STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Sampling methods and estimation

Research articles

Other articles:

Multivariate Statistical Analysis 2015, Łódź. Conference Papers

Research Communicates and Letters

Conference Announcement

Volume 18, Number 1, March 2017

Editorial Office **ISSN 1234-7655**

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, March 2017
Vol. 18, No. 1, pp. 1−4

FROM THE EDITOR

In this first issue of the 18th volume of the journal, the reader will find few novelties and changes, some of which have already been announced, but some emerged as an unexpected event. The latter are marked by the mourning sign around the name of Sir Anthony Atkinson, who passed away on January 1st 2017. We were honoured to have Tony Atkinson among members of the Editorial Board. In research and analysis in 'economics and related disciplines', promotion of which belongs to one of the journal's key objectives, his name personified sensitivity to the data dimension as well as its relevance and quality, among other things. Professor A. B. Atkinson was one of the very first scholars awarded with the Jerzy Neyman medal, established on the occasion of the 100th anniversary of the Polish Statistical Association (in 2012).

On behalf of the journal's Editorial Office and its Editorial Board I am pleased to welcome Prof. Carl-Erik Särndal, Statistics Sweden, as a new member of the Editorial Board. We all are grateful to C-E Särndal for his generosity and willingness to share with us his renowned expertise, and by the same token to contribute to international prestige of the journal.

There are some changes in the panel of Associate Editors as well. I am pleased to welcome new collaborators who have accepted our invitation to serve as members of this body: Professor Eugeniusz Gatnar, who is a Member of the Monetary Policy Council of the National Bank of Poland, and Professor Waldemar Tarczyński, Dean of the Faculty of Economics and Management of the University of Szczecin.

Another type of change concerns the structure of the journal. With this issue onward it is extended by an additional section – *Research Communicates and Letters* – that is meant to serve as a platform for sharing new ideas or research results on problems and approaches which may still be under refinement or needs further verification or improvement. Introduction of such a section was motivated by increasingly felt need to create a kind of window for more and more often submitted texts which may not be fully matured yet, or have a character of practice-oriented voice in a debate on statistics, but deserve to be published for their originality and potential interest to other experts. Although such a form of facilitations seems to be especially suitable for younger researchers, there is no status or experience-related constrains envisioned to be used for the selection of texts for which this section is specifically devoted. For example, besides of strictly scientific papers some policy-oriented articles by experts in issues associated with different aspects of public statistics or infrastructure of statistical process are welcome.

This issue contains ten articles arranged in four sections, starting with *Sampling methods and estimation.* The first paper, by **Katarzyna Budny,** ***Estimation of the central moments of a random vector based on the definition of the power of a vector*** discusses the moments of a random vector based on the definition of the power of a vector (proposed by J. Tatar), i.e. as scalar and vector characteristics of a multivariate distribution, but in analogy to the univariate case. Som characteristics of a multivariate distribution, such as index of skewness and kurtosis, have been introduced by using the central moments of a random vector. This paper presents the consistent estimators of the central moments of a random vector, for which essential characteristics have been found, such as a mean vector and a mean squared error. In these formulas, the relevant orders of approximation have been taken into account.

In their paper ***On the performance of the biased estimators in a misspecified model with correlated regressors,*** **Shalini Chandra, Gargi Tyagi,** discuss the effect of misspecification due to omission of relevant variables on the dominance of the $r-(k,d)$ class estimator proposed by Özkale (2012) over the ordinary least squares (OLS) estimator, and some other competing estimators, when some of the regressors in the linear regression model are correlated. Using the mean squared error criterion they conducted a Monte Carlo simulation study with numerical example to compare the performance of the estimators for some selected values of the parameters involved. One of the conclusions is that the MSE of the estimators increases in the misspecified model as compared to the model assumed to be true.

Housila P. Singh, Vishal Mehta, in the paper on ***Improved estimation of the scale parameter for log-logistic distribution using balanced ranked set sampling*** suggest estimators under a situation where the units in a sample can be ordered by judgement method without any error. They have also suggested some linear shrinkage estimator of a scale parameter of LLD. Efficiency comparisons are also made in this work. In general, the estimator are more efficient than Lesitha and Thomas (2012) estimators and MMSE estimators.

The next section, *Research articles,* begins with the paper by **Marta Dziechciarz-Duda,** and **Anna Król,** ***An application of multivariate statistical analysis for the valuation of durable goods brands.*** The authors attempt to improve the process of analysing the position and value of brands using selected multivariate statistical analysis methods (hedonic regression, multidimensional scaling, classification and linear ordination methods). The measurement have been performed on two levels: the product level, in which the prices of branded products were compared, and the consumer level, where the perception and attitudes of consumers towards the brands were studied. The analyses have been carried out on two sets of data, which enabled fuller and more comprehensible understanding of decision rules that guide consumers in choosing the brand.

Henryk Gurgul's and **Artur Machno's** paper, ***Trade Pattern on Warsaw stock exchange and prediction of number of trade***, presents the method for describing and predicting trade intensity on the Warsaw Stock Exchange. Their approach is based on generalized linear models, the variable selection is performed using shrinkage methods such as the Lasso or Ridge regression. The variable under investigation is the number of trades of a particular stock 5-minute interval. The main conclusion is that the number of trades during short intervals is predictable in the sense that the prediction, even based on relatively simple models, is with respect to statistical properties better than the prediction based on the random walk, which is used as a benchmark model.

Jan Kordos, in the paper ***The challenges of the population census round of 2020. Outline of the methods of quality assessment of population census data,*** discusses the challenges and gives an outline of the methods of quality assessment of population census data. After a synthetic overview of two Eurostat documents (2007, 2009) and UN Statistics Division (2010) monograph relating to census data quality assessment, the three methods of census quality assessment are discussed: (i) demographic method, (ii) post-enumeration survey and (iii) comparison with existing household surveys. In this context, some Polish experience in these fields is also discussed, along with some suggestions for the 2021 Polish Census of Population and Housing.

The third section, *Other articles,* contains two papers based on presentations at the *Multivariate Statistical Analysis (MAS) Conference* (Łódź, 2015). In the first one **Mirosława Sztemberg-Lewandowska** discusses ***The achievements of students at the II-IV stages of education using functional principal component analysis.*** The research covers the average exam results received on graduation from the second, third and fourth stage of education. The level of knowledge of students at the subsequent stages of education in the period 2009-2015 was measured. Functional principal component analysis, which is based on functional data, is applied in the study to include dynamic data, showing both the tendency as well as the pace of changes in time.

In another post-MAS conference paper, ***The application of Buhlmann Straub model to the estimation of net premium rates depending on the age of the insured in the motor third liability insurance***, by **Anna Szymańska** the problem of the age of the insured in the tariff calculation of premiums in motor liability insurance is discussed. The article proposes a method of rate estimation of net premiums in the age groups of the motor third liability insurance portfolio of individuals using one of the maximum likelihood models, called the Bűhlmann-Straub model.

The new section, *Research Communicates and Letters*, is inaugurated by **Jacek Białek's** and **Elżbieta Roszko-Wójtowicz's** paper entitled ***Evaluation of the EU countries innovative potential – multivariate approach***. It is aimed at working out a synthetic measure for estimating country’s innovation potential (CIP) of EU economies using data from Eurostat and several indicators arranged into four areas of analysis: investment expenditure, education, labour market and effects. A tendency toward convergence among some of the ratings based on different indexes and methods was shown.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 5

SUBMISSION INFORMATION FOR AUTHORS

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

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

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

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

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

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

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

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

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 7–26

ESTIMATION OF THE CENTRAL MOMENTS OF A RANDOM VECTOR BASED ON THE DEFINITION OF THE POWER OF A VECTOR

Katarzyna Budny[1]

ABSTRACT

The moments of a random vector based on the definition of the power of a vector, proposed by J. Tatar, are scalar and vector characteristics of a multivariate distribution. Analogously to the univariate case, we distinguish the uncorrected and the central moments of a random vector. Other characteristics of a multivariate distribution, i.e. an index of skewness and kurtosis, have been introduced by using the central moments of a random vector. For the application of the mentioned quantities for the analysis of multivariate empirical data, it appears desirable to construct their respective estimators.

This paper presents the consistent estimators of the central moments of a random vector, for which essential characteristics have been found, such as a mean vector and a mean squared error. In these formulas, the relevant orders of approximation have been taken into account.

Key words: central moment of a random vector, estimator, multivariate distribution, power of a vector.

1. Introduction

One of the fundamental characteristics of the univariate random variable are the ordinary (raw, uncorrected) and the central moments (e.g. Shao, 2003, Jakubowski and Sztencel, 2004). Even order moments are measures of dispersion of the distribution of the random variable, the moments of odd order characterize their location. In the analysis of multivariate distributions the product moments (about zero), the central mixed moments or collections thereof, e.g. mean vector, covariance matrix, are considered as classical generalizations of the above quantities (e.g. Johnson, Kotz and Kemp, 1992; Fujikoshi, Ulyanov and Shimizu, 2010). The uncorrected and central moments of a random vector are also considered as the expectations of relevant Kronecker products of a random vector (e.g. Holmquist, 1988). Thus, from this definition, they are size vector quantities

[1] Cracow University of Economics. E-mail: budnyk@uek.krakow.pl.

(collections of product moments (about zero) and central mixed moments of the corresponding orders).

On the basis of the definition of the power of a vector, Tatar (1996, 1999) suggested multivariate generalizations of the uncorrected and the central moments of a random variable, which are different from the above. Let us recall the basic definitions.

Definition 1.1. [Tatar 1996, 1999] Let $(H, R, +, \cdot)$ be a Hilbert space with the inner product $\langle \cdot, \cdot \rangle$. For any vector $v \in H$ and for any number $r \in N_{\circ} = N \cup \{0\}$ the r-th power of the vector v is defined as follows:

$$v^{o} = 1 \in R \quad \text{ and } v^{r} = \begin{cases} v^{r-1} \cdot v & \text{for } r - \text{odd} \\ \langle v^{r-1}, v \rangle & \text{for } r - \text{even} \end{cases}.$$

Let $L_k^r(\Omega)$ be a space of random vectors whose absolute value raised to the r-th power has finite integral, that is:

$$L_k^r(\Omega) = \left\{ \mathbf{X} : \Omega \to R^k : \int_{\Omega} \|\mathbf{X}\|^r \, dP < +\infty \right\}.$$

In the literature, the measure $E\left[\|\mathbf{X}\|^r\right] = \int_{\Omega} \|\mathbf{X}\|^r \, dP$ is sometimes called the moment of the order r of a random vector $\mathbf{X}$ and designated as $E\left[\mathbf{X}^r\right]$ (see Bilodeau and Brenner, 1999). Tatar (2002), however, by analogy with the univariate case, defines this expression as the absolute moment of order r of a random vector $\mathbf{X}$. In this study, we will also regard these values as the absolute moments of a random vector.

Therefore, let us assume that for the vector $\mathbf{X} : \Omega \to R^k$ an absolute moment of order r exists.

Definition 1.2. [Tatar 1996, 1999] The ordinary (raw, uncorrected) moment of order r of the random vector $\mathbf{X} : \Omega \to R^k$ is defined as

$$\alpha_{r,k} = E\left[\mathbf{X}^r\right]. \tag{1.1}$$

Let us note that the uncorrected moment of the first order is the mean vector, that is:

$$\alpha_{1,k} = m = E\mathbf{X}.$$

The definition and basic properties of the central moments of a random vector based on the definition of the power of a vector will be presented in the next section.

2. The central moments of a random vector

Definition 2.1. [Tatar 1996, 1999] The central moment of order r of a random vector $\mathbf{X}:\Omega\to R^k$, for which the absolute moment of order r exists, is given as

$$\mu_{r,k}=E\left[(\mathbf{X}-E\mathbf{X})^r\right]. \tag{2.1}$$

Remark 2.1. According to the concept of the power of the vector, if r is an even number, then $\alpha_{r,k},\mu_{r,k}\in R$, which means they are scalar quantities. However, if r is an odd number, then $\alpha_{r,k},\mu_{r,k}\in R^k$, so we obtain vectors.

Ordinary and central moments of a random vector based on the definition of the power of the vector are defined at an arbitrarily fixed inner product in the R^k space. In the following part of the paper we will consider the Hilbert space R^k where we define the Euclidean inner product $\langle v,w\rangle=\sum_{i=1}^{k}v_iw_i$ where $v=(v_1,...,v_k),\ w=(w_1,...,w_k)\in R^k$.

Remark 2.2. Let us observe that from the basic properties of the power of a vector and multinomial theorem we get the following formulas:

$$\mu_{2l,k}=E\left[\left(\sum_{i=1}^{k}(X_i-EX_i)^2\right)^l\right]=\sum_{l_1+...+l_m=l}\binom{l}{l_1,..,l_m}E\left[\prod_{t=1}^{k}(X_t-EX_t)^{2l_t}\right]$$

and

$$\mu_{2l+1,k}=E\left[\left(\sum_{i=1}^{k}(X_i-EX_i)^2\right)^l\cdot(X_1,...,X_k)\right]=$$

$$\left(\sum_{l_1+..+l_m=l}\binom{l}{l_1,..,l_m}E\left[\prod_{t=1}^{k}(X_t-EX_t)^{2l_t}(X_1-EX_1)\right],...,\sum_{l_1+..+l_m=l}\binom{l}{l_1,..,l_m}E\left[\prod_{t=1}^{k}(X_t-EX_t)^{2l_t}(X_k-EX_k)\right]\right)$$

where $\binom{l}{l_1,..,l_m}=\frac{l!}{l_1!\cdot...\cdot l_m!}$ is a multinomial coefficient.

By this inner product, the second order central moment is called the total variance of the random vector (see Bilodeau and Brenner, 1999) or the variance of the random vector (see Tatar 1996, 1999). According to these terms, $D^2\mathbf{X}$ will denote the central moment of the second order of a random vector.

Let us recall that the variance of a random vector was used to present multivariate generalization of Chebyshev's inequality (see Osiewalski and Tatar, 1999).

By using the central moments, characteristics of a multivariate distribution such as index of skewness and kurtosis have also been introduced.

Definition 2.2. [Tatar, 2000] The index of skewness of a random vector $\mathbf{X}:\Omega\to R^k$, for which there is an absolute moment of third order, is called

$$\gamma_{1,k}(\mathbf{X})=\frac{\mu_{3,k}}{(\mu_{2,k})^{\frac{3}{2}}}=\frac{E\left[(\mathbf{X}-E\mathbf{X})^3\right]}{(D^2\mathbf{X})^{\frac{3}{2}}}. \qquad (2.2)$$

Definition 2.3. [Budny and Tatar, 2009, Budny, 2009] Kurtosis of a random vector

$\mathbf{X}:\Omega\to R^k$, for which an absolute moment of fourth order exists, is a quantity defined as

$$\beta_{2,k}(\mathbf{X})=\frac{\mu_{4,k}}{\mu_{2,k}^2}=\frac{E\left[(\mathbf{X}-E\mathbf{X})^4\right]}{(D^2\mathbf{X})^2}. \qquad (2.3)$$

Assuming further that the random vector $\mathbf{N}:\Omega\to R^k$ has a multivariate normal distribution, we obtain the form

$$\beta_{2,k}(\mathbf{N})=1+\frac{2\sum_{i=1}^{k}\sigma_i^4+2\sum_{\substack{i,j=1\\ i\neq j}}^{k}\rho_{ij}^2\sigma_i^2\sigma_j^2}{\sum_{i,j=1}^{k}\sigma_i^2\sigma_j^2}. \qquad (2.4)$$

The formula (2.4) was determined (Budny, 2012) using Isserlis theorem (Isserlis, 1919), setting the algorithm for determination of the central mixed moments of the normally distributed random vector.

