance of $\hat{\theta}_w$ equals ($a_\theta = 0$ and $b_\theta = 1$):

$$\begin{aligned} &\frac{(N_2-(1-f)n)}{(N_2-1)(1-f)n}\theta(1-\theta)+ \\ &-\frac{N_1\left(2(N+1)f^2+(3NN_2+N_2-N_1-2n(N+1))f-N_1(N_2-1)\right)}{6N^2(N_2-1)nf(1-f)} \end{aligned} \tag{15}$$

To obtain explicit formula for variance of $\hat{\theta}_w$ for $1-w_1 < \theta < 1$ it is sufficient to replace θ by $1-\theta$ in (13). Observe that variance $D^2_\theta\hat{\theta}_w$ depends on size n of the sample, size N of the population, contribution w_1 of the first strata in population and contribution f of the first strata in the sample. In Figure 1 variances of $\hat{\theta}_w$ and $\hat{\theta}_c$ are drawn against θ, for $N = 100000$, $n = 100$, $w_1 = 0.4$ and $f = 0.3$.

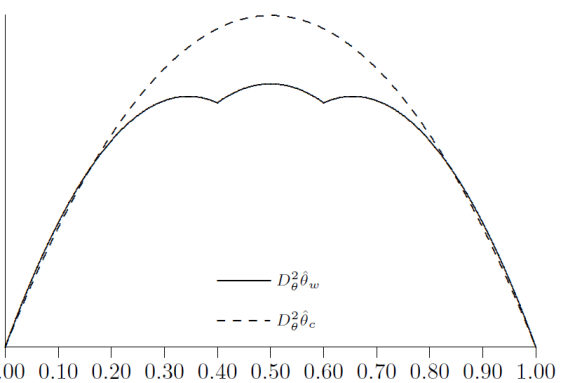


Figure 1. Variances of $\hat{\theta}_c$ and $\hat{\theta}_w$ for $w_1 = 0.4$ and $f = 0.3$

Source: Own calculations.

It is easy to note that $D^2_\theta \hat{\theta}_w = D^2_{1-\theta} \hat{\theta}_w$ and $D^2_0 \hat{\theta}_w = 0$.

Maximum of variance $D^2_\theta \hat{\theta}_w$ determines for which value of unknown parameter θ estimation of θ is the worst one. After the analysis of variance of $\hat{\theta}_w$, it is seen that the maximal variance may be in the one of the intervals: $(0, w_1)$, $(w_1, 1-w_1)$ or $(1-w_1, 1)$. It depends on the values of w_1 and f. In Figures 2, 3, 4 and 5 variance of $\hat{\theta}_w$ as well as variance of $\hat{\theta}_c$ is drawn for $N = 100000$, $n = 100$, $w_1 = 0.4$ and $f = 0.2, 0.4, 0.6, 0.9$.

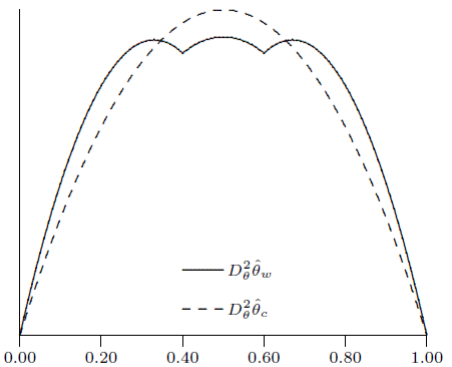


Figure 2. Variances of $\hat{\theta}_c$ and $\hat{\theta}_w$ for $w_1 = 0.4$ and $f = 0.2$

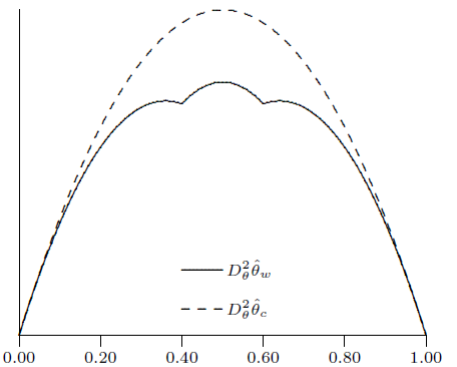


Figure 3. Variances of $\hat{\theta}_c$ and $\hat{\theta}_w$ for $w_1 = 0.4$ and $f = 0.4$

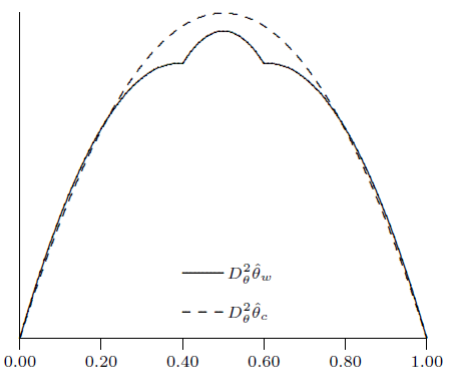


Figure 4. Variances of $\hat{\theta}_c$ and $\hat{\theta}_w$ for $w_1 = 0.4$ and $f = 0.6$

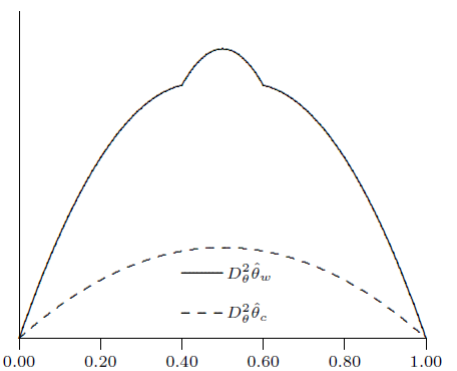


Figure 5. Variances of $\hat{\theta}_w$ and $\hat{\theta}_c$ for $w_1 = 0.4$ and $f = 0.9$

Source: Own calculations.