For the application of the central moments for the analysis of multivariate, empirical data, it appears desirable to construct their respective estimators. The next section will present their form along with a discussion of basic properties.

3. The multivariate sample central moments

3.1. Construction and basic properties

At the beginning let us recall the form of multivariate sample raw moments with their basic properties useful in the next part of this paper (Budny, 2014).

Suppose that $\mathbf{X}^1:\Omega\to R^k,\ldots,\mathbf{X}^n:\Omega\to R^k$ is a random sample from multivariate distribution, i.e. a set of n independent, identically distributed (i.i.d.) random vectors, with a finite $r-$th absolute moment.

Definition 3.1.1. The multivariate sample raw moment of order r (the estimator of the $r-\mathrm{th}$ raw moment of a random vector) is called:

$$a_{r,k}=\frac{\sum_{i=1}^{n}\left(\mathbf{X}^{i}\right)^{r}}{n}. \tag{3.1.1}$$

Multivariate sample raw moments are consistent and unbiased estimators, and their central moments satisfy the condition

$$E\left[\left(a_{r,k}-\alpha_{r,k}\right)^{2s}\right]=O\left(n^{-s}\right), \tag{3.1.2}$$

for all $r,s\in N$.

Let us therefore proceed to formulate the forms of estimators of the central moments of a random vector.

Definition 3.1.2. The multivariate sample central moment of order r (the estimator of the $r-\mathrm{th}$ central moment of a random vector) is defined as

$$m_{r,k}=\frac{\sum_{i=1}^{n}\left(\mathbf{X}^{i}-\overline{\mathbf{X}}\right)^{r}}{n}. \tag{3.1.3}$$

Remark 3.1.1. According to the definition of the power of a vector: if r is an even number, then $m_{r,k}$ is a univariate random variable, and if r is an odd number, then a random vector is obtained as $m_{r,k}$.

Remark 3.2.1. In the following discussion, while examining the properties of multivariate sample central moments, we will assume, without loss of generality, that the mean vector is a zero vector, i.e. $\alpha_{1,k}=m=0$.

We will begin the analysis of the properties of estimators of the central moments of a random vector by determining the form of their expected values. To do this, we will first introduce some forms of multivariate sample central moments, useful in further considerations.

Theorem 3.1.1. Multivariate sample central moments can be represented as follows:

- estimator of central moments of even order:

$$m_{2s,k}=a_{2s,k}+\sum_{p=1}^{s}\binom{s}{p}a_{2s-2p,k}\overline{\mathbf{X}}^{2p}-2\sum_{p=1}^{s}p\binom{s}{p}\left\langle a_{2s-(2p-1),k},\overline{\mathbf{X}}^{2p-1}\right\rangle+$$
$$+\frac{\sum_{i=1}^{n}\sum_{p=2}^{s}\sum_{l=0}^{p-2}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}\left(\mathbf{X}^{i}\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}}{n}, \tag{3.1.4}$$

- estimator of central moments of odd order:

$$m_{2s+1,k} =$$

$$= a_{2s+1,k} + \sum_{p=1}^{s}\binom{s}{p}\overline{\mathbf{X}}^{2p}\cdot a_{2s+1-2p,k} - \frac{2\sum_{i=1}^{n}\sum_{p=1}^{s} p\binom{s}{p}\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle \overline{\mathbf{X}}^{2p-2}\cdot\left(\mathbf{X}^i\right)^{2s+1-2p}}{n} +$$

$$+\frac{\sum_{i=1}^{n}\sum_{p=2}^{s}\sum_{l=0}^{p-2}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}\overline{\mathbf{X}}^{2l}\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\left(\mathbf{X}^i\right)^{2s+1-2p}}{n} - m_{2s,k}\cdot\overline{\mathbf{X}}\,. \qquad (3.1.5)$$

Proof: It is easily seen that from the definition of the power of a vector, we get above formulas.

The computation leading to explicit forms of the expected value of multivariate sample central moments is tedious and does not bring any relevant elements for further consideration. So, the next theorem will present their form with appropriate order of approximation (for univariate case - see Cramér 1958, p. 336). Prior to the formulation of this result, let us consider the following lemma.

Lemma 3.1.1. Assume that $r \in N\setminus\{0\}$. Let us consider a multivariate distribution (in population) for which the absolute moment of order $2r$ exists. Then, for every $t\in\{1,...r-1\}$ we get

$$E\left[a_{r-t,k}\circ\overline{\mathbf{X}}^t\right] = O\left(n^{-\frac{t}{2}}\right), \qquad (3.1.6)$$

where the operator "$\circ$" is defined (Tatar, 2008) as follows:

$$v^k\circ w^l = \begin{cases} v^k w^l & \text{for } k,l\text{ - even} \\ v^k\cdot w^l & \text{for } k-\text{even, } l-\text{odd} \\ w^l\cdot v^k & \text{gdy } k-\text{odd, } l-\text{even} \\ \left\langle v^k, w^l\right\rangle & \text{gdy } k,l\text{ - odd}\end{cases}.$$

Proof: see Appendix.

Property $(3.1.6)$ will play a key role in the study of property of unbiasedness of multivariate sample central moments. It will be used in the proof of the theorem, which presents forms of their expected values with the relevant order of approximation.

Theorem 3.1.2. Under the assumptions of Lemma 3.1.1.

$$E\left[m_{r,k}\right] = \mu_{r,k} + O\left(n^{-1}\right). \qquad (3.1.7)$$

Proof: First, let us consider multivariate sample central moments of even orders. Towards $(3.1.4)$ we get:

$$E[m_{2s,k}] = E[a_{2s,k}] + \sum_{p=1}^{s}\binom{s}{p}E\left[a_{2s-2p,k}\overline{\mathbf{X}}^{2p}\right] - 2sE\left[\left\langle a_{2s-1,k},\overline{\mathbf{X}}\right\rangle\right] +$$

$$-2\sum_{p=2}^{s}p\binom{s}{p}E\left[\left\langle a_{2s-(2p-1),k},\overline{\mathbf{X}}^{2p-1}\right\rangle\right] +$$

$$+\frac{\sum_{i=1}^{n}\sum_{p=2}^{s}\sum_{l=0}^{p-2}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}E\left[\left(\mathbf{X}^{i}\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left\langle \mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\right]}{n}. \qquad (3.1.8)$$

By the assumption $\alpha_{1,k} = m = 0$, an easy computation shows that

$$E\left[\left\langle a_{2s-1,k},\overline{\mathbf{X}}\right\rangle\right] = \frac{\mu_{2s,k}}{n}, \qquad (3.1.9)$$

$$E\left[a_{2s,k}\cdot\overline{\mathbf{X}}\right] = \frac{\mu_{2s+1,k}}{n}. \qquad (3.1.10)$$

Note that for each $i \in \{1,...,n\}$ and $t \in N$ we obtain the following property

$$E\left[\left|\left\langle \mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle\right|^{t}\right] = O\left(n^{-\frac{t}{2}}\right). \qquad (3.1.11)$$

Indeed, due to Minkowski's inequality in the $L_1^t(\Omega)$, Schwarz's inequality in $L_1^2(\Omega)$ and property $(3.1.2)$, applied to the coordinates of univariate sample mean (Cramér, 1958, p. 332), we get:

$$E\left[\left|\left\langle \mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle\right|^{t}\right] = \left(\left(E\left[\left|\sum_{j=1}^{k}X_j^i\overline{X}_j\right|^{t}\right]\right)^{\frac{1}{t}}\right)^{t} \le \left(\sum_{j=1}^{k}\left(E\left[\left|X_j^i\overline{X}_j\right|^{t}\right]\right)^{\frac{1}{t}}\right)^{t} \le$$

$$\le \left(\sum_{j=1}^{k}\left(\left(E\left[\left(X_j^i\right)^{2t}\right]\right)^{\frac{1}{2}}\left(E\left[\left(\overline{X}_j\right)^{2t}\right]\right)^{\frac{1}{2}}\right)^{\frac{1}{t}}\right)^{t} \le \left(\sum_{j=1}^{k}\left(\sqrt{E\left[\left(X_j^i\right)^{2t}\right]}\left(\frac{A}{n^{t}}\right)^{\frac{1}{2}}\right)^{\frac{1}{t}}\right)^{t} =$$

$$= \left(\sum_{j=1}^{k}\left(\sqrt{E\left[\left(X_j^i\right)^{2t}\right]A}\right)^{\frac{1}{t}}\cdot n^{-\frac{1}{2}}\right)^{t} = \frac{C}{n^{\frac{t}{2}}},$$

for each $n \ge n_0$ where $A, C > 0$.

Let us also note that by Jensen's inequality and Hölder's inequality, after taking into account properties $(3.1.2)$ and $(3.1.11)$we have:

$$\left|E\left[\left(\mathbf{X}^i\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\right]\right| \le E\left[\left(\mathbf{X}^i\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left|\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle\right|^{p-l}\right] \le$$

$$\le \left(E\left[\left(\left(\mathbf{X}^i\right)^2\right)^s\right]\right)^{\frac{s-p}{s}}\left(E\left[\left(\overline{\mathbf{X}}^2\right)^s\right]\right)^{\frac{l}{s}}\left(E\left[\left|\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle\right|^s\right]\right)^{\frac{p-l}{s}} \le \frac{C}{n^{\frac{p+l}{2}}}$$

for each $i \in \{1,...,n\}$, $p \in \{2,...,s\}$, $l \in \{0,..., p-2\}$ and $n \ge n_0$ where $C > 0$.

Clearly, this leads to the conclusion that

$$\frac{\sum_{i=1}^{n}\sum_{p=2}^{s}\sum_{l=0}^{p-2}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}E\left[\left(\mathbf{X}^i\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\right]}{n} = O\left(n^{-1}\right). \quad (3.1.12)$$

Finally, the use of the properties in the following order: unbiasedness of multivariate sample raw moments, $(3.1.6)$, $(3.1.9)$ and $(3.1.2)$, to the equation $(3.1.8)$ implies

$$E\left[m_{2s,k}\right]=\mu_{2s,k}+O\left(n^{-1}\right). \quad (3.1.13)$$

In order to determine the form of the expected value of the multivariate sample central moments of odd orders, let us note that

$$E\left[m_{2s+1,k}\right]= E\left[a_{2s+1,k}\right]+\sum_{p=1}^{s}\binom{s}{p}E\left[\overline{\mathbf{X}}^{2p}\cdot a_{2s+1-2p,k}\right]+$$

$$-\frac{2\sum_{i=1}^{n}\sum_{p=1}^{s}p\binom{s}{p}E\left[\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle\overline{\mathbf{X}}^{2p-2}\cdot\left(\mathbf{X}^i\right)^{2s+1-2p}\right]}{n}+$$

$$+\frac{\sum_{i=1}^{n}\sum_{p=2}^{s}\sum_{l=0}^{p-2}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}E\left[\overline{\mathbf{X}}^{2l}\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\left(\mathbf{X}^i\right)^{2s+1-2p}\right]}{n}-E\left[m_{2s,k}\cdot\overline{\mathbf{X}}\right].$$

$$(3.1.14)$$

At the beginning we shall show that for each $i \in \{1,...,n\}$, $p \in \{2,...,s\}$ and $l \in \{0,..., p-2\}$ there is a property

$$E\left[\overline{\mathbf{X}}^{2l}\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\left(\mathbf{X}^i\right)^{2s+1-2p}\right]=O\left(n^{-\left(\frac{p+l}{2}\right)}\right). \quad (3.1.15)$$

Indeed, Jensen's and Hölder's inequalities, the properties $(3.1.2)$ and $(3.1.11)$ imply

$$\left\| E\left[\overline{\mathbf{X}}^{2l}\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\left(\mathbf{X}^i\right)^{2s+1-2p}\right]\right\|^2 \le E\left[\overline{\mathbf{X}}^{4l}\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{2(p-l)}\left(\mathbf{X}^i\right)^{4s+2-4p}\right]\le$$

$$\le\left(E\left[\left(\overline{\mathbf{X}}^2\right)^{2s-p+l+1}\right]\right)^{\frac{2l}{2s-p+l+1}}\left(E\left[\left(\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^2\right)^{2s-p+l+1}\right]\right)^{\frac{p-l}{2s-p+l+1}}\left(\mu_{4s-2p+2l+2,k}\right)^{\frac{2s+1-2p}{2s-p+l+1}}\le$$

$$\le\left(\mu_{4s-2p+2l+2,k}\right)^{\frac{2s+1-2p}{2s-p+l+1}}\left(\frac{A_1}{n^{2s-p+l+1}}\right)^{\frac{2l}{2s-p+l+1}}\left(\frac{A_2}{n^{2s-p+l+1}}\right)^{\frac{p-l}{2s-p+l+1}}=\frac{C}{n^{p+l}},$$

for each $n\ge n_0$ where $A_1,A_2,C>0$, which obviously is equivalent to $(3.1.15)$.

This implies, therefore, a property

$$\frac{\sum_{i=1}^{n}\sum_{p=2}^{s}\sum_{l=0}^{p-2}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}E\left[\overline{\mathbf{X}}^{2l}\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\left(\mathbf{X}^i\right)^{2s+1-2p}\right]}{n}=O\left(n^{-1}\right).$$

$(3.1.16)$

Note that the reasoning analogous to the one carried out above leads to another property expressed as

$$E\left[\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle\overline{\mathbf{X}}^{2p-2}\cdot\left(\mathbf{X}^i\right)^{2s+1-2p}\right]=O\left(n^{-\left(p-\frac{1}{2}\right)}\right), \qquad (3.1.17)$$

for each $i\in\{1,...,n\}$ and $p\in\{1,...,s\}$.

Furthermore, the elementary computation establishes the equality

$$E\left[\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle\cdot\left(\mathbf{X}^i\right)^{2s-1}\right]=\frac{\mu_{2s+1,n}}{n}, \qquad (3.1.18)$$

for every $i\in\{1,...,n\}$.