The point at which $D^2_\theta \hat{\theta}_w$ takes on the maximal value may be located in interval $(0, w_1)$ or in interval $(w_1, 1-w_1)$. Hence, to find the global maximum due to θ, we have to find local maximum in both intervals. Denote by θ^* a local maximum point in interval $(0, w_1)$ (local maximum point in interval $(1-w_1, 1)$ is $1-\theta^*$). In an interval $(w_1, 1-w_1)$ local maximum is achieved at $\theta = 1/2$. Let $\widetilde{\theta}$ denote a global

maximum point, i. e. $\widetilde{\theta} = 1/2$ or $\widetilde{\theta} = \theta^*$, hence

$$\max_{\theta \in \langle 0,1 \rangle} D^2_\theta \hat{\theta}_w = \max \left\{ D^2_{0.5} \hat{\theta}_w, D^2_{\theta^*} \hat{\theta}_w \right\}. \tag{16}$$

Regardless of which point is the global maximum point ($1/2$ or θ^*), the maximum of the variance $D^2_\theta \hat{\theta}_w$ depends on size n of the sample, size N of the population, contribution w_1 of the first strata in the population and the contribution f of the first strata in the sample. Values N, n, w_1 are treated as given. It may be seen that for given w_1, variance $D^2_\theta \hat{\theta}_w$ may be smaller as well as greater than $D^2_\theta \hat{\theta}_c$. We would like to find optimal f, which minimizes maximal variance $D^2_{\widetilde{\theta}} \hat{\theta}_w$.

3. Results

A general formula for the optimal f is unobtainable, because of complexity of symbolic computation. But for given N, w_1 and n numerical solution is easy to obtain. Table 1 shows some numerical results for $N = 100000$ and $n = 100$.

Table 1. Maximal variances $D^2_{\widetilde{\theta}} \hat{\theta}_w$

w_1	f^{opt}	n_1^{opt}	$D^2_{\widetilde{\theta}} \hat{\theta}_w$	$D^2_{0.5} \hat{\theta}_c$	$\left(1 - \frac{D^2_{\widetilde{\theta}} \hat{\theta}_w}{D^2_{0.5} \hat{\theta}_c}\right) \cdot 100\%$
0.05	0.018	2	0.0004645	0.0025	81%
0.10	0.041	4	0.0008404	0.0025	66%
0.15	0.071	7	0.0011328	0.0025	55%
0.20	0.111	11	0.0013493	0.0025	46%
0.25	0.166	17	0.0015004	0.0025	40%
0.30	0.250	25	0.0015984	0.0025	36%
0.35	0.350	35	0.0017045	0.0025	32%
0.40	0.400	40	0.0017982	0.0025	28%
0.45	0.450	45	0.0018544	0.0025	26%
0.50	0.500	50	0.0018731	0.0025	25%

Source: Own calculations.

In the first column of Table 1. the values of w_1 are given. In the second column, optimal contribution of the first strata in the sample is shown. It is a value f, which gives minimum of $D^2_{\widetilde{\theta}} \hat{\theta}_w$. Column n_1^{opt} shows optimal sample size from the first strata (called averaged sample allocation). The values of minimal (maximal) variances $D^2_{\widetilde{\theta}} \hat{\theta}_w$ are given in the fourth column. The next column contains maximal variance $D^2_{0.5} \hat{\theta}_c$. The last column shows how much estimator $\hat{\theta}_w$ is better than $\hat{\theta}_c$.

4. Summary

In the paper a new approach to the sample allocation between strata was proposed. Two estimators of an unknown fraction θ in the finite population were considered: standard estimator $\hat{\theta}_c$ and stratified estimator $\hat{\theta}_w$. It was shown that both estimators are unbiased. Their variances were compared. It appears that for a given sample size there exists its optimal allocation between strata, i.e. the allocation for which variance of $\hat{\theta}_w$ is smaller than variance of $\hat{\theta}_c$. Since a theoretical comparison seems to be impossible, hence a numerical example was presented. In that example it was shown that variance of the stratified estimator may be smaller at least 25% with respect to variance of the classical estimator. For such an approach there is no need to estimate unknown θ_1 and θ_2 by preliminary sample. It will be interesting to generalize the above results to the case of more than two “subpopulations". Work on the subject is in progress.

REFERENCES

ARMITAGE, P., (1947). A Comparison of Stratified with Unrestricted Random Sampling from a Finite Population, Biometrika, 34, 3/4 , pp. 273–280.

BRACHA, CZ., (1998). Metoda reprezentacyjna w badaniach opinii publicznej i marketingu. PWN, Warszawa.

COCHRAN, W. G., (1977). Sampling Techniques (3rd ed.), New York: John Wiley.

NEYMAN, J., (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection, Journal of the Royal Statistical Society, 97, pp. 558–606.

SIERADZKI, D., (2016). Estimation of proportion in finite population divided into two strata, master thesis, WZIiM SGGW Warszawa (in polish).

ZIELIŃSKI, W., (2010). Estymacja wskaźnika struktury, Wydawnictwo SGGW, Warszawa.

ZIELIŃSKI, W., (2016). A remark on estimating defectiveness in sampling acceptance inspection, Colloquium Biometricum, 46, pp. 9–14.

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 549–558, DOI 10. 21307

REMARKS ON THE ESTIMATION OF POSITION PARAMETERS

Czesław Domański[1]

ABSTRACT

The article contains some theoretical remarks about selected models of position parameters estimation as well as numerical examples of the problem. We ask a question concerning the existence of possible measures of the quality of interval estimation and we mention some popular measures applied to the task. Point estimation is insufficient in practical problems and it is rather interval estimation that is in wide use. Too wide interval suggests that the information available is not sufficient to make a decision and that we should look for more information, perhaps by increasing the sample size.

Key words: estimation, the positional parameters, statistical models

1. Introduction

When it is impossible to state what the level of accuracy of estimation of random variable parameter is, the question arises whether there are any methods which help to determine the distance between the estimator assessment and the real value of parameter. The answer to this question is provided by J. Neyman – the author of the interval estimation (1937). Sometimes the interval we obtain is too wide. Too wide intervals allow us to draw a conclusion that the available information is not sufficient to take a decision, and therefore we need to search for more information, either by widening the scope of research or by running another series of experiments.

The interval estimation includes almost all types of statistical analyses. In public opinion polls, for instance, when we state that 58% of citizens of the Republic of Poland trust the president usually a footnote should be added stating that the poll is biased with „an error of plus or minus 3%". This means that 58% of the interviewees trust the president. As the research was based on a representative sample, the parameter sought is the percentage of all people who think in this way. Due to a small sample size a reasonable "guess" is that the

[1] University of Lodz, Chair of Statistical Methods. E-mail: czedoman@uni.lodz.pl

parameter can encompass the interval 55% (54% minus 3%) to 61% (57% plus 3%).

How should the results of the interval estimation be interpreted? Can probability assumptions be made on the basis of interval estimation? How certain is the researcher that the parameter searched for will be included in a given interval?

Neyman (1935) proposed an accessible way of constructing interval estimation, defining how accurate the estimation is and calling the new procedure „confidence intervals", and the ends of confidence intervals – „confidence limits".

Neyman (1937) went back to the frequency definition of a real probability. In his later works he provided a more detailed explanation of confidence intervals stating that they should be perceived not as an individual conclusion but rather as a process. In the long term the statistician who always calculates 95% confidence intervals will see that in 95% of cases the real value of parameter can be found in the determined intervals. It is worth mentioning that Neyman was right saying that the probability connected with confidence interval was not a probability. It rather represents the frequency of correct conclusions drawn by a statistician using this method over a longer period of time but says nothing about the „accuracy" of the current estimation.

Majority of researchers find 90% or 95% confidence limits and continue as if they were certain that the interval encompassed the real value of parameter.

2. Statistical models

Every statistical analysis of a certain real phenomenon must be based on a mathematical model (i.e. a model expressed in the form of mathematical dependencies where the way of obtaining information was taken into account).

The researcher should aim at a situation where the applied model is a modest description of nature. This means that the functional form of the model should be simple and the number of its parameters and elements as small as possible.

As we know there are no perfect models which perfectly copy the behaviour of the modelled object. Each new observation and an analysis of the discrepancy between the mathematical model and the real object leads to new, more accurate mathematical models. The main reasons for the discrepancy between the model and the modelled phenomenon are as follows (Domański et al. (2014)):

1) the present state of knowledge on the examined phenomenon;
2) high level of dependence of the modelled phenomenon, which prevents the application of the mathematical model encompassing all qualities of the object;
3) variety and changeability of the object's environment where modelling of the real reasons for the object's condition becomes impossible;
4) costs related to the model's application can become a barrier to the model's complexity. It may occur that a simpler model despite being less accurate

turns out to be better, as the profits connected with giving up complicated measurement often exceed the losses resulting from using a less accurate model.

The starting point for our discussions is always a certain random element X (random variable, finite or non-finite series of random variables). Most frequently it will be called an experiment result, a measurement result, an observation result or, simply an observation. The set of all values of the random element X will be denoted by χ and called space χ of the sample. Space χ will be a finite or a countable set, or a certain area in a finite dimensional space R^n.

Let Ω be a set of elementary events and let $\aleph$ be σ - a body of subsets of the set Ω. An ordered triple $(\Omega, \aleph, P)$ is called a **probabilistic space**, where P denotes probability.