Owing to the conditions $(3.1.17)$ and $(3.1.18.)$ we get the property

$$\frac{\sum_{i=1}^{n}\sum_{p=1}^{s}p\binom{s}{p}E\left[\left\langle \mathbf{X}^i,\overline{\mathbf{X}}\right\rangle\overline{\mathbf{X}}^{2p-2}\cdot\left(\mathbf{X}^i\right)^{2s+1-2p}\right]}{n}=O\left(n^{-1}\right). \qquad (3.1.19)$$

The proof of the theorem will be completed by determining the order of approximation of a quantity $E\left[m_{2s,k}\cdot\overline{\mathbf{X}}\right]$, that takes the form

$$E\left[m_{2s,k}\cdot\overline{\mathbf{X}}\right]=E\left[a_{2s,k}\cdot\overline{\mathbf{X}}\right]+\frac{\sum_{i=1}^{n}\sum_{p=1}^{s}\sum_{l=0}^{p}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}E\left[\left(\mathbf{X}^{i}\right)^{2(s-p)}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\overline{\mathbf{X}}^{2l+1}\right]}{n}$$

Reasoning analogous to that shown in the proof of the property $(3.1.15)$ justifies the expression

$$E\left[\left(\mathbf{X}^{i}\right)^{2(s-p)}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\overline{\mathbf{X}}^{2l+1}\right]=O\left(n^{-\left(\frac{p+l+1}{2}\right)}\right) \qquad (3.1.20)$$

for each $i\in\{1,...,n\}$, $p\in\{1,...,s\}$ and $l\in\{0,...,p\}$. Thus, we get the condition

$$\frac{\sum_{i=1}^{n}\sum_{p=1}^{s}\sum_{l=0}^{p}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}E\left[\left(\mathbf{X}^{i}\right)^{2(s-p)}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\overline{\mathbf{X}}^{2l+1}\right]}{n}=O\left(n^{-1}\right),$$

which, together with $(3.1.10)$ leads to the property

$$E\left[m_{2s,k}\cdot\overline{\mathbf{X}}\right]=O\left(n^{-1}\right) \qquad (3.1.21)$$

Finally, after the consecutive application of unbiasedness of multivariate sample raw moments and the forms $(3.1.6)$, $(3.1.19)$, $(3.1.16)$ and $(3.1.21)$ to the equation $(3.1.14)$ we get

$$E\left[m_{2s+1,k}\right]=\mu_{2s+1,k}+O\left(n^{-1}\right), \qquad (3.1.22)$$

which completes the proof of the theorem.

Corollary 3.1.1 Multivariate sample central moments are asymptotically unbiased estimators of the central moments of a random vector.

3.2. Consistency of the multivariate sample central moments

Our considerations in this section will focus on finding orders of approximations of mean squared errors of multivariate sample central moments, i.e. quantities of the form $E\left[\left(m_{r,k}-\mu_{r,k}\right)^{2}\right]$.

At the beginning we will take into account even-order sample central moments. Let us note that the expression $E\left[\left(m_{2s,k}-\mu_{2s,k}\right)^2\right]$ from the formula $(3.1.4)$ can be presented in the following form:

$$E\left[\left(m_{2s,k}-\mu_{2s,k}\right)^2\right]=E\left[\left(a_{2s,k}-\mu_{2s,k}\right)^2\right]+$$
$$+\frac{\sum_{i=1}^{n}\sum_{p=1}^{s}\sum_{l=0}^{p}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}E\left[\left(a_{2s,k}-\mu_{2s,k}\right)\left(\mathbf{X}^{i}\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\right]}{n}+$$
$$+\frac{\sum_{i_1,i_2=1}^{n}\sum_{p_1,p_2=1}^{s}\sum_{l_1=0}^{p_1}\sum_{l_2=0}^{p_2}E\left[\prod_{z=1}^{2}\binom{s}{p_z}\binom{p_z}{l_z}(-2)^{p_z-l_z}\left(\mathbf{X}^{i_z}\right)^{2(s-p_z)}\overline{\mathbf{X}}^{2l_z}\left\langle\mathbf{X}^{i_z},\overline{\mathbf{X}}\right\rangle^{p_z-l_z}\right]}{n}.$$

$(3.2.1)$

Based on the property $(3.1.2)$, we have

$$E\left[\left(a_{2s,k}-\mu_{2s,k}\right)^2\right]=O\left(n^{-1}\right). \qquad (3.2.2)$$

Let us also note that

$$E\left[\left(a_{2s,k}-\mu_{2s,k}\right)\left(\mathbf{X}^{i}\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\right]=O\left(n^{-\frac{1+p+l}{2}}\right), \qquad (3.2.3)$$

for each $i\in\{1,...,n\}$, $p\in\{1,...,s\}$, $l\in\{0,...,p\}$.

In fact, considering Schwarz's inequality and Hölder's inequality we get

$$E^2\left[\left(a_{2s,k}-\mu_{2s,k}\right)\left(\mathbf{X}^{i}\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\right]\leq$$
$$\leq E\left[\left(a_{2s,k}-\mu_{2s,k}\right)^2\right]E\left[\overline{\mathbf{X}}^{4l}\left(\mathbf{X}^{i}\right)^{4(s-p)}\left\langle\mathbf{X}^{i_z},\overline{\mathbf{X}}\right\rangle^{2(p-l)}\right]\leq$$
$$\leq E\left[\left(a_{2s,k}-\mu_{2s,k}\right)^2\right]\left(E\left[\overline{\mathbf{X}}^{2d}\right]\right)^{\frac{2l}{d}}\left(E\left[\left(\mathbf{X}^{i}\right)^{2d}\right]\right)^{\frac{2(s-p)}{d}}\left(E\left[\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{2d}\right]\right)^{\frac{p-l}{d}},$$

where $d=2s-(p-l)$.

Owing to the properties $(3.1.2)$, $(3.1.11)$ and the condition $m=\alpha_{1,k}=0$, we get

$$E^2\left[\left(a_{2s,k}-\mu_{2s,k}\right)\left(\mathbf{X}^{i}\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\right]\leq$$
$$\leq\frac{A_1}{n}\left(\frac{A_2}{n^d}\right)^{\frac{2l}{d}}\left(\mu_{2d,k}\right)^{\frac{2(s-p)}{d}}\left(\frac{A_3}{n^d}\right)^{\frac{p-l}{d}}\leq\frac{C}{n^{1+p+l}},$$

where $A_1, A_2, A_3, C > 0$, which justifies $(3.2.3)$.

Property $(3.2.3)$ leads to condition

$$\frac{\sum_{i=1}^{n}\sum_{p=1}^{s}\sum_{l=0}^{p}\binom{s}{p}\binom{p}{l}(-1)^{p-l}2^{p-l}E\left[\left(a_{2s,k}-\mu_{2s,k}\right)\left(\mathbf{X}^{i}\right)^{2(s-p)}\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\right]}{n}=O\left(n^{-1}\right). \tag{3.2.4}$$

Reasoning analogous to that shown above justifies the expression

$$E\left[\prod_{z=1}^{2}\left(\mathbf{X}^{i_z}\right)^{2(s-p_z)}\overline{\mathbf{X}}^{2l_z}\left\langle\mathbf{X}^{i_z},\overline{\mathbf{X}}\right\rangle^{p_z-l_z}\right]=O\left(n^{-\frac{p_1+p_2+l_1+l_2}{2}}\right), \tag{3.2.5}$$

for each $i_1, i_2 \in \{1,...,n\}$, $p_1, p_2 \in \{1,...,s\}$, $l_1 \in \{0,..., p_1\}$, $l_2 \in \{0,..., p_2\}$.

Therefore,

$$\frac{\sum_{i_1,i_2=1}^{n}\sum_{p_1,p_2=1}^{s}\sum_{l_1=0}^{p_1}\sum_{l_2=0}^{p_2}E\left[\prod_{z=1}^{2}\binom{s}{p_z}\binom{p_z}{l_z}(-2)^{p_z-l_z}\left(\mathbf{X}^{i_z}\right)^{2(s-p_z)}\overline{\mathbf{X}}^{2l_z}\left\langle\mathbf{X}^{i_z},\overline{\mathbf{X}}\right\rangle^{p_z-l_z}\right]}{n}=O\left(n^{-1}\right). \tag{3.2.6}$$

Taking into account the form $(3.2.1)$ and the properties $(3.2.2)$, $(3.2.4)$ and $(3.2.6)$ we get

$$E\left[\left(m_{2s,k}-\mu_{2s,k}\right)^2\right]=O\left(n^{-1}\right). \tag{3.2.7}$$

Now, we will take into consideration the estimators of central moments of odd order. Let us note that on the basis of $(3.1.5)$, the expression $\left(m_{2s+1,k}-\mu_{2s+1,k}\right)^2$ can be presented as the relevant linear combination of the components of the form $\left(a_{2s+1,k}-\mu_{2s+1,k}\right)^2$,

$$\left\langle\left(a_{2s+1,k}-\mu_{2s+1,k}\right),\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\left(\mathbf{X}^{i}\right)^{2(s-p)+1}\right\rangle,$$

$$\left\langle\left(a_{2s+1,k}-\mu_{2s+1,k}\right),\left(\mathbf{X}^{i}\right)^{2(s-p)}\left\langle\mathbf{X}^{i},\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\overline{\mathbf{X}}^{2l+1}\right\rangle,$$

$$\left\langle\overline{\mathbf{X}}^{2l_1}\left\langle\mathbf{X}^{i_1},\overline{\mathbf{X}}\right\rangle^{p_1-l_1}\cdot\left(\mathbf{X}^{i_1}\right)^{2(s-p_1)+1},\left(\mathbf{X}^{i_2}\right)^{2(s-p_2)}\left\langle\mathbf{X}^{i_2},\overline{\mathbf{X}}\right\rangle^{p_2-l_2}\cdot\overline{\mathbf{X}}^{2l_2+1}\right\rangle.$$

Moving on to determine the order of approximation of the expected values of these components, we note that $(3.1.2)$ implies

$$E\left[\left(a_{2s+1,k}-\mu_{2s+1,k}\right)^2\right] = O\left(n^{-1}\right). \tag{3.2.8}$$
$$\left\|E\left[\left(a_{2s+1,k}-\mu_{2s+1,k}\right)^2\right]\right\| \le E\left[\left(a_{2s+1,k}-\mu_{2s+1,k}\right)^2\right].$$

Carrying out further analysis we, in turn, apply Jensen's, Schwarz's and Hölder's inequalities and, thanks to them, we obtain

$$\left\|E\left[\left\langle\left(a_{2s+1,k}-\mu_{2s+1,k}\right),\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\left(\mathbf{X}^i\right)^{2(s-p)+1}\right\rangle\right]\right\|^2 \le$$
$$E\Big[\big[\left(a_{2s+1,k}-\mu_{2s+1,k}\right)^2\overline{\mathbf{X}}^{4l}\left\langle\mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{2(p-l)}\left(\mathbf{X}^i\right)^{4(s-p)+2}\Big]\le$$
$$\le\left(E\left[\left(a_{2s+1,k}-\mu_{2s+1,k}\right)^{2d}\right]\right)^{\frac{1}{d}}\left(E\left[\left(\overline{\mathbf{X}}\right)^{2d}\right]\right)^{\frac{2l}{d}}\left(E\left[\left\langle\mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{2d}\right]\right)^{\frac{(p-l)}{d}}\left(E\left[\left(\mathbf{X}^i\right)^{2d}\right]\right)^{\frac{2(s-p)+1}{d}},$$

where $d=2s+2-(p-l)$.

This estimation, after taking into account the properties $(3.1.2)$, $(3.1.11.)$ and the fact that $m=\alpha_{1,k}=0$ leads to

$$\left\|E\left[\left\langle\left(a_{2s+1,k}-\mu_{2s+1,k}\right),\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\left(\mathbf{X}^i\right)^{2(s-p)+1}\right\rangle\right]\right\|^2 \le$$
$$\le\left(\frac{A_1}{n^d}\right)^{\frac{1}{d}}\left(\frac{A_2}{n^d}\right)^{\frac{2l}{d}}\left(\frac{A_3}{n^d}\right)^{\frac{p-l}{d}}\left(\mu_{2d,k}\right)^{\frac{2(s-p)+1}{d}}\le\frac{C}{n^{1+p+l}},$$

which is equivalent to the property

$$E\left[\left\langle\left(a_{2s+1,k}-\mu_{2s+1,k}\right),\overline{\mathbf{X}}^{2l}\left\langle\mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\left(\mathbf{X}^i\right)^{2(s-p)+1}\right\rangle\right]=O\left(n^{-\frac{1+p+l}{2}}\right),$$

for each $i\in\{1,...,n\}$, $p\in\{1,...,s\}$, $l\in\{0,...,p\}$.

A slight change in the proof above shows that

$$E\left[\left\langle\left(a_{2s+1,k}-\mu_{2s+1,k}\right),\left(\mathbf{X}^i\right)^{2(s-p)}\left\langle\mathbf{X}^i,\overline{\mathbf{X}}\right\rangle^{p-l}\cdot\overline{\mathbf{X}}^{2l+1}\right\rangle\right]=O\left(n^{-\frac{2+p+l}{2}}\right),$$

for each $i\in\{1,...,n\}$, $p\in\{1,...,s\}$, $l\in\{0,...,p\}$ and

$$E\left[\left\langle \overline{\mathbf{X}}^{2l_1}\left\langle \mathbf{X}^{i_1},\overline{\mathbf{X}}\right\rangle^{p_1-l_1}\cdot\left(\mathbf{X}^{i_1}\right)^{2(s-p_1)+1},\left(\mathbf{X}^{i_2}\right)^{2(s-p_2)}\left\langle \mathbf{X}^{i_2},\overline{\mathbf{X}}\right\rangle^{p_2-l_2}\cdot\overline{\mathbf{X}}^{2l_2+1}\right\rangle\right]=O\left(n^{-\frac{2+p_1+p_2+l_1+l_2}{2}}\right)$$

for each $i_1,i_2\in\{1,\dots,n\}$, $p_1,p_2\in\{1,\dots,s\}$, $l_1\in\{0,\dots,p_1\}$, $l_2\in\{0,\dots,p_2\}$.

The above properties leads, thus, to

$$E\left[\left(m_{2s+1,k}-\mu_{2s+1,k}\right)^2\right]=O\left(n^{-1}\right). \qquad (3.2.9)$$

The next theorem will include the summary of the above consideration.

Theorem 3.2.1. Assume that for any natural number $r\geq 2$, a quantity $E\left[\left(m_{r,k}-\mu_{r,k}\right)^2\right]$ exists. Then, the mean squared error of the multivariate sample central moment of order r is of the form

$$\mathrm{MSE}m_{r,k}=E\left[\left(m_{r,k}-\mu_{r,k}\right)^2\right]=O\left(n^{-1}\right).$$

From this theorem and multivariate generalization of Chebyshev's inequality (Osiewalski and Tatar, 1999) we obtain a very important corollary.

Corollary 3.2.1 The multivariate sample central moments are consistent estimators of the central moments of a random vector.

4. Numerical illustration

Multivariate sample central moments, and also their respective functions, can be useful tools for the analysis of multivariate empirical data, for instance for the multivariate financial items, which have been considered in Budny and Tatar (2014).

Budny, Szklarska and Tatar (2014) have presented an analysis of socio-demographic conditions of Poland taking into account the multivariate (four-dimensional) data from the Central Statistical Office of Poland from 2012, and administrative division of the country into counties (*powiats*). The coordinates of the data are: the total marriage rate, the total divorce rate, the total fertility rate and the total mortality rate.

For this data, the second, third and fourth multivariate sample central moments in selected regions of Poland are as follows.

- North region (72 counties of the voivodeships: West Pomeranian, Kuyavian.