Let A be a distinguished σ-body of subsets of the set $X \subset R^n$, and X a measured transformation $(\Omega, \aleph) \to (\chi, A)$. Distribution $P^X(A) = P\left(X^{-1}(A)\right)$ is a measure on space (χ, A). In statistical problems it is assumed that distribution P belongs to a certain defined class of distributions $\mathcal{P}$ on (χ, A). Knowing the class and having the results of observation of the random variable *X*, we want to draw correct conclusions about an unknown distribution *P*. Thus, a mathematical basis for statistical research is a measured space (χ, A) and a family of distributions $\mathcal{P}$. Probabilistic space $(\Omega, \aleph, P)$ plays a subordinate role. The term: a probabilistic space $(\Omega, \aleph, P)$ is given, which means that a probabilistic model of a certain phenomenon or experiment is known i.e. we know what are the possible results of the experiment, what events are distinguished and what probabilities are assigned to these events. To sum up, the *a priori* knowledge of the subject of research is given in the form of certain probabilistic models. Probability may result from the very nature of the examined phenomenon or it can be introduced by a researcher.

Let us note that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of distributions of probability on a given σ-body of random events in χ.

The sample space together with a family of distributions $\mathcal{P}$, i.e. the object:

$$(\chi, \{P_\theta : \theta \in \Theta\}) \tag{1}$$

is called a **statistical model** (statistical space), while representations from χ in R^k – statistics or k-dimensional statistics.

If $\mathbf{X} = (X_1, X_2, ..., X_n)^T$, while $X_1, X_2, ..., X_n$ are independent random variables with a uniform distribution, we will also use a denotation:

$$(\chi, \{P_\theta : \theta \in \Theta\})^n \tag{2}$$

where $\mathcal{X}$ is a set of values of the random variable X (and each of variables $X_1, X_2, ..., X_n$) and P_θ is a distribution of the random variable. It is also accepted to use the following terms: $X_1, X_2, ..., X_n$ is a sample from distribution P_θ or a sample from population P_θ for a given $\theta \in \Theta$.

3. Confidence intervals for expected value μ

To estimate a certain unknown, real parameter μ we get suitable observations $X_1, \ldots, X_n$ of this value. Each observation $X_j, j = 1, \ldots, n$, was different from μ by a certain random value ε_j (statistical observation error). If nothing is known about the nature of the error ε, then consequently nothing can be said about the size of μ. However, if we can describe the random error ε in terms of the theory of probability, i.e. if we can say something about the distribution of the probability of this random error, then we can in the same terms answer various questions about parameter μ. Thus, the statistical inference becomes a result of the prior knowledge about the parameter and the knowledge obtained from the sample $X_1, \ldots, X_n$.

Let a distribution of random error probability μ be denoted by F; then the sample has a distribution F_μ so that $F_\mu(x) = F(x - \mu)$.

Let us now, on the other hand, analyse four general models of our observations $X_1, \ldots, X_n$.

- Model 1: F is a normal distribution $N(0, \sigma)$ with a known standard deviation σ.
- Model 2: F is a normal distribution $N(0, \sigma)$ with an unknown standard deviation σ.
- Model 3: F is a known distribution with a continuous and strictly ascending distribution function.
- Model 4: F is an unknown distribution with a continuous and strictly ascending distribution function. In this case it seems that „in actual fact we know nothing", yet it turns out that knowing that the distribution function is continuous and strictly monotonous is sufficient to say something more interesting about the parameter μ, especially when we combine this with data from observation $X_1, \ldots, X_n$.

In the first model the estimation of parameter μ by a mean value from observation

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \tag{3}$$

It is assumed that X has a distribution $N(\mu, \sigma)$, then the mean $\bar{X}_n$ is a random variable with a normal distribution $N(\mu, \sigma/\sqrt{n})$, in other words $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is

a random variable with a normal distribution $N(0,1)$ and for an arbitrarily selected $\gamma \in (0,1)$ we get

$$P_\mu\{|\sqrt{n}(\bar{X}_n - \mu)/\sigma| \leq u_{(1+\gamma)/2}\} = \gamma \tag{4}$$

where u_α is a quantile of an order α of a normal distribution $N(0,1)$.

This can be denoted in the form

$$P_\mu\left\{\bar{X}_n - u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\right\} = \gamma \tag{5}$$

and interpreted in the following way: with a selected probability γ, a random interval

$$\left(\bar{X}_n - u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\right) \tag{6}$$

includes the unknown, estimated value of parameter μ.

In the second model the estimation of parameter μ is based on the t Student distribution. In the case under consideration we deal with a random variable

$$\frac{\frac{\bar{X}_n - \mu}{\sigma}\sqrt{n}}{\sqrt{\frac{nS^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}_n - \mu}{S}\sqrt{n-1} \tag{7}$$

with the t Student distribution and with $(n-1)$ degrees of freedom.

The possibility of inference on parameter μ changes, because the random variable $\frac{\bar{X}_n - \mu}{S}\sqrt{n-1}$ with the t Student distribution is more dispersed around zero than the random variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ with the normal distribution.

Then, for the estimated parameter μ we get a confidence interval at a given level of confidence γ of the form :

$$\left(\bar{X}_n - t_{n-1}\left(\frac{1+\gamma}{2}\right)\frac{S}{\sqrt{n-1}}, \bar{X}_n + t_{n-1}\left(\frac{1+\gamma}{2}\right)\frac{S}{\sqrt{n-1}}\right) \tag{8}$$

where $t_{n-1}(\alpha)$ is a quantile of order α of the t Student distribution with $n-1$ degrees of freedom.

When the standard deviation σ was known like in the first model, the length of the confidence interval (2d) at the confidence level γ could be expressed with the formula $2\,u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}$ and on this basis the required accuracy of the estimation of parameter μ could be obtained. If the unknown standard deviation σ is replaced with its estimation S, then the length of interval calculated in this way will be random. The problem consists in selecting n, in such a way that the random variable never exceeds the pre-assigned number 2d. There are various methods of solving this problem. The simplest and the most transparent method is the so-called two-stage Stein procedure (1956).

In the third model it is the median $M_n$which is the third estimated position parameter. Median μ of the distribution of observations will be estimated with the

use of median M_n from a sample $X_1, \ldots, X_n$. According to a generally accepted agreement the median M_n from a sample is expressed by the following formula:

$$M_n = \begin{cases} \frac{1}{2}\left(\mathrm{X}_{\frac{\mathrm{n}}{2}:\mathrm{n}} + \mathrm{X}_{\frac{\mathrm{n}}{2}+1:\mathrm{n}}\right), \text{for even n,} \\ \mathrm{X}_{\frac{\mathrm{n}+1}{2}:\mathrm{n}}, \text{for uneven n .} \end{cases} \tag{9}$$

Let us now analyse the problem of the biasedness of estimator M_n. The basic definition where estimator T is called the unbiased parameter θ if $E_\theta T = \theta$ for every θ, cannot be applied here due to the fact that the median M_n cannot have the expected value. We can introduce the notion of median unbiasedness. We say that estimator T is the median-unbiased estimator of parameter θ if for every θ its median is $Med_\theta T = \theta$. In other words, T is the median-unbiased estimator of parameter θ if

$$P_\theta\{T \le \theta\} = P_\theta\{T \ge \theta\} = \frac{1}{2}, \text{ for every } \theta \tag{10}$$

under the assumption that, similarly to the distribution of observation X, also the distribution of estimator T has a continuous and strictly ascending distribution function, that is an unambiguous median.

If the sample $X_1, \ldots, X_n$ has an uneven number of elements n, then the median M_n from the sample is a median-unbiased estimator of median μ of distribution F_μ of observation X. It can be noticed that the distribution function of the k-th position statistics $X_{k,n}$, when the sample comes from a distribution with the distribution function F takes the following form:

$$F_{k,n}(x) = \sum_{j=k}^{n} \binom{n}{j} F^j(x)\big(1 - F(x)\big)^{n-j} \tag{11}$$

Let us recall here the formula combining binominal distribution with beta distribution:

$$\sum_{j=1}^{n} \binom{n}{j} x^j (1-x)^{n-j} = B(x; k, n-k+1) \tag{12}$$

Following from (11) and (12) the distribution function of median M_n is given by the formula:

$$P_\mu\{M_n \le x\} = B\left(F(x-\mu); \frac{n+1}{2}, \frac{n+1}{2}\right), \tag{13}$$

therefore

$$P_\mu\{M_n \le x\} = B\left(F(0); \frac{n+1}{2}, \frac{n+1}{2}\right) = B\left(\frac{1}{2}; \frac{n+1}{2}, \frac{n+1}{2}\right) = \frac{1}{2} \tag{14}$$

In the case of the sample $X_1, \ldots, X_n$ with the even number of elements the median M_n, which was defined by formula (9), is not the median-unbiased estimator of the median μ and for some distributions F_μ of observations X the difference between the median of estimator M_n and the median μ can be very significant.

Our considerations are now limited to the case of uneven number of observations n in a sample. For the case like this the distribution of the median from a sample is given by the formula (13).

Now, let $x_\gamma(M_n)$ be the quantile of the order γ of estimator M_n, i.e. such a number that

$$P_\mu\{M_n \le x_\gamma(M_n)\} = \gamma \tag{15}$$

On the basis of (13) we get:

$$x_\gamma(M_n) = \mu + F^{-1}\left(B^{-1}\left(\gamma; \tfrac{n+1}{2}, \tfrac{n+1}{2}\right)\right) \tag{16}$$

and hence the unilateral confidence interval on the confidence level γ takes the form:

$$\left(M_n - F^{-1}\left(B^{-1}\left(\gamma; \tfrac{n+1}{2}, \tfrac{n+1}{2}\right)\right), +\infty\right). \tag{17}$$

Similarly, taking as a basis the relation

$$P_\mu\left\{|M_n| \le x_{\frac{1+\gamma}{2}}(M_n)\right\} = \gamma, \tag{18}$$

we get a bilateral confidence interval at the confidence level γ:

$$\left(M_n - F^{-1}\left(B^{-1}\left(\tfrac{1+\gamma}{2}; \tfrac{n+1}{2}, \tfrac{n+1}{2}\right)\right), M_n + F^{-1}\left(B^{-1}\left(\tfrac{1+\gamma}{2}; \tfrac{n+1}{2}, \tfrac{n+1}{2}\right)\right)\right) \tag{19}$$

where F is a normal distribution $N(0, \sigma)$.

In the fourth model the confidence interval for median is presented. First, we consider constructing the confidence interval for a quantile $x_q = F^{-1}(q)$ of an arbitrary order $q \in (0,1)$, then the confidence interval for the median is a special case for $q = \frac{1}{2}$.