Pomernian, Pomeranian, Warmian-Masurian, without city counties):

$$m_{2,k} = 1{,}4744, \qquad m_{3,k} = \begin{bmatrix} \text{-}0.0022 \\ \text{-}0.1227 \\ 0.0171 \\ \text{-}0.3623 \end{bmatrix}, \qquad m_{4,k} = 5.3820.$$

- East region (74 counties of the voivodeships: Podlaskie, Lublin, Subcarpathian.

Lesser Poland, without city counties):

$$m_{2,k} = 2.8700, \qquad m_{3,k} = \begin{bmatrix} \text{-}0.3286 \\ 0.2943 \\ \text{-}0.0531 \\ 3.3462 \end{bmatrix}, \qquad m_{4,k} = 27.7891.$$

- West region (97 counties of the voivodeships: Greater Poland, Lubusz, Lower.

Silesian, Silesian, Opole, without city counties):

$$m_{2,k} = 1.8058, \qquad m_{3,k} = \begin{bmatrix} -0.2345 \\ 0.1702 \\ -0.0448 \\ 1.1132 \end{bmatrix}, \qquad m_{4,k} = 9.1024.$$

- Central region (71 counties of the voivodeships: Masovian, Świętokrzyskie, Łódź, without city counties):

$$m_{2,k} = 2.3571, \qquad m_{3,k} = \begin{bmatrix} -0.3748 \\ 0.1529 \\ 0.0188 \\ -1.4188 \end{bmatrix}, \qquad m_{4,k} = 14.0671.$$

- City counties (65 counties):

$$m_{2,k} = 2.9811, \qquad m_{3,k} = \begin{bmatrix} -0.0125 \\ -0.0828 \\ 0.00003 \\ 0.69907 \end{bmatrix}, \qquad m_{4,k} = 22.4378.$$

Dispersion of the distribution of the vector of the socio-demographic situation of Poland is measured by the central moments of the even orders. Note that the highest dispersion of test vector was observed in city counties while the smallest in the counties of the northern region (see Budny, Szklarska, Tatar, 2014). Central moments of odd order of multivariate distribution are parameters of location.

They can also be considered as a measure of the asymmetry (see Tatar, 2000) and as a vector measures indicate the direction of asymmetry. The above considerations can be supplemented (see Budny, Szklarska, Tatar, 2014) using the functions of the central moments of a random vector, e.g. index of skewness (Tatar 2000) and kurtosis (Budny, Tatar 2009, Budny 2009). Let us mention that the index of skewness shows the direction of asymmetry while its square informs also about the size of the asymmetry.

The problem of estimation of this characteristics of multivariate distribution is left for further study.

5. Conclusions

In this paper we have proposed consistent and asymptotically unbiased estimators of the central moments of a random vector based on the power of a vector. Essential characteristics such as mean vectors and mean squared errors with the relevant orders of approximation have been established for them. The central moments of even order are parameters of dispersion of the distribution of a random vector. The moments of odd order characterize its location. These quantities can be useful tools for the analysis of multivariate data.

Acknowledgement

Publication was financed from the funds granted to the Faculty of Finance and Law at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

REFERENCES

BILODEAU, M., BRENNER, D., (1999). Theory of Multivariate Statistics. Springer-Verlag, New York.

BUDNY, K., (2009). Kurtoza wektora losowego. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. Ekonometria (26), Vol. 76, s. 44–54. [Kurtosis of a random vector. Research Papers of Wrocław University of Economics. Ekonometria (26), Vol. 76, pp. 44–54].

BUDNY, K., (2012). Kurtoza wektora losowego o wielowymiarowym rozkładzie normalnym. w: S. FORLICZ (red.). Zastosowanie metod ilościowych w finansach i ubezpieczeniach. CeDeWu, Warszawa, s. 41–54. [Kurtosis of normally distributed random vector. in: S. FORLICZ (ed.). The application of quantitative methods in finance and insurance. CeDeWu, Warsaw, pp. 41–54].

BUDNY, K., (2014). Estymacja momentów zwykłych wektora losowego opartych na definicji potęgi wektora. Folia Oeconomica Cracoviensia, Vol. 55, s. 81–96. [Estimation of the raw moments of a random vector based on the definition of the power of a vector. Folia Oeconomica Cracoviensia, Vol. 55, pp. 81–96].

BUDNY, K., TATAR, J., (2009). Kurtosis of a random vector – special types of distributions. Statistics in Transiton - new series, Vol. 10, No. 3, pp. 445–456.

BUDNY, K., TATAR, J., (2014). Charakterystyki wielowymiarowych wielkości finansowych oparte na definicji potęgi wektora. Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, Vol. 189, s. 27–39. [Characteristics of multivariate financial items based on definition of the power of a vector. Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, Vol. 189, pp. 27–39].

BUDNY, K., SZKLARSKA, M., TATAR, J., (2014). Wielowymiarowa analiza sytuacji społeczno-demograficznej Polski. Studia i Materiały "Miscellanea Oeconomicae", Rok 18, Nr 1/2014, s. 273–288. [Multivariate analysis of the socio-demographic situation of Poland, Studia i Materiały "Miscellanea Oeconomicae", Vol. 18, No. 1/2014, pp. 273–288].

CRÁMER, H., (1958). Metody matematyczne w statystyce. Wyd.1. PWN, Warszawa. [Mathematical Methods of Statistics. 1st ed. PWN, Warsaw].

FUJIKOSHI, Y., ULYANOV, V. V., SHIMIZU, R., (2010). Mutivariate Statistics: high-dimensional and large-sample approximations, John Wiley & Sons, Inc.

HOLMQUIST, B., (1988). Moments and cumulants of the multivariate normal distribution, Stochastic Analysis and Applications, 6, pp. 273–278.

ISSERLIS, L., (1918). On a Formula for the Product-Moment Coefficient of any Order of a Normal Frequency Distribution in any Number of Variables, Biometrika, Vol. 12, No. 1/2, Nov.

JAKUBOWSKI, J., SZTENCEL, R., (2004). Wstęp do rachunku prawdopodobieństwa. Wyd. 3. Script, Warszawa. [Introduction to Probability Theory. 3rd ed. Script, Warsaw].

JOHNSON, N. J., KOTZ, S., KEMP, A. W., (1992). Univariate discrete distributions:, 2nd ed. John Wiley & Sons, Inc.

OSIEWALSKI, J., TATAR, J., (1999). Multivariate Chebyshev inequality based on a new definition of moments of a random vector, Statistical Review, Vol. 46, No. 2, pp. 257–260.

SHAO, J., (2003). Mathematical statistics, 2nd ed. Springer.

TATAR, J., (1996). O niektórych miarach rozproszenia rozkładów prawdopodobieństwa. Przegląd Statystyczny, Vol. 43, No. 3–4, s. 267–274. [On certain measures of the diffusion of probability distribution. Statistical Review, Vol. 43, No. 3–4, pp. 267–274].

TATAR, J. (1999). Moments of a random variable in a Hilbert space. Statistical Review, Vol. 46, No. 2, pp. 261–271.

TATAR, J., (2000). Asymetria wielowymiarowych rozkładów prawdopodobieństwa. Materiały z XXXV Konferencji Statystyków, Ekonometryków i Matematyków Akademii Ekonomicznych Polski Południowej. [Asymmetry of multivariate probability distributions. Conference Proceedings, Cracow University of Economics].

TATAR, J., (2002). Nierówność Lapunowa dla wielowymiarowych rozkładów prawdopodobieństwa. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, Vol. 549, s. 5–10. [Lapunav's inequality for multivariate probability distribution. Cracow Review of Economics and Management, Vol. 549, pp.5–10].

TATAR, J., (2008). Miary zależności wektorów losowych o różnych wymiarach. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, Vol. 780, s. 53–60 [Measures of dependence of random vectors of different sizes. Cracow Review of Economics and Management, Vol. 780, pp. 53–60].

APPENDIX

Proof of lemma 3.11.: The property $(3.1.6)$ will be proved by considering the cases of even and odd powers of random vectors. Reasoning (in each case) will be based on Schwarz's inequality considered in the relevant Hilbert's space $L_k^2(\Omega)$ or $L_1^2(\Omega)$.

At the beginning, let $r = 2s$ and $t = 2p - 1$ where $p \in \{1,...s\}$. It is, therefore, necessary to prove the property

$$E\left[\left\langle a_{2s-(2p-1),k}, \overline{\mathbf{X}}^{2p-1}\right\rangle\right] = O\left(n^{-\left(p-\frac{1}{2}\right)}\right), \qquad (1)$$

For the proof we will apply Schwarz's inequality in Hilbert's space $L_k^2(\Omega)$ with the inner product $\langle \mathrm{X}, \mathrm{Y}\rangle_{L_k^2} := E[\langle \mathrm{X}, \mathrm{Y}\rangle]$ that leads to

$$E^2\left[\left\langle a_{2s-(2p-1),k}, \overline{\mathbf{X}}^{2p-1}\right\rangle\right] \le E\left[a^2_{2s-(2p-1),k}\right] E\left[\overline{\mathbf{X}}^{4p-2}\right].$$

Let us note that $E\left[a^2_{r,k}\right] = D^2 a_{r,k} + \left(E\left[a_{r,k}\right]\right)^2 = O(1)$. Thus, taking into account the property $(3.1.2)$ we get

$$E^2\left[\left\langle a_{2s-(2p-1),k}, \overline{\mathbf{X}}^{2p-1}\right\rangle\right] = O\left(n^{-(2p-1)}\right),$$

which is equivalent to the condition (1).

Now, we consider the even numbers $r = 2s$ and $t = 2p$ where $p \in \{1,...s-1\}$ In view of this, the property

$$E\left[a_{2s-2p,k}\overline{\mathbf{X}}^{2p}\right] = O\left(n^{-p}\right). \qquad (2)$$

requires the proof.

Note that Schwarz's inequality in the space $L_1^2(\Omega)$ of the property $(3.1.2)$ justifies the estimations

$$E^2\left[a_{2s-2p,k}\overline{\mathbf{X}}^{2p}\right] \le E\left[a^2_{2s-2p,k}\right] E\left[\overline{\mathbf{X}}^{4p}\right] \le \frac{C}{n^{2p}},$$

which obviously leads to (2).

In turn, for odd natural numbers $r = 2s - 1$ and $t = 2p - 1$ where $p \in \{1,...s-1\}$ it is necessary to prove the property

$$E\left[a_{2s-2p,k} \cdot \overline{\mathbf{X}}^{2p-1}\right] = O\left(n^{-\left(p-\frac{1}{2}\right)} \right). \qquad (3)$$

For this proof, let us note that Jensen's inequality, Schwarz's inequality and the condition $(3.1.2)$ imply a sequence of inequalities, respectively.

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 27–52

ON THE PERFORMANCE OF SOME BIASED ESTIMATORS IN A MISSPECIFIED MODEL WITH CORRELATED REGRESSORS

Shalini Chandra[1], **Gargi Tyagi**[2]

ABSTRACT

In this paper, the effect of misspecification due to omission of relevant variables on the dominance of the $r-(k,d)$ class estimator proposed by Özkale (2012), over the ordinary least squares (OLS) estimator and some other competing estimators when some of the regressors in the linear regression model are correlated, have been studied with respect to the mean squared error criterion. A simulation study and numerical example have been demostrated to compare the performance of the estimators for some selected values of the parameters involved.

Key words: omission of relevant variables, multicollinearity, $r-(k,d)$ class estimator, mean squared error.

1. Introduction

In multiple linear regression, the presence of multicollinearity inflates sampling variance of the ordinary least squares estimator and may also produce wrong signs of the estimator. Many authors have witnessed the presence of multicollinearity in the various fields of application, including Hamilton (1972), Mahajan *et al.* (1977), Heikkila (1988), Graham (2003), among others. To cope up with the problem of multicollinearity several alternative methods to the OLS have been proposed, *viz.* ordinary ridge regression (ORR) by Hoerl and Kennard (1970); principal component regression (PCR) by Massy (1965). In the hope that combining two estimators will contain the properties of both gave rise to the development of the $r-k$ class, the two-parameter class and the $r-(k,d)$ class estimators (see Baye and Parker (1984); Kaciranlar and Sakallioglu (2001); Özkale and Kaciranlar (2007) and Özkale (2012)).

[1] Department of Mathematics & Statistics, Banasthali Vidyapith, Banasthali - 304022 India. E-mail: chandrshalini@gmail.com.

[2] Department of Mathematics & Statistics, Banasthali Vidyapith, Banasthali - 304022 India. E-mail: tyagi.gargi@gmail.com.

The performance of these estimators have been evaluated under various comparison criteria like mean squared error (MSE), matrix MSE, Pitman's closeness criterion and the Mahalanobis loss function. Nomura and Ohkubo (1985) derived the dominance conditions of the $r-k$ class estimator over the OLS and ORR estimators and Sarkar (1996) obtained conditions of the superiority of the $r-k$ class estimator over the other estimators under matrix MSE criterion. Özkale and Kaciranlar (2008) compared the $r-k$ class estimator with the OLS estimator under Pitman's closeness criterion. Özkale (2012) proposed the $r-(k,d)$ class estimator and compared this estimator with the other biased estimators under the MSE criterion. Sarkar and Chandra (2015) studied the performance of the $r-(k,d)$ class estimator over the OLS, PCR and the two-parameter class estimator under the Mahalanobis loss function and derived tests to verify the conditions.

In these studies, it has been assumed inherently that the model is correctly specified. However, in practice, some of the relevant regressors may get excluded from the model, i.e. the model does not remain correctly specified, known as misspecified model. The omission of relevant regressors causes biased and inconsistent estimation. The effect of the omission of relevant regressors on the performance of the estimators have been studied by several authors, for example, Kadiyala (1986); Trenkler and Wijekoon (1989) and Wijekoon and Trenkler (1989). Although not much work has been done when some of the regressors are omitted and multicollinearity is also present, Sarkar (1989) studied the performance of the $r-k$ class estimator and compared it with the OLS, ORR and PCR estimators under MSE criterion when the model is misspecified due to omission of relevant regressors.

In this paper, misspecification due to omission of relevant regressors and multicollinearity have been studied simultaneously and the effect of misspecification on the dominance of the $r-(k,d)$ class estimator over the other biased estimators has been studied under the MSE criterion. The plan of this paper is as follows: in Section 2, the model and the estimators under study are given. Section 3 provides the comparison of the estimators and a Monte Carlo simulation has been given in Section 4. A numerical example is given in Section 5 to see the effect of misspecification on the estimators, which in turn exhibits the utility of the estimators. The paper is concluded in Section 6.

2. Model structure and the estimators

Let us consider the regression model as:

$$y = X\beta + Z\gamma + \varepsilon, \tag{2.1}$$

where y is an $n \times 1$ vector of dependent variable, X and Z are $n \times p$ and $n \times q$ full column rank matrices of regressors respectively such that $X'X$ and $Z'Z$ are ill-conditioned, $p+q<n$, β and γ are the corresponding $p \times 1$ and $q \times 1$ vectors of

parameters associated with X and Z, respectively. ε is an $n \times 1$ vector of disturbance term, and it is assumed that $\varepsilon \sim N(0, \sigma^2 I_n)$. Suppose that an investigator has unknowingly excluded regressors of Z matrix, thus the misspecified model is given by:

$$y = X\beta + u, \tag{2.2}$$

where $u = Z\gamma + \varepsilon$. Misspecification occurs when the investigator assumes the disturbance vector u to be normally distributed with mean vector 0 and variance $\sigma^2 I_n$.