As we analyse the unilateral interval of the form $(X_{i:n}, +\infty)$ with an assumed level of confidence γ, we should choose index $i \in \{1,2,\dots,n\}$ so that $P_F\{X_{i:n} \le x_q\} \ge \gamma$ for every $F \in \mathcal{F}$. As $X_{i:n} < X_{j:n}$, when $i < j$, it is reasonable to choose the biggest number $i = i(n.\gamma)$ which satisfies the given condition. Making use of the distribution of the i −th position statistics from a sample $X_1, \dots, X_n$, of the form (11), we get:

$$\begin{aligned} P_F\{X_{i:n} \le x_q\} &= P_F\{X_{i:n} \le F^{-1}(q)\} \\ &= \sum_{j=i}^{n} \binom{n}{j} \left(F\left(F^{-1}(q)\right)\right)^j \left(1 - F\left(F^{-1}(q)\right)\right)^{n-j} \\ &= \textstyle\sum_i^n \binom{n}{j} q^j (1-q)^{n-j}. \end{aligned} \tag{20}$$

The solution is the biggest $i = i(n,q)$ so that

$$\sum_{j=i(n,\gamma)}^{n} \binom{n}{j} q^j (1-q)^{n-j} \geq \gamma \tag{21}$$

The confidence interval at the level γ for the quantile of the order $q \in (0,1)$ only exists when

$$\sum_{j=i}^{n} \binom{n}{j} q^j (1-q)^{n-j} \geq \gamma \tag{22}$$

i.e. when $(1-q)^n \leq 1-\gamma$.

As a conclusion we get the unilateral confidence interval for median $(X_{i:n}, +\infty)$, where $i = i\left(n, \frac{1}{2}\right) \in \{1, \dots, n\}$ is the biggest number such that

$$2^{-n} \sum_{s=i(n,\gamma)}^{n} \binom{n}{s} \geq \gamma \tag{23}$$

Due to the discreteness of the distribution the actual confidence interval

$$\gamma^* = 2^{-n} \sum_{j=i(n,\gamma)}^{n} \binom{n}{j} \tag{24}$$

can obviously be bigger than the assumed γ.

The bilateral confidence interval $(X_{i:n}, X_{j:n})$ takes the form:

$$P_F\{X_{i:n} \leq F^{-1}(q) \leq X_{j:n}\} = P_F\{X_{i:n} \leq F^{-1}(q)\} - P_F\{X_{j:n} > F^{-1}(q)\}$$

$$= \sum_{s=1}^{j-1} \binom{n}{s} q^s (1-q)^{n-s} \tag{25}$$

and the problem of selection of indexes (i,j) arises, so that

$$\sum_{s=i}^{j-1} \binom{n}{s} q^s (1-q)^{n-s} \geq \gamma.$$

An attempt of solving this problem was presented in the work of Zieliński (2011). In our research we assume that:

$$P\{X_{i:n} \leq F^{-1}(q) \leq X_{j:n}\} = \left(\frac{1}{2}\right)^n \sum_{s=1}^{j=1} \binom{n}{s} \approx \gamma.$$

Applications of other estimators are given in the monograph of Lehmann (1991).

4. Assessment of accuracy of position parameters estimation

Let us now follow the obtained results and assess the accuracy of statistical inference in the four models under consideration. The accuracy of inference will be assessed with the use of the width of confidence interval for μ. Obviously, it depends on the distribution F of error and on the size n of the sample $X_1, \dots, X_n$.

Confidence intervals of models (1) and (3) have a deterministic length depending only on n. Half of their length is denoted by D (1) and D (3),

respectively. Intervals (2) and (4) have a random length so for further consideration the expected values of their lengths will be taken and denoted by D (2) and D (4), respectively. Then, we get:

$$D(2) = t_{n-1}\left(\frac{1+\gamma}{2}\right)\frac{E(S)}{\sqrt{n-1}}$$

$$E(S) = \sqrt{\frac{2}{\pi}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

For $D(4)$ we get:

$$D(4) = \frac{1}{2}\left(E_{N(0,1)}X_{j:n} - E_{N(0,1)}X_{i:n}\right),$$

where by $E_{N(0,1)}X_{j:n}$ we denoted the expected value of the j-th position statistics from the sample $X_1, \ldots, X_n$, when the sample comes from the standard normal distribution $N(0,1)$.

Table 1. Assessment of accuracy of position parameters estimation

n	γ	$D(1)$	$D(2)$	$D(3)$	$D(4)$
15	0.90	0.424699	0.462405	0.524439	0.515701
	0.95	0.506061	0.563081	0.625379	0.714877
	0.99	0.665076	0.781524	0.823391	0.947689
25	0.90	0.328971	0.345613	0.408676	0.408597
	0.95	0.391993	0.416926	0.487204	0.463971
	0.99	0.515166	0.565007	0.641052	0.700479
30	0.90	0.300308	0.312812	0.373624	0.382351
	0.95	0.357839	0.376531	0.445383	0.473288
	0.99	0.470280	0.507456	0.585921	0.672498
50	0.90	0.232617	0.238289	0.290265	0.304216
	0.95	0.277180	0.285621	0.345961	0.356962
	0.99	0.364277	0.380902	0.454954	0.494328
100	0.90	0.164485	0.166455	0.205701	0.214301
	0.95	0.195996	0.198918	0.245140	0.252810
	0.99	0.257583	0.263298	0.322272	0.331143

Source: own calculations

The numbers included in Table 1 clearly show a great significance of both the choice of the statistical model and the statistics, that is the estimator of a suitable position parameter (expected value, median or an arbitrary quantile). The statistics serves as a basis for statistical inference on values which are of interest to the researcher. What is particularly striking are the differences in assessment of accuracy of position parameters for sample sizes $n \leq 30$.