Let us consider the following transformation for the model in (2.2):

$$y = XTT'\beta + u = X^*\alpha + u, \tag{2.3}$$

where $X^* = XT$, $T'\beta = \alpha$, $T = (t_1, t_2, \dots, t_p)$ is a $p \times p$ orthogonal matrix with $T'X'XT = \Lambda$ and $\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_p)$ is a $p \times p$ diagonal matrix of eigen values of $X'X$ matrix such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Now, let $T_r = (t_1, t_2, \dots, t_r)$ be $p \times r$ orthogonal matrix after deleting last $p - r$ columns from T matrix, where $r \leq p$. Thus, $T_r{}'X'XT_r = \Lambda_r$ where $\Lambda_r = diag(\lambda_1, \lambda_2, \dots, \lambda_r)$ and $T_{p-r}{}'X'XT_{p-r} = \Lambda_{p-r}$, where $\Lambda_{p-r} = diag(\lambda_{r+1}, \lambda_{r+2}, \dots, \lambda_p)$. Also, $T'T = T_r{}'T_r + T_{p-r}{}'T_{p-r}$ and let $N = \{1,2, \dots, r; r+1, \dots, p\}$ be a set of first p integers such that $N = \{N_r; N_{p-r}\}$ where $N_r = \{1,2, \dots, r\}$ and $N_{p-r} = \{r+1, r+2, \dots, p\}$.

Özkale (2012) introduced an estimator by grafting the two-parameter class estimator and the PCR estimator together, known as the $r - (k,d)$ class estimator to deal with the problem of multicollinearity. For the misspecified model in (2.2) the $r - (k,d)$ class estimator is given by:

$$\hat{\beta}_r(k,d) = T_r(T_r{}'X'XT_r + kI)^{-1}(T_r{}'X'y + kdT_{r\prime}\hat{\beta}_r) \quad k \geq 0,\ 0 < d < 1 \tag{2.4}$$

which can be rewritten as:

$$\hat{\beta}_r(k,d) = T_r S_r(k)^{-1} \Lambda_r^{-1} S_r(kd) T_{r\prime} X'y k \geq 0,\ 0 < d < 1, \tag{2.5}$$

where $S_r(k) = \Lambda_r + kI_r$ and $S_r(kd) = \Lambda_r + kdI_r$. This is a general estimator which includes the OLS, ORR, PCR, $r - k$ class and the two-parameter class estimators as its special cases as:

1. $\hat{\beta}_p(0,0) = \hat{\beta} = (X'X)^{-1}X'y$, is the OLS estimator,
2. $\hat{\beta}_p(k,0) = \hat{\beta}(k) = (X'X + kI)^{-1}X'y$, is the ORR estimator,
3. $\hat{\beta}_r(0,0) = \hat{\beta}_r = T_r(T_{r\prime}X'XT_r)^{-1}T_{r\prime}X'y$, is the PCR estimator,
4. $\hat{\beta}_r(k,0) = \hat{\beta}_r(k) = T_r(T_{r\prime}X'XT_r + kI)^{-1}T_{r\prime}X'y$, is the $r - k$ class estimator,
5. $\hat{\beta}_p(k,d) = \hat{\beta}(k,d) = (X'X + kI)^{-1}(X'y + kd\hat{\beta})$, is the two-parameter class estimator.

2.1. Properties of the estimator

From (2.5), the bias and the variance of $\hat{\beta}_r(k,d)$ can be obtained as:

$$\begin{aligned}\text{Bias}(\hat{\beta}_r(k,d)) = \big(k(d-1)T_rS_r(k)^{-1}T_r{}' - T_{p-r}T_{p-r}{}'\big)\beta + \\ T_rS_r(k)^{-1}\Lambda_r^{-1}S_r(kd)T_{r'}\delta \end{aligned} \tag{2.6}$$

where $\delta = X'Z\gamma$, and

$$\text{Var}(\hat{\beta}_r(k,d)) = \sigma^2 T_rS_r(k)^{-2}\Lambda_r^{-1}S_r(kd)^2T_r{}' \tag{2.7}$$

respectively.

It is clear from (2.6) and (2.7) that the bias of the $r-(k,d)$ class estimator increases due to omission of relevant regressors whereas the variance of the estimator is not affected by the misspecification.

Further, the MSE for an estimator $\tilde{\beta}$ of β is defined as:

$$MSE(\tilde{\beta}) = E(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta) = tr(\text{Var}(\tilde{\beta})) + \big[\text{Bias}\big(\tilde{\beta}\big)\big]'\big[\text{Bias}\big(\tilde{\beta}\big)\big] \tag{2.8}$$

By substituting (2.6) and (2.7) in (2.8) and on simplification, we get:

$$\begin{aligned} MSE(\hat{\beta}_r(k,d)) &= \sigma^2 tr[S_r(k)^{-1}S_r(kd)\Lambda_r^{-1}S_r(kd)S_r(k)^{-1}] \\ &+\beta'(k(1-d)T_rS_r(k)^{-1}T_r{}' + T_{p-r}T_{p-r}{}')(k(1-d)\,T_rS_r(k)^{-1}T_r{}' \\ &+ T_{p-r}T_{p-r}{}')\beta \\ &-2\beta'(k(1-d)T_rS_r(k)^{-1}T_r{}' + T_{p-r}T_{p-r}{}')T_rS_r(k)^{-1}\Lambda_r^{-1}S_r(kd)T_r{}'\delta \\ &+\delta'T_rS_r(k)^{-1}\Lambda_r^{-1}S_r(kd)S_r(k)^{-1}\Lambda_r^{-1}S_r(kd)T_{r'}\delta \end{aligned} \tag{2.9}$$

which can be rewritten as:

$$\begin{aligned} MSE(\hat{\beta}_r(k,d)) = \underbrace{\textstyle\sum_{i=1}^{r} \frac{\sigma^2(\lambda_i+kd)^2+k^2(d-1)^2\lambda_i\alpha_i^2}{\lambda_i(\lambda_i+k)^2} + \sum_{i=r+1}^{p}\alpha_i^2} \\ + \underbrace{\textstyle\sum_{i=1}^{r} \frac{(\lambda_i+kd)^2\eta_i^2-2k(1-d)\lambda_i(\lambda_i+kd)\alpha_i\eta_i}{\lambda_i^2(\lambda_i+k)^2}} \end{aligned} \tag{2.10}$$

where $T'\delta = \eta = \{\eta_1,\eta_2,\dots,\eta_p\}$. Following Özkale(2012), the first under-bracket is the MSE obtained when there is no misspecification and the second under-bracket is the contribution of omission of relevant regressors.

The MSE of other estimators can be obtained by substituting the suitable values of r, k and d in (2.10). From the risk expression in (2.10), it can be seen that the effect of omission of relevant regressors on the MSE values will depend on the sign of the second term. If $\alpha_i\eta_i$ is negative for all values of $i \in N_r$, the second term in (2.10) will be negative and thus the MSE of the $r-(k,d)$ class estimator will increase due to omission of relevant regressors. However, if $\alpha_i\eta_i$ is non-negative for some values of $i \in N_r$ no definite conclusion can be made regarding the effect of misspecification.

2.2. Optimum values of k and d

The selection of the unknown biasing parameters k and d in the $r-(k,d)$ class estimator is an important problem. The optimum values of k and d in $r-(k,d)$ class estimator can be obtained by minimizing the MSE of the estimator with respect to k and d. To find a pair (k,d) of optimum values of k and d, we will use the technique of maxima and minima in calculus.

Let the two-dimensional function $MSE(\hat{\beta}_r(k,d))$ have its minimum value at (k_0,d_0) and have a continuous partial derivative at this point, then $\frac{\partial MSE(\hat{\beta}_r(k_0,d_0))}{\partial k}=0$ and $\frac{\partial MSE(\hat{\beta}_r(k_0,d_0))}{\partial d}=0$. The points k_0 and d_0 can be found as follows:

On differentiating $MSE(\hat{\beta}_r(k,d))$ in (2.10) with respect to d keeping r and k fixed, we obtain

$$\frac{\partial MSE(\hat{\beta}_r(k,d))}{\partial d}=2k\sum_{i=1}^{r}\frac{\sigma^2(\lambda_i+kd)-k(1-d)\lambda_i\alpha_i^2+(\lambda_i+2kd-k)\alpha_i\eta_i+(\lambda_i+kd)\eta_i^2/\lambda_i}{\lambda_i(\lambda_i+k)^2} \qquad (2.11)$$

and equating (2.11) to zero, we get:

$$d_0=\frac{\sum_{i=1}^{r}\frac{k\alpha_i^2-\sigma^2}{(\lambda_i+k)^2}-\sum_{i=1}^{r}\frac{(\lambda_i-k)\alpha_i\eta_i+\eta_i^2}{\lambda_i(\lambda_i+k)^2}}{\sum_{i=1}^{r}\frac{k(\sigma^2+\lambda_i\alpha_i^2)}{\lambda_i(\lambda_i+k)^2}+\sum_{i=1}^{r}\frac{k(2\alpha_i\eta_i+\eta_i^2/\lambda_i)}{\lambda_i(\lambda_i+k)^2}}. \qquad (2.12)$$

Assuming that $\alpha_i\eta_i>0$ for all $i \in N_r$, if $k\alpha_i^2-\sigma^2>0$ and $\lambda_i>k$ for all $i \in N_r$, the upper bound of d_0 is given by

$$\frac{\sum_{i=1}^{r}(k\alpha_i^2-\sigma^2)/(\lambda_i+k)^2}{\sum_{i=1}^{r}k(\sigma^2+\lambda_i\alpha_i^2)/\lambda_i(\lambda_i+k)^2} \qquad (2.13)$$

which is the optimum value of d when there is no misspecification due to omission of relevant regressors. Thus, if $k{\alpha_i}^2-\sigma^2>0$ and $\lambda_i>k$ for all $i \in N_r$, the optimum value of d in the misspecified model is less than that in the case of no misspecification. Moreover, for d_0 to be a positive value $\lambda_i(k{\alpha_i}^2-\sigma^2)-(\lambda_i-k)\alpha_i\eta_i+\eta_i^2$ should be positive for $i \in N_r$.

Further, differentiating (2.10) with respect to k keeping r and d fixed, we obtain:

$$\frac{\partial MSE(\hat{\beta}_r(k,d))}{\partial k}=-2(1-d)\sum_{i=1}^{r}\frac{\sigma^2(\lambda_i+kd)-k(1-d)\lambda_i\alpha_i^2}{(\lambda_i+k)^3}$$
$$-2(1-d)\sum_{i=1}^{r}\frac{(\lambda_i-k+2kd)\alpha_i\eta_i+(\lambda_i+kd)\eta_i^2/\lambda_i}{(\lambda_i+k)^3} \qquad (2.14)$$

From (2.14) and (2.12), we have:

$$\frac{\partial MSE(\widehat{\beta}_r(k,d_0))}{\partial k}=\frac{-2\sum_{i=1}^{r}(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i)/(\lambda_i(\lambda_i+k))}{k\left[\sum_{i=1}^{r}(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)/(\lambda_i(\lambda_i+k)^2)\right]^2}$$

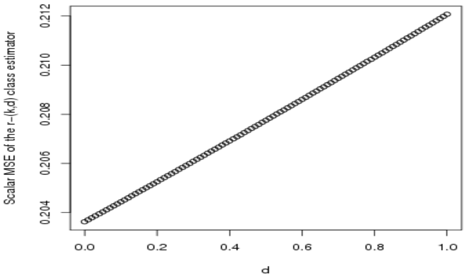
$$\times\left[\sum_{i=1}^{r}\frac{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)}{\lambda_i(\lambda_i+k)^2}\sum_{i=1}^{r}\frac{\lambda_i(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i)-k(\lambda_i\alpha_i^2+\alpha_i\eta_i)}{(\lambda_i+k)^3}\right.$$

$$\left.+\sum_{i=1}^{r}\frac{k(\lambda_i\alpha_i^2+\alpha_i\eta_i)-\lambda_i(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i)}{\lambda_i(\lambda_i+k)^2}\sum_{i=1}^{r}\frac{\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i}{(\lambda_i+k)^3}\right]. \quad (2.15)$$

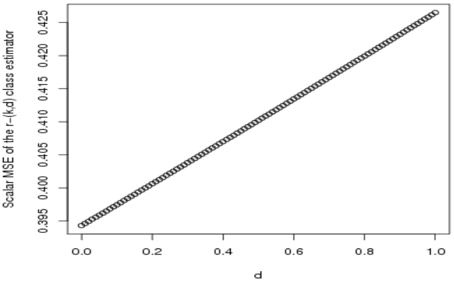
Clearly, $\frac{\partial MSE(\widehat{\beta}_r(k,d_0))}{\partial k}$ is zero, when

$$k_0=\frac{\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i}{\alpha_i^2+\alpha_i\eta_i/\lambda_i}, \quad fori=1,2,\dots,r. \quad (2.16)$$

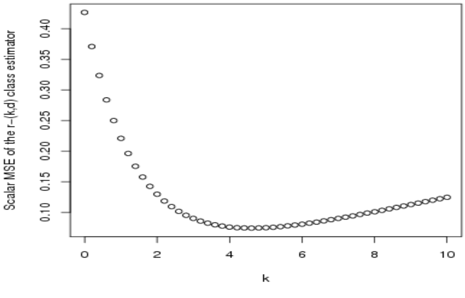
Then (k_0,d_0) is the expected point which minimizes $MSE(\hat{\beta}_r(k,d))$ where k_0 and d_0 are given as (2.16) and (2.12) respectively. However, when we substitute $k=k_0$ in (2.12) d_0 becomes zero. Therefore, a point (k_0,d_0) which satisfies $k>0$, $0<d<1$ and minimizes $MSE(\hat{\beta}_r(k,d))$ cannot be found (see Fig.1). In order to find an appropriate value of k and d, the behaviour of the MSE of the estimator at boundary points can be studied. This conclusion has been illustrated through the graph reported below:



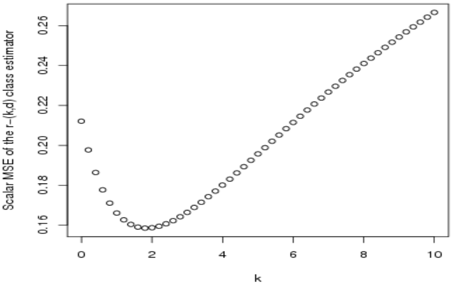
(a) In the case of no misspecification when $k=0.1$