5. Final remarks

In any statistical research we have a set statistical observations and some incomplete information about the distribution of these observations.

It is necessary to analyze the questions which we expect to answer by applying a suitable statistical procedure and the initial assumptions that have to be made so that our answers would be justified. A procedure dependent on some prior assumptions impossible to be verified by the observations collected or logically derived cannot be applied here. Statistical methods, therefore, should be treated not as a tool for a given detailed model but rather as an assisting tool to interpret data for different models.

This article presents certain problems connected with the choice of the procedure appropriate for the assumed statistical model along with the verification of its assumptions on the one hand, and the assessment of the data set and their distribution on the other. It is very important to analyze the behaviour of statistical procedures in very varied conditions.

Acknowledgment

The author would like to thank the anonymous referees for suggestions, which helped to improve the text in several points.

REFERENCES

DOMAŃSKI, CZ., PEKASIEWICZ, D., BASZCZYŃSKA, A., WITASZCZYK, A., (2014). Testy statystyczne w procesie podejmowania decyzji, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

LEHMANN, E.L., (1991). Teoria estymacji punktowej, Wydawnictwo Naukowe PWN, Warszawa.

NEYMAN, J., (1935). On the problem of confidence intervals, The Annals of Mathematical Statistics 6, p. 111.

NEYMAN, J., (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Phil. Trans. Royal Soc. London, A 236, p. 333

STEIN, C., (1956). Efficient nonparametric testing and estimation, Proc. Third Berkeley Symp. Math. Statist. Probab. 1, pp. 187–195.

ZIELIŃSKI, R., (2011). Statystyka matematyczna stosowana. Elementy, Centrum Studiów Zaawansowanych Politechniki Warszawskiej, Warszawa.

STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 559–562, DOI 10. 21307

ABOUT THE AUTHORS

Beck Krzysztof, PhD, is an economist and econometrician, received PhD in Economics from the Poznań University of Economics, recipient of two Lazarski Rector's Award for outstanding accomplishments in the field of economics. Assistant professor, head of Applied Economics Institute, researcher, and lecturer at Lazarski University. Author and co-author of more than thirty scientific articles, books and book chapters in the area of macroeconomics, macroeconomic policy, international economics, monetary unions, international macroeconomics, business cycle synchronization, business cycle analysis, model averaging and Bayesian econometrics. Participant of several research and application projects, with national and international scope. Research interest include: macroeconomics, international economics, international macroeconomics, international macroeconomic policy coordination, international trade, monetary union, business cycle synchronization, spectral analysis, robust econometrics, Bayesian econometrics and time series analysis.

Domański Czesław is a Professor at the Department of Statistical Methods in University of Lodz, Poland. His research interests are multivariate statistical analysis, construction of tests based on the theory of runs and order statistics construction of statistical tables for selected non-parametric statistics based on exact distributions and recursive formulas, non-classical methods of statistical inference including the Bayes and bootstrap analysis and non-parametric inference, analysis of properties of multivariate normality tests, small area statistics, medical statistics, statistical methods on capital and insurance market, tests based on stochastic processes. Professor Domański has published more than 220 research papers in international/national journals and conferences. An author or co-author of 22 books including 15 monographs. He is an active member of many scientific professional bodies.

Domitrz Adrian graduated from Faculty of Economic Sciences, University of Warsaw. His master thesis focused on subjective poverty and income micro-simulation topics. Currently engaged in practical usage of econometrics in a media agency, developing marketing mix modelling and attribution modelling methods.

Dudek Andrzej, dr hab., is an associate professor of Wrocław University of Economics. His main field(s) of research interest is: classification and data analysis, multivariate statistical analysis, symbolic data analysis. Additional info: Maintainer and co-author of R packages: clusterSim, symbolicDA and mdsOpt.

Górecki Tomasz received his MSc in mathematics from the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznań, Poland, in 2001. There he received his PhD in 2005. He obtained habilitation in computer science from Systems Research Institute Polish Academy of Sciences in 2015. Currently, he is an assistant professor at this University. His research interests include machine learning, times series classification and data mining.

Hindls Richard is a Professor of Statistics. In 2001-2006 he was the Dean of the Faculty of Informatics and Statistics of the University of Economics in Prague, and in 2006-2014 he was the President of the University of Economics in Prague. He focuses on economic aggregates, the analysis of time series and the application of statistical methods in auditing. He has published 40 books and about 250 articles. He is active in many institutions and scientific councils (Czech Statistical Council, Association of National Accounting Paris, European Advisory Committee for Economic and Social Statistics and others). In 2012, he received the Medal for the development of cooperation between Poland and the Czech Republic in the area of statistics (awarded by the Polish Statistical Society on the occasion of its 100th anniversary).