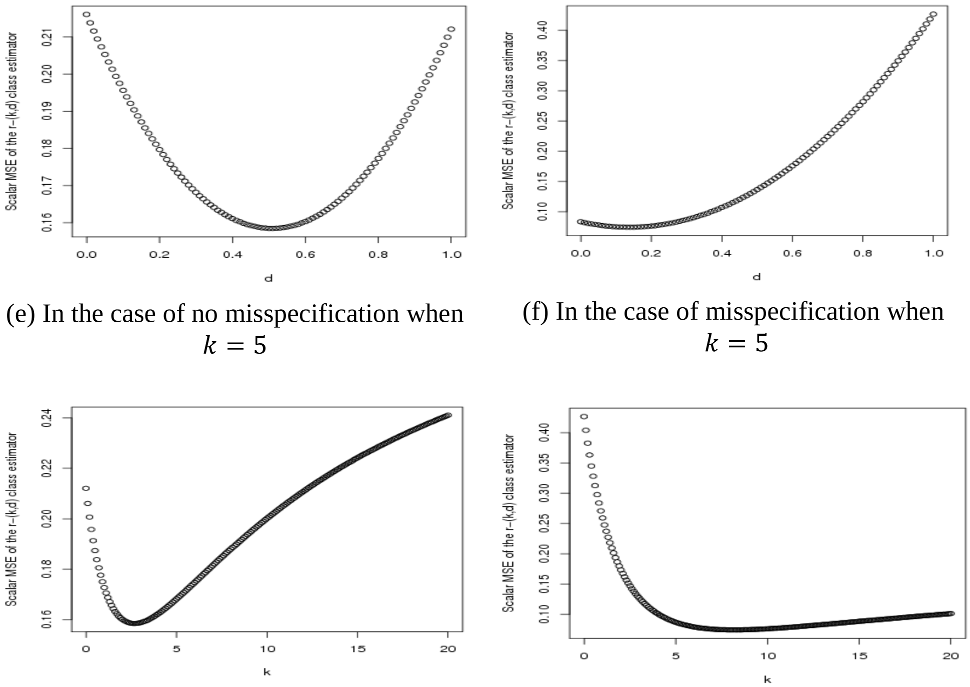
(b) In the case of misspecification when $k=0.1$



(c) In the case of no misspecification when $d=0.1$



(d) In the case of misspecification when $d=0.1$



(e) In the case of no misspecification when $k = 5$

(f) In the case of misspecification when $k = 5$

(g) In the case of no misspecification when $d = 0.3$

(h) In the case of misspecification when $d = 0.3$

Figure 1. MSE of the $r - (k, d)$ class estimator for the true and misspecified model

From Figure 1 the effect of misspecification on the optimum values of k and d for fixed values of d and k respectively can be observed, and, also the pair of values of k and d may not be found out for which the $r - (k, d)$ class estimator has minimum MSE. Further, we note that for the fixed values of d, the MSE of $\hat{\beta}_r(k, d)$ takes the minimum value for smaller value of k in the misspecified model when compared with the true model. However, for small value of k (see Fig. (a) and Fig. (b)), no variations are observed in the MSE values of $\hat{\beta}_r(k, d)$ for both the models, whereas for $k = 5$, the MSE of $\hat{\beta}_r(k, d)$ takes the minimum value for a smaller value of d in the misspecified model.

3. Comparison of the estimators under mse criterion

In this section, we compare the $r - (k, d)$ class estimator with other biased estimators when the model is misspecified due to omission of relevant regressors, and also study the effect of misspecification on the dominance conditions.

3.1. Comparison of the $r-(k,d)$ class estimator with the OLS estimator

The MSE of the OLS estimator in the misspecified model can be obtained by substituting $r=p, k=0$ in (2.10), as:

$$MSE(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} + \sum_{i=1}^{p} \frac{\eta_i^2}{\lambda_i^2}. \tag{3.1}$$

The difference of MSEs of the $r-(k,d)$ class estimator and the OLS estimator, say Δ_1, can be written as:

$$\begin{aligned}\Delta_1 &= MSE(\hat{\beta}) - MSE(\hat{\beta}_r(k,d)) \\ &= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} - \sum_{i=1}^{r} \frac{\sigma^2(\lambda_i+kd)^2 + k^2(1-d)^2\lambda_i\alpha_i^2}{\lambda_i(\lambda_i+k)^2} - \sum_{i=r+1}^{p} \alpha_i^2 \\ &+ \sum_{i=1}^{p} \frac{\eta_i^2}{\lambda_i^2} - \sum_{i=1}^{r} \frac{-2k(1-d)(\lambda_i+kd)\alpha_i\eta_i + (\lambda_i+kd)^2\eta_i^2/\lambda_i}{\lambda_i(\lambda_i+k)^2}.\end{aligned} \tag{3.2}$$

On further simplification, the difference can be rewritten as:

$$\begin{aligned}\Delta_1 = k(1-d)\sum_{i=1}^{r} \frac{[2\lambda_i(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i) + k((\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i) + d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)]}{\lambda_i(\lambda_i+k)^2} \\ + \sum_{i=r+1}^{p} \frac{\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i}{\lambda_i}.\end{aligned} \tag{3.3}$$

It is clear from the above expression that $\Delta_1 \geq 0$ that is, the $r-(k,d)$ class estimator dominates the OLS estimator, for all $k>0$, $0<d<1$ if $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i \geq 0$, for all $i \in N$. From (3.3), it can also be observed that when there is no misspecification due to omission of relevant regressors (i.e. $\eta_i = 0$ for all $i \in N$) the condition reduces to $\sigma^2-\lambda_i\alpha_i^2 \geq 0$ for all $i \in N$, which is the same as that of obtained by Ozkale (2012)). It is evident that due to addition of a positive term η_i^2/λ_i the odds for $\Delta_1 \geq 0$ are higher in the misspecified model.

Further, if $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i < 0,\ i \in N_r$ and $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i \geq 0, \forall\ i \in N_{p-r}$ then for a fixed k there exists a d in the range:

$$\frac{\sum_{i=1}^{r}\frac{k(\alpha_i^2\lambda_i-\sigma^2)-2\lambda_i\sigma^2}{\lambda_i(\lambda_i+k)^2} - \sum_{i=1}^{r}\frac{2\lambda_i\alpha_i\eta_i+(2\lambda_i+k)\eta_i^2\lambda_i}{\lambda_i(\lambda_i+k)^2}}{\sum_{i=1}^{r}\frac{k(\alpha_i^2\lambda_i+\sigma^2)}{\lambda_i(\lambda_i+k)^2} + \sum_{i=1}^{r}\frac{k(2\alpha_i\eta_i+\eta_i^2/\lambda_i)}{\lambda_i(\lambda_i+k)^2}} < d < 1 \tag{3.4}$$

such that the $r-(k,d)$ class estimator dominates the OLS estimator. If $\alpha_i\eta_i > 0$ for all $i \in N_r$, then the lower limit of d decreases due to omission of relevant regressors and thus the dominance range of the $r-(k,d)$ class estimator over the OLS estimator increases.

Furthermore, if $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i < 0$, for some $i = r+1, r+2, \ldots, p$, no definite conclusion can be drawn regarding dominance of one over the other. The results obtained are reported in the form of the following theorem.

Theorem 3.1

(i) If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i \geq 0$, for all $i \in N$, the $r-(k,d)$ class estimator dominates the OLS estimator for all $k > 0$ and $0 < d < 1$. The odds for superiority of $r-(k,d)$ class estimator over the OLS estimator increases in the misspecified model.

(ii) If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i < 0$, $i \in N_r$ and $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i \geq 0$, for all $i \in N_{p-r}$, the $r-(k,d)$ class estimator dominates the OLS estimator for all $k > 0$ and d such that it satisfies (3.4). The range of dominance of $r-(k,d)$ class estimator increases in the misspecified model provided $\alpha_i\eta_i$ is positive for all $i \in N_r$.

(iii) If $\sigma^2 - \lambda_i\alpha_i^2 + \frac{\eta_i^2}{\lambda_i} < 0$, for some $i = r+1, r+2, \ldots, p$, no definite conclusion can be drawn regarding their dominance.

3.2. Comparison of the $r-(k,d)$ class estimator with the ORR estimator

The MSE of the ORR estimator can be obtained by substituting $r = p$ and $d = 0$ in (2.10), given as:

$$MSE(\hat{\beta}(k)) = \sum_{i=1}^{p} \frac{\sigma^2\lambda_i + k^2\alpha_i^2}{(\lambda_i+k)^2} + \sum_{i=1}^{p} \frac{\eta_i^2 - 2k\alpha_i\eta_i}{(\lambda_i+k)^2}. \qquad (3.5)$$

Using (2.10) and (3.5), the difference between the MSEs,; say Δ_2, is given by:

$$\begin{aligned}\Delta_2 &= MSE(\hat{\beta}(k)) - MSE(\hat{\beta}_r(k,d))\\ &= \sum_{i=1}^{p} \frac{\sigma^2\lambda_i + k^2\alpha_i^2}{(\lambda_i+k)^2} - \sum_{i=1}^{r} \frac{\sigma^2(\lambda_i+kd)^2 + k^2(1-d)^2\lambda_i\alpha_i^2}{\lambda_i(\lambda_i+k)^2} - \sum_{i=r+1}^{p} \alpha_i^2\\ &+ \sum_{i=1}^{p} \frac{-2k\alpha_i\eta_i + \eta_i^2}{(\lambda_i+k)^2} - \sum_{i=1}^{r} \frac{-2k(1-d)(\lambda_i+kd)\alpha_i\eta_i + (\lambda_i+kd)^2\eta_i^2/\lambda_i}{\lambda_i(\lambda_i+k)^2}. \qquad (3.6)\end{aligned}$$

On further simplification, we obtain:

$$\begin{aligned}\Delta_2 &= kd\sum_{i=1}^{r} \frac{[k[\lambda_i\alpha_i^2 + 2\alpha_i\eta_i - d(\lambda_i\alpha_i^2 + 2\alpha_i\eta_i + \sigma^2 + \eta_i^2)] - 2\lambda_i(\sigma^2 + \alpha_i\eta_i + \eta_i^2)]}{\lambda_i(\lambda_i+k)^2}\\ &+ \sum_{i=r+1}^{p} \frac{(\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i) - 2k(\alpha_i^2 + \alpha_i\eta_i/\lambda_i)}{(\lambda_i+k)^2}. \qquad (3.7)\end{aligned}$$

From (3.7), it can be noticed that the $r-(k,d)$ class estimator dominates the ORR estimator if both summations are positive, that is:

$$k\left(2\lambda_i\alpha_i^2 + \alpha_i\eta_i\right) - kd\left(\sigma^2 + \lambda_i\alpha_i^2 + \alpha_i\eta_i + \lambda_i\eta_i^2\right) - 2\lambda_i\left(\sigma^2 + \alpha_i\eta_i + \lambda_i\eta_i^2\right) > 0 \quad \text{for all } i \in N_r \qquad (3.8)$$

and

$$\left(\sigma^2 - \lambda_i\alpha_i^2 + \frac{\eta_i^2}{\lambda_i}\right) - 2k\left(\alpha_i^2 + \frac{\alpha_i\eta_i}{\lambda_i}\right) > 0 \quad \text{forall } i \in N_{p-r}. \qquad (3.9)$$

If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i > 0$ for all $i \in N_{p-r}$ then (3.9) holds when:

$$k < \frac{\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i}{2(\alpha_i^2+\alpha_i\eta_i/\lambda_i)} \quad \text{forall } i \in N_{p-r}. \tag{3.10}$$

If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i < 0$ for all $i \in N_{p-r}$, then (3.9) does not hold true. However, if $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i < 0$ for some $i \in N_{p-r}$, a positive k can be found such that the second summation in Δ_2, i.e. $\sum_{i=r+1}^{p} \frac{(\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i)-2k(\alpha_i^2+\alpha_i\eta_i/\lambda_i)}{(\lambda_i+k)^2}$ is positive.

Further, from (3.8) we obtain:

$$d < \frac{k(2\lambda_i\alpha_i^2+\alpha_i\eta_i)-2\lambda_i(\sigma^2+\alpha_i\eta_i+\lambda_i\eta_i^2)}{k(\sigma^2+\lambda_i\alpha_i^2+\alpha_i\eta_i+\lambda_i\eta_i^2)} \quad \text{for all } i \in N_r. \tag{3.11}$$

For d to be a positive in (3.11) , $k(2\lambda_i\alpha_i^2 + \alpha_i\eta_i) - 2\lambda_i(\sigma^2 + \alpha_i\eta_i + \lambda_i\eta_i^2)$ should be a positive value for all $i \in N_r$, i.e.

$$k > \frac{\sigma^2+\alpha_i\eta_i+\lambda_i\eta_i^2}{\alpha_i^2+\alpha_i\eta_i/2\lambda_i} \quad \text{for all } i \in N_r. \tag{3.12}$$

If upper bound of d in (3.11) is greater than 1, any value smaller than 1 can be taken, which satisfies (3.11) and $0 < d < 1$.

The conditions of dominance of the $r-(k,d)$ class estimator over the ORR estimator under MSE criterion is stated below in the form of the following theorem:

Theorem 3.2

(i) If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i > 0$ for all $i \in N_r$ and $k > \frac{\sigma^2+\alpha_i\eta_i+\lambda_i\eta_i^2}{\alpha_i^2+\alpha_i\eta_i/2\lambda_i}$ for all $i \in N_r$, the $r-(k,d)$ class estimator dominates the ORR estimator if $k < \min_{i\in N_{p-r}} \frac{\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i}{2(\alpha_i^2+\alpha_i\eta_i/\lambda_i)}$ and $0 < d < min\left\{1, \min_{i\in N_r} \frac{k(2\lambda_i\alpha_i^2+\alpha_i\eta_i)-2\lambda_i(\sigma^2+\alpha_i\eta_i+\lambda_i\eta_i^2)}{k(\sigma^2+\lambda_i\alpha_i^2+\alpha_i\eta_i+\lambda_i\eta_i^2)}\right\}$.

(ii) If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i > 0$ for some $i \in N_{p-r}$ and $k > \frac{\sigma^2+\alpha_i\eta_i+\lambda_i\eta_i^2}{\alpha_i^2+\alpha_i\eta_i/2\lambda_i}$ for all $i \in N_r$, the $r-(k,d)$ class estimator dominates the ORR estimator for a value of k such that $\sum_{i=r+1}^{p} \frac{\lambda_i((\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i)-2k(\alpha_i^2+\alpha_i\eta_i/\lambda_i))}{(\lambda_i+k)^2}$ is positive and $0 < d < min\left\{1, \min_{i\in N_r} \frac{k(2\lambda_i\alpha_i^2+\alpha_i\eta_i)-2\lambda_i(\sigma^2+\alpha_i\eta_i+\lambda_i\eta_i^2)}{k(\sigma^2+\lambda_i\alpha_i^2+\alpha_i\eta_i+\lambda_i\eta_i^2)}\right\}$.

3.3. Comparison of the$r-(k,d)$ class estimator with the PCR estimator

On substituting $k=0$ in (2.10), the MSE of the PCR estimator can be obtained as:

$$MSE(\hat{\beta}_r)=\textstyle\sum_{i=1}^{r}\frac{\sigma^2}{\lambda_i}+\sum_{i=r+1}^{p}\alpha_i^2+\sum_{i=1}^{r}\frac{\eta_i^2}{\lambda_i^2}. \qquad (3.13)$$

From (2.10) and (3.13), the difference in MSEs, say Δ_3, is given by:

$$\Delta_3=MSE(\hat{\beta}_r)-MSE(\hat{\beta}_r(k,d))$$
$$=\textstyle\sum_{i=1}^{r}\frac{\sigma^2}{\lambda_i}+\sum_{i=r+1}^{p}\alpha_i^2-\sum_{i=1}^{r}\frac{\sigma^2(\lambda_i+kd)^2+k^2(d-1)^2\lambda_i\alpha_i^2}{\lambda_i(\lambda_i+k)^2}-$$
$$\textstyle\sum_{i=r+1}^{p}\alpha_i^2$$
$$+\textstyle\sum_{i=1}^{r}\frac{\eta_i^2}{\lambda_i^2}-\sum_{i=1}^{r}\frac{2k(1-d)\lambda_i(\lambda_i+kd)\alpha_i\eta_i-(\lambda_i+kd)^2\eta_i^2}{\lambda_i^2(\lambda_i+k)^2}.$$

On further simplifying it, we get:

$$\Delta_3=\ k(1-d)\textstyle\sum_{i=1}^{r}\frac{\left[2\lambda_i\left(\sigma^2+\alpha_i\eta_i+\frac{\eta_i^2}{\lambda_i}\right)+k\left\{\left(\sigma^2-\lambda_i\alpha_i^2+\frac{\eta_i^2}{\lambda_i}\right)+d\left(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\frac{\eta_i^2}{\lambda_i}\right)\right\}\right]}{\lambda_i(\lambda_i+k)^2}. \qquad (3.14)$$

It can be observed from the above expression that if $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i>0$ for all $i\in N_r$, the $r-(k,d)$ class estimator dominates the PCR estimator for all $k>0$ and $0<d<1$. Evidently, the odds for superiority of the $r-(k,d)$ class estimator over the PCR estimator increases due to misspecification.

If $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i<0$ for all $i\in N_r$, then Δ_3 is positive when $2\lambda_i(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i)+k((\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i)+d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)$ is positive for all $i\in N_r$, which can be rewritten as

$$2\lambda_i(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i)-k((\lambda_i\alpha_i^2-\sigma^2-\eta_i^2/\lambda_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)>0\ \text{ for all }\ i\in N_r \qquad (3.15)$$

If $\left(\lambda_i\alpha_i^2-\sigma^2-\frac{\eta_i^2}{\lambda_i}\right)-d\left(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\frac{\eta_i^2}{\lambda_i}\right)<0,$ forall $i\in N_r$, i.e.

$$d>\frac{(\lambda_i\alpha_i^2-\sigma^2-\eta_i^2\lambda_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)}\quad \text{for all} i\in N_r \qquad (3.16)$$

then Δ_3 is positive for all $k>0$. It is noticeable that the lower limit of d decreases due to misspecification, thus a wider range for the dominance of the $r-(k,d)$ class estimator over the PCR estimator is obtained as compared with no misspecification.

Further, if

$$d < \frac{(\lambda_i\alpha_i^2-\sigma^2-\eta_i^2/\lambda_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)} \quad \text{forall } i \in N_r \tag{3.17}$$

holds, then Δ_3 is positive for k such that

$$k < \frac{2\lambda_i(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i)}{((\lambda_i\alpha_i^2-\sigma^2-\eta_i^2/\lambda_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)} \quad \text{forall } i \in N_r. \tag{3.18}$$

By rewriting (3.18), it is observed that the upper limit of k increases due to misspecification. Additionally, the upper limit of d decreases. Thus, due to misspecification a wider range of k for a shorter range of d in which the $r-(k,d)$ class estimator dominates the PCR estimator is obtained.

The comparisons can be concluded in the following theorem.

Theorem 3.3

(i) If $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i>0$ for all $i \in N_r$, the $r-(k,d)$ class estimator dominates the PCR estimator for all $k>0$ and $0<d<1$. The odds for superiority of the $r-(k,d)$ class estimator over the PCR estimator increases due to misspecification.

(ii) If $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i<0$ for all $i \in N_r$ and $\max_{i\in N_r}\{(\lambda_i\alpha_i^2-\sigma^2-\eta_i^2/\lambda_i)/(\lambda_i\alpha_i^2+\sigma^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)<d<1$, then the $r-(k,d)$ class estimator dominates the PCR estimator for all $k>0$. The dominance range of the $r-(k,d)$ class estimator over the PCR estimator increases due to misspecification.

(iii)If $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i<0$ for all $i \in N_r$ and $0<d<\min_{i\in N_r}\{(\lambda_i\alpha_i^2-\sigma^2-\eta_i^2/\lambda_i)/(\lambda_i\alpha_i^2+\sigma^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)$, then the $r-(k,d)$ class estimator dominates the PCR estimator if $k<\min_{i\in N_r}\{2\lambda_i(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i)/((\lambda_i\alpha_i^2-\sigma^2-\eta_i^2/\lambda_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i))\}$. The range of k increases while the range of d decreases due to misspecification, in which the $r-(k,d)$ class estimator dominates the PCR estimator.

3.4. Comparison of the $r-(k,d)$ class estimator with the $r-k$ class estimator

The MSE of the $r-k$ class estimator can be obtained by substituting $d=0$ in (2.10), given as:

$$MSE(\hat{\beta}_r(k)) = \sum_{i=1}^{r}\frac{\lambda_i\sigma^2+k^2\alpha_i^2}{(\lambda_i+k)^2}+\sum_{i=r+1}^{p}\alpha_i^2+\sum_{i=1}^{r}\frac{\eta_i^2-2k\alpha_i\eta_i}{(\lambda_i+k)^2}. \tag{3.19}$$

From (2.10) and (3.19), the difference between the MSEs, say Δ_4, is given by:

$$\begin{aligned}\Delta_4 &= MSE(\hat{\beta}_r(k)) - MSE(\hat{\beta}_r(k,d))\\ &= \sum_{i=1}^{r} \frac{\lambda_i\sigma^2+k^2\alpha_i^2}{(\lambda_i+k)^2} - \sum_{i=1}^{r} \frac{\sigma^2(\lambda_i+kd)^2+k^2(d-1)^2\lambda_i\alpha_i^2}{\lambda_i(\lambda_i+k)^2}\\ &+ \sum_{i=1}^{r} \frac{\eta_i^2-2k\alpha_i\eta_i}{(\lambda_i+k)^2} - \sum_{i=1}^{r} \frac{(\lambda_i+kd)^2\eta_i^2-2k(1-d)\lambda_i(\lambda_i+kd)\alpha_i\eta_i}{\lambda_i^2(\lambda_i+k)^2}. \qquad (3.20)\end{aligned}$$

On further simplification, we get:

$$\Delta_4 = kd\sum_{i=1}^{r} \frac{\left[k(2(\lambda_i\alpha_i^2-\alpha_i\eta_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i-\eta_i^2/\lambda_i))-2\lambda_i(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i)\right]}{\lambda_i(\lambda_i+k)^2}. \qquad (3.21)$$

From (3.21), the $r-(k,d)$ class estimator dominates the $r-k$ class estimator when

$$k\left(2\left(\lambda_i\alpha_i^2-\alpha_i\eta_i\right)-d\left(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i-\frac{\eta_i^2}{\lambda_i}\right)\right)-2\lambda_i\left(\sigma^2+\alpha_i\eta_i+\frac{\eta_i^2}{\lambda_i}\right)> 0 \quad \text{forall } i\in N_r. \qquad (3.22)$$

If $\lambda_i\alpha_i^2-\alpha_i\eta_i>0$ for all $i\in N_r$ and $2(\lambda_i\alpha_i^2-\alpha_i\eta_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i-\eta_i^2/\lambda_i)>0$ for all $i\in N_r$, that is:

$$d<\frac{2(\lambda_i\alpha_i^2-\alpha_i\eta_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i-\eta_i^2/\lambda_i)} \quad \text{forall } i\in N_r \qquad (3.23)$$

then Δ_4 is positive when k is such that:

$$k>\frac{2\lambda_i(\sigma^2+\alpha_i\eta_i+\eta_i^2/\lambda_i)}{(2(\lambda_i\alpha_i^2-\alpha_i\eta_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i-\eta_i^2/\lambda_i))} \quad \text{forall } i\in N_r. \qquad (3.24)$$

If $\lambda_i\alpha_i^2-\alpha_i\eta_i>0$ for all $i\in N_r$ and $2(\lambda_i\alpha_i^2-\alpha_i\eta_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i-\eta_i^2/\lambda_i)<0$ for all $i\in N_r$, that is:

$$d>\frac{2(\lambda_i\alpha_i^2-\alpha_i\eta_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i-\eta_i^2/\lambda_i)} \quad \text{forall } i\in N_r \qquad (3.25)$$

then Δ_4 is negative for all $k>0$.

If $\lambda_i\alpha_i^2-\alpha_i\eta_i<0$ for all $i\in N_r$, then Δ_4 is negative for all $k>0$ and $0<d<1$.

The results obtained are given in the following theorem.

Theorem 3.4

(i) If $\lambda_i\alpha_i^2-\alpha_i\eta_i>0$ for all $i\in N_r$ and $0<d<\min_{i\in N_r}\left\{\frac{2(\lambda_i\alpha_i^2-\alpha_i\eta_i)}{\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i-\eta_i^2/\lambda_i}\right\}$, then the $r-(k,d)$ class estimator dominates the $r-k$ class estimator for values of k such that

$$k > \max_{i \in N_r} \left\{ \frac{2\lambda_i(\sigma^2 + \alpha_i\eta_i + \eta_i^2/\lambda_i)}{(2(\lambda_i\alpha_i^2 - \alpha_i\eta_i) - d(\sigma^2 + \lambda_i\alpha_i^2 + 2\alpha_i\eta_i - \eta_i^2/\lambda_i))} \right\}.$$

(ii) If $\lambda_i\alpha_i^2 - \alpha_i\eta_i > 0$ for all $i \in N_r$ and $\max_{i \in N_r} \left\{ \frac{2(\lambda_i\alpha_i^2 - \alpha_i\eta_i)}{\sigma^2 + \lambda_i\alpha_i^2 + 2\alpha_i\eta_i - \eta_i^2/\lambda_i} \right\} < d < 1$, then the $r-(k,d)$ class estimator dominates the $r-k$ class estimator for all values of $k > 0$.

(iii)If $\lambda_i\alpha_i^2 - \alpha_i\eta_i < 0$ for all $i \in N_r$, the $r-k$ class estimator dominates the $r-(k,d)$ class estimator for all values of $k > 0$ and $0 < d < 1$.

3.5. Comparison of the $r-(k,d)$ class estimator with the two-parameter class estimator

The MSE of the two-parameter class estimator can be obtained by substituting $r = p$ in (2.10), given as:

$$MSE(\hat{\beta}(k,d)) = \sum_{i=1}^{p} \frac{\sigma^2(\lambda_i + kd)^2 + k^2(1-d)^2\lambda_i\alpha_i^2}{\lambda_i(\lambda_i + k)^2}$$
$$+ \sum_{i=1}^{p} \frac{(\lambda_i + kd)^2\eta_i^2 - 2k(1-d)\lambda_i(\lambda_i + kd)\alpha_i\eta_i}{\lambda_i^2(\lambda_i + k)^2}. \quad (3.26)$$

From (2.10) and (3.26), the difference in the MSEs, denoted as Δ_5, is given by:

$$\Delta_5 = MSE(\hat{\beta}(k,d)) - MSE(\hat{\beta}_r(k,d))$$
$$= \sum_{i=1}^{p} \frac{\sigma^2(\lambda_i + kd)^2 + k^2(1-d)^2\lambda_i\alpha_i^2}{\lambda_i(\lambda_i + k)^2} - \sum_{i=1}^{r} \frac{\sigma^2(\lambda_i + kd)^2 + k^2(d-1)^2\lambda_i\alpha_i^2}{\lambda_i(\lambda_i + k)^2} - \sum_{i=r+1}^{p} \alpha_i^2$$
$$+ \sum_{i=1}^{p} \frac{(\lambda_i + kd)^2\eta_i^2 - 2k(1-d)\lambda_i(\lambda_i + kd)\alpha_i\eta_i}{\lambda_i^2(\lambda_i + k)^2} - \sum_{i=1}^{r} \frac{(\lambda_i + kd)^2\eta_i^2 - 2k(1-d)\lambda_i(\lambda_i + kd)\alpha_i\eta_i}{\lambda_i^2(\lambda_i + k)^2}.$$

which can be further simplified as:

$$\Delta_5 = \sum_{i=r+1}^{p} \frac{(\lambda_i + kd)\left[\lambda_i(\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i) + k(d(\sigma^2 + \lambda_i\alpha_i^2 + 2\alpha_i\eta_i + \eta_i^2/\lambda_i) - 2(\lambda_i\alpha_i^2 + \alpha_i\eta_i))\right]}{\lambda_i(\lambda_i + k)^2}. \quad (3.27)$$

From (3.27), it is evident that Δ_5 is positive if

$$\lambda_i\left(\sigma^2 - \lambda_i\alpha_i^2 + \frac{\eta_i^2}{\lambda_i}\right) + k\left(d\left(\sigma^2 + \lambda_i\alpha_i^2 + 2\alpha_i\eta_i + \frac{\eta_i^2}{\lambda_i}\right) - 2\left(\lambda_i\alpha_i^2 + \alpha_i\eta_i\right)\right) > 0 \quad \text{forall } i \in N_{p-r} \quad (3.28)$$

If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i > 0$ for all $i \in N_{p-r}$, Δ_5 is positive when $d(\sigma^2 + \lambda_i\alpha_i^2 + 2\alpha_i\eta_i + \eta_i^2/\lambda_i) - 2(\lambda_i\alpha_i^2 + \alpha_i\eta_i) > 0$ for all $i \in N_{p-r}$, i.e.

$$d > \frac{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)} \quad \text{forall } i \in N_{p-r} \tag{3.29}$$

for all values of $k > 0$.

However, when

$$d < \frac{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)} \quad \text{forall } i \in N_{p-r} \tag{3.30}$$

Δ_5 is positive for the values of k such that

$$k < \frac{\lambda_i(\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i)}{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)} \quad \text{forall } i \in N_{p-r} \tag{3.31}$$

Furthermore, if $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i < 0$ for all $i \in N_{p-r}$ and d satisfies (3.29), Δ_5 is positive for the values of k, which satisfies

$$k > \frac{\lambda_i(\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i)}{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)} \quad \text{forall } i \in N_{p-r}. \tag{3.32}$$

And, if d satisfies (3.30), Δ_5 is negative for all values of $k > 0$ and $0 < d < 1$. The comparisons can be concluded in the following theorem.

Theorem 3.5

(i) If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i > 0$ for all $i \in N_{p-r}$ and $\max_{i\in N_{p-r}}\left\{\frac{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)}\right\} < d < 1$, then $\Delta_5 > 0$ for all $k > 0$.