Hronová Stanislava is a Professor at the Department of Economic Statistics in University of Economics, Prague. From 2001 to 2006 she was a Vice-Dean for Research of the Faculty of Informatics and Statistics, in 2006–2014 the Vice-President for Research of the University of Economics, Prague. From 2010 to 2014 she was a member of the Government Research and Development Council; now she is a Vice-president of the Czech Science Foundation. She is interested in national accounts and economics statistics. She has co-operated with the Czech Statistical Office in the area of statistical methodology. She was awarded the French Ordre des Palmes Académiques. She is an active member of many scientific bodies. She is an Editor-in-Chief of Statistika journal and a member of Editorial Board of Politická ekonomie journal and Silesian Statistical Review.

Lubos Marek, an associate professor of Statistics. He works as the Dean of the Faculty of Informatics and Statistics, University of Economics, Prague, the Czech Republic. He studied the Mathematical Statistics at Charles University, Prague. His main research field concentrates on data analysis, probability, stochastic processes and time series analysis. He worked as a member of some scientific projects. He worked 7 years as the Head of the research project „Methods of knowledge acquisition from data and their use in economic decision-making“. He is an author of several textbooks and monographs, many conference papers and journal articles. He is a member of Czech and International statistic societies, and member of several scientific boards.

Łuczak Maciej received MSc and PhD degrees in mathematics from the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznań Poland, in 2001 and 2005, respectively. He is currently an assistant professor at the Faculty of Civil Engineering, Environmental and Geodetic

Sciences of the Koszalin University of Technology, Poland. His research interests are in the area of machine learning, time series and evolutionary algorithms.

Morawski Leszek is an Associate Professor at Vistula University (Akademia Finansów i Bankowości) and Institute of Economics at Polish Academy of Sciences. For almost 20 years his main research interests have included financial motivation to work and tax and benefit policies. He is a co-author of the Polish tax and benefit microsimulation model "SIMPL" and the Polish module for the European tax and the benefit model EUROMOD. He is currently interested in issues related to the impact of demographic change on social policy and life satisfaction.

Pandey Ranjita is an Assistant Professor at the Department of Statistics, University of Delhi. She received her dr hab. degree from University of Allahabad in 2002. Her research interests include ecological modelling, demography, imputation methods and Bayesian inference.

Prasad S. is currently working as an assistant professor at the Department of Statistics, University College, Thiruvananthapuram, India and pursuing research under the guidance of Professor P. G. Sankaran in the area of regression modelling of lifetime data. His research interests include survival analysis and distribution theory. He has published some research papers in national/international journals. He is a member of several academic/professional bodies.

Sankaran P. G. is a Professor at the Department of Statistics, Cochin University of Science and Technology, Cochin, India. His research interests include survival analysis, reliability analysis and distribution theory. Professor Sankaran has published more than 130 research papers in international/national journals and conferences. He has also published three books/monographs. He has successfully supervised nine PhD students and is an active member in many scientific professional bodies. He is serving as a visiting scientist at many reputed international institutions and is a reviewer of many international journals.

Sieradzki Dominik is a PhD Student at the Department of Econometrics and Statistics of the Warsaw University of Life Sciences. His main research includes applications of quantitative methods in economics, in particular sample allocation in estimation of proportion. He is an author of several research papers.

Smaga Łukasz, PhD, received his MSc degree in mathematics from the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznań in 2009. There he obtained his PhD degree in 2013. Currently he is an assistant professor at this university. His research interests are in the area of mathematical and computational statistics and its applications. He has published over 20 papers and given about 30 talks at conferences and seminars.

Zieliński Wojciech is a Full Professor in the Department of Econometrics and Statistics of the University of Life Sciences in Warsaw. His main research area is classical inference and its applications, especially problems of interval estimation

and testing statistical hypotheses as well as robust inference. He has published more than 80 articles and textbooks.

Walesiak Marek is a Professor at Wroclaw University of Economics in Department of Econometrics and Computer Science. His main field(s) of research interest is classification and data analysis, multivariate statistical analysis, marketing research. Professional achievements: Head of Department of Econometrics and Computer Science (1997-) Dean of Faculty of Regional Economics and Tourism (since 2012 Faculty of Economics, Management and Tourism): 1996-2002, 2008-2016 Polish Statistical Association (PTS): the member (1992-) of The Council of Section on Classification and Data Analysis of Polish Statistical Society (SKAD): chairman (2009-2012), vice-chairman (2013-2018), secretary (1993-1998), the member (1999-2008) of International Federation of Classification Societies (IFCS) – the member of Council (2002-2006) The Committee of the Statistics and Econometrics of Polish Academy of Sciences (PAN) – the member (2003-2006, 2007-2010, 2011-2014, 2015-2018) Vice Editor-in-Chief of Polish Statistical Review „Przegląd Statystyczny" (2009-).

Yadav Kalpana is a research scholar in the Department of Statistics, University of Delhi. Her research area is imputation methods in sample surveys. She has participated in a number of workshops/conferences at national and international level. She has published several research papers in national and international journals.

Żądło Tomasz is employed as an associate professor at the Department of Statistics, Econometrics and Mathematics at the University of Economics in Katowice. His research interests focus on small area estimation, survey sampling and mixed models. He is an elected member of the International Statistical Institute, a country representative of the International Association of Survey Statisticians and an associate editor of "Mathematical Population Studies" published by Taylor & Francis.

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).