(ii) If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i > 0$ for all $i \in N_{p-r}$ and $0 < d < \min_{i\in N_{p-r}}\left\{\frac{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)}\right\}$, then $\Delta_5 > 0$ when k is such that

$$k < \min_{i\in N_{p-r}}\left\{\frac{\lambda_i(\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i)}{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)}\right\}$$

(iii) If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i > 0$ for all $i \in N_{p-r}$ and $\max_{i\in N_{p-r}}\left\{\frac{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)}\right\} < d < 1$, then $\Delta_5 > 0$ when k is such that

$$k > \max_{i\in N_{p-r}}\left\{\frac{\lambda_i(\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i)}{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)-d(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)}\right\}.$$

(iv) If $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i > 0$ for all $i \in N_{p-r}$ and $0 < d < \min_{i\in N_{p-r}}\left\{\frac{2(\lambda_i\alpha_i^2+\alpha_i\eta_i)}{(\sigma^2+\lambda_i\alpha_i^2+2\alpha_i\eta_i+\eta_i^2/\lambda_i)}\right\}$, then $\Delta_5 < 0$ for all values of $k > 0$.

In this section, conditions for dominance of the $r-(k,d)$ class estimator over the OLS, ORR, PCR, $r-k$ class estimator and the two-parameter class estimator under the MSE criterion in the misspecified model have been obtained. However, the range of dominance does not remain the same in the misspecified model as it is in the model assumed to be correct. Moreover, the depletion or enlargement of the dominance range for the $r-(k,d)$ class estimator over the other competing estimators depend on certain parametric conditions. For instance, if $\sigma^2-\lambda_i\alpha_i^2+\eta_i^2/\lambda_i\geq 0$, for all $i\in N$, the range of dominance of the $r-(k,d)$ class estimator over the OLS estimator increases in the misspecified model. Furthermore, a Monte Carlo study has been conducted to understand the effect of misspecification on the dominance of the $r-(k,d)$ class estimator over the other competing estimators.

4. Monte Carlo simulation

To compare the dominance of the estimators in true (when there is no misspecification in the model) and misspecified model, the regressors have been generated by the method given in McDonald and Galarneau (1975) and Gibbons(1981), which is defined as:

$$X_i=(1-\rho^2)^{\frac{1}{2}}w_i+\rho w_{p+1},\qquad i=1,2,\dots,p,$$
$$Z_j=(1-\rho^2)^{1/2}w_j+\rho w_{q+1},\qquad j=1,2,\dots,q.$$

where w_i and w_j are $n\times 1$ vectors of independent standard normal pseudo-random numbers, ρ is specified so that the correlation between any two regressors is given by ρ^2. The dependent variable y has been generated as follows:

$$y=D\zeta+u=X\beta+Z\gamma+u;\qquad u{\sim}N(0,\sigma^2I)\tag{4.1}$$

where $\zeta=[\beta\gamma]$. u is a vector of normal pseudo-random numbers with standard deviation σ. Following McDonald Galarneau (1975), Gibbons (1981), Kibria (2003) and others, ζ has been chosen as the normalized eigenvector corresponding to the largest eigenvalue of the $D'D$ matrix. As this study is aimed at studying the effect of the omission of relevant regressors on the performance of some competing estimators of β, the following is estimated for the model (4.1) and the misspecified model: when there is no misspecification, both X and Z have been used in estimation and when there is misspecification due to the omission of relevant regressors, information in Z matrix has not been used to estimate β. For example, the OLS estimator for the misspecified model is obtained by:

$$\hat{\beta}_M=(X'X)^{-1}X'y$$

and the OLS estimate of β in the case of no misspecification is obtained by taking first p components of the OLS estimate of ζ, given as:

$$\hat{\zeta}=(D'D)^{-1}D'y=[\hat{\beta}_T\quad\hat{\gamma}_T]'.$$

In this study, simulation is done for some selected values of $n,\ p,\ q,\ \rho,\ \sigma^2$, k and d to compare the performance of the estimators. The values of the parameters are taken as: $n = 50$; $p = 5$; $q = 3$; $\rho = 0.95,0.99$; $\sigma = 0.5,1$; $k = 0.1,0.5,0.9,1.5,5$ and $d = 0.1,0.5,0.9$. The value of r is decided by a scree plot, which is drawn between eigenvalues and components (see Johnson and Wichern (2007)). For each parametric combination, the simulation process has been repeated 2500 times and the estimated MSE (EMSE) is calculated by the following formula

$$EMSE(\hat{\beta}) = \frac{1}{2500}\sum_{i=1}^{2500} \left(\hat{\beta}_{(i)} - \beta\right)'\left(\hat{\beta}_{(i)} - \beta\right), \tag{4.2}$$

where $\hat{\beta}_{(i)}$ is the estimated value of β in i^{th} iteration. The results of the simulation are shown in Tables 1 to 4, where EMSE of estimators in true model and in misspecified model are denoted by $EMSE_T$ and $EMSE_M$, respectively, and $\tilde{\beta}$, $\tilde{\beta}(k)$, $\tilde{\beta}_r$, $\tilde{\beta}_r(k)$, $\tilde{\beta}(k,d)$ and $\tilde{\beta}_r(k,d)$ denote the OLS, ORR, PCR, $r-k$ class, two-parameter class and $r-(k,d)$ class estimators respectively. The following remarks are made from simulation results:

Table 1. Estimated MSE of the estimators for true and misspecified model when $\rho = 0.95$ and $\sigma = 0.5$

		$\boldsymbol{d = 0.1}$		$\boldsymbol{d = 0.5}$		$\boldsymbol{d = 0.9}$	
$\boldsymbol{k}$		$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$
0.1	$\tilde{\boldsymbol{\beta}}$	0.2801295	0.4486118	0.2801295	0.4486118	0.2801295	0.4486118
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.2640343	0.4350360	0.2640342	0.4350360	0.2640342	0.4350360
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.2656187	0.4363773	0.2720124	0.4417786	0.2784949	0.4472379
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0003777	0.1855473	0.0003777	0.1855473	0.0003777	0.1855473
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0003774	0.1851850	0.0003774	0.1851850	0.0003774	0.1851850
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0003774	0.1852212	0.0003775	0.1853661	0.0003776	0.1855110
0.5	$\tilde{\boldsymbol{\beta}}$	0.2801295	0.4486118	0.2801295	0.4486118	0.2801295	0.4486118
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.2131531	0.3899955	0.2131531	0.3899955	0.2131531	0.3899955
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.2193777	0.3955276	0.2453274	0.4183884	0.2729588	0.4424207
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0003777	0.1855473	0.0003777	0.1855473	0.0003777	0.1855473
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0003768	0.1837417	0.0003768	0.1837417	0.0003768	0.1837417
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0003768	0.1839219	0.0003770	0.1846434	0.0003775	0.1853663
0.9	$\tilde{\boldsymbol{\beta}}$	0.2801295	0.4486118	0.2801295	0.4486118	0.2801295	0.4486118
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.1769369	0.3558756	0.1769369	0.3558756	0.1769369	0.3558756
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.1860480	0.3642675	0.2251774	0.3997944	0.2686021	0.4384565
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0003777	0.1855473	0.0003777	0.1855473	0.0003777	0.1855473
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0003773	0.1823080	0.0003773	0.1823080	0.0003773	0.1823080
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0003771	0.1826306	0.0003768	0.1839241	0.0003774	0.1852221
1.5	$\tilde{\boldsymbol{\beta}}$	0.2801295	0.4486118	0.2801295	0.4486118	0.2801295	0.4486118
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.1388449	0.3179755	0.1388449	0.3179755	0.1388449	0.3179755
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.1504966	0.3291388	0.2026073	0.3780150	0.2635243	0.4336479
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0003777	0.1855473	0.0003777	0.1855473	0.0003777	0.1855473
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0003799	0.1801752	0.0003799	0.1801752	0.0003799	0.1801752
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0003790	0.1807089	0.0003770	0.1828514	0.0003772	0.1850065

Table 1. Estimated MSE of the estimators for true and misspecified model when $\rho = 0.95$ and $\sigma = 0.5$ (cont.)

$\boldsymbol{k}$		$\boldsymbol{d = 0.1}$		$\boldsymbol{d = 0.5}$		$\boldsymbol{d = 0.9}$	
		$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$
5	$\tilde{\boldsymbol{\beta}}$	0.2801295	0.4486118	0.2801295	0.4486118	0.2801295	0.4486118
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.0526366	0.2207759	0.0526366	0.2207759	0.0526366	0.2207759
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.0666946	0.2360594	0.1422406	0.3138604	0.2486889	0.4183282
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0003777	0.1855473	0.0003777	0.1855473	0.0003777	0.1855473
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0004400	0.1681494	0.0004400	0.1681494	0.0004400	0.1681494
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0004267	0.1698506	0.0003892	0.1767411	0.0003768	0.1837689

Table 2. Estimated MSE of the estimators for true and misspecified model when $\rho = 0.95$ and $\sigma = 1$

$\boldsymbol{k}$		$\boldsymbol{d = 0.1}$		$\boldsymbol{d = 0.5}$		$\boldsymbol{d = 0.9}$	
		$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$
0.1	$\tilde{\boldsymbol{\beta}}$	1.1205180	1.1783938	1.1205180	1.1783938	1.1205180	1.1783938
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	1.0561408	1.1288003	1.0561408	1.1288003	1.0561408	1.1288003
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	1.0624786	1.1336995	1.0880518	1.1534299	1.1139803	1.1733743
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0015107	0.1888630	0.0015107	0.1888630	0.0015107	0.1888630
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0015098	0.1884980	0.0015098	0.1884980	0.0015098	0.1884980
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0015099	0.1885345	0.0015103	0.1886804	0.0015106	0.1888264
0.5	$\tilde{\boldsymbol{\beta}}$	1.1205180	1.1783938	1.1205180	1.1783938	1.1205180	1.1783938
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.8526281	0.9646305	0.8526281	0.9646305	0.8526281	0.9646305
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.8775255	0.9847905	0.9813187	1.0681334	1.0918374	1.1558011
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0015107	0.1888630	0.0015107	0.1888630	0.0015107	0.1888630
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0015066	0.1870440	0.0015066	0.1870440	0.0015066	0.1870440
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0015069	0.1872255	0.0015085	0.1879524	0.0015103	0.1886807
0.9	$\tilde{\boldsymbol{\beta}}$	1.1205180	1.1783938	1.1205180	1.1783938	1.1205180	1.1783938
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.7077687	0.8408833	0.7077687	0.8408833	0.7077687	0.8408833
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.7442128	0.8713751	0.9007244	1.0005850	1.0744122	1.1413835
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0015107	0.1888630	0.0015107	0.1888630	0.0015107	0.1888630
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0015044	0.1855997	0.0015044	0.1855997	0.0015044	0.1855997
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0015048	0.1859248	0.0015069	0.1872278	0.0015099	0.1885353
1.5	$\tilde{\boldsymbol{\beta}}$	1.1205180	1.1783938	1.1205180	1.1783938	1.1205180	1.1783938
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.5554001	0.7045712	0.5554001	0.7045712	0.5554001	0.7045712
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.6020087	0.7449187	0.8104502	0.9219414	1.0541030	1.1239768
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0015107	0.1888630	0.0015107	0.1888630	0.0015107	0.1888630
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0015030	0.1834511	0.0015030	0.1834511	0.0015030	0.1834511
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0015032	0.1839887	0.0015051	0.1861471	0.0015093	0.1883182
5	$\tilde{\boldsymbol{\beta}}$	1.1205180	1.1783938	1.120518	1.1783938	1.120518	1.1783938
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.2103956	0.3713567	0.2103956	0.3713567	0.2103956	0.3713567
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.2666748	0.4242071	0.5689763	0.6975050	0.9947728	1.0698368
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0015108	0.1888630	0.0015108	0.1888630	0.0015108	0.1888630
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0015404	0.1713349	0.0015404	0.1713349	0.0015404	0.1713349
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0015303	0.1730490	0.0015059	0.1799914	0.0015066	0.1870714

Table 3. Estimated MSE of the estimators for true and misspecified model when $\rho = 0.99$ and $\sigma = 0.5$

		$\boldsymbol{d = 0.1}$		$\boldsymbol{d = 0.5}$		$\boldsymbol{d = 0.9}$	
$\boldsymbol{k}$		$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$
0.1	$\widetilde{\boldsymbol{\beta}}$	1.3733302	1.4155047	1.3733302	1.4155047	1.3733302	1.4155047
	$\widetilde{\boldsymbol{\beta}}(\boldsymbol{k})$	1.0501954	1.1614661	1.0501954	1.1614661	1.0501954	1.1614661
	$\widetilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	1.0802694	1.1854445	1.2055422	1.2845257	1.3387773	1.3886753
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0003490	0.2090398	0.0003490	0.2090398	0.0003490	0.2090398
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0003487	0.2086612	0.0003487	0.2086612	0.0003487	0.2086612
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0003487	0.2086991	0.0003488	0.2088504	0.0003489	0.2090019
0.5	$\boldsymbol{\beta}^{\sim}$	1.3733302	1.4155047	1.3733302	1.4155047	1.3733302	1.4155047
	$\boldsymbol{\beta}^{\sim}(\boldsymbol{k})$	0.4930231	0.6729832	0.4930231	0.6729832	0.4930231	0.6729832
	$\boldsymbol{\beta}^{\sim}(\boldsymbol{k},\boldsymbol{d})$	0.5592052	0.7310081	0.8724862	0.9991682	1.2634509	1.3250253
	$\boldsymbol{\beta}^{\sim}\_\boldsymbol{r}$	0.0003490	0.2090398	0.0003490	0.2090398	0.0003490	0.2090398
	$\boldsymbol{\beta}^{\sim}\_\boldsymbol{r}\ (\boldsymbol{k})$	0.0003482	0.207153	0.0003482	0.207153	0.0003482	0.207153
	$\boldsymbol{\beta}^{\sim}\_\boldsymbol{r}\ (\boldsymbol{k},\boldsymbol{d})$	0.0003482	0.2073413	0.0003484	0.2080953	0.0003488	0.2088507
0.9	$\widetilde{\boldsymbol{\beta}}$	1.3733302	1.4155047	1.3733302	1.4155047	1.3733302	1.4155047
	$\widetilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.2931221	0.4843209	0.2931221	0.4843209	0.2931221	0.4843209
	$\widetilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.3626887	0.5471788	0.7264088	0.865856	1.2268551	1.2921258
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0003490	0.2090398	0.0003490	0.2090398	0.0003490	0.2090398
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0003486	0.2056542	0.0003486	0.2056542	0.0003486	0.2056542
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0003484	0.2059915	0.0003482	0.2073435	0.0003487	0.2087000
1.5	$\widetilde{\boldsymbol{\beta}}$	1.3733302	1.4155047	1.3733302	1.4155047	1.3733302	1.4155047
	$\widetilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.1645441	0.3596925	0.1645441	0.3596925	0.1645441	0.3596925
	$\widetilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.2297062	0.4196822	0.6141691	0.7609554	1.1967351	1.2643319
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0003490	0.2090398	0.0003490	0.2090398	0.0003490	0.2090398
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0003509	0.2034235	0.0003509	0.2034235	0.0003509	0.2034235
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0003501	0.2039817	0.0003483	0.2062221	0.0003486	0.2084747
5	$\widetilde{\boldsymbol{\beta}}$	1.3733302	1.4155047	1.3733302	1.4155047	1.3733302	1.4155047
	$\widetilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.0280793	0.2169173	0.0280793	0.2169173	0.0280793	0.2169173
	$\widetilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.0697681	0.2569569	0.4428262	0.5944911	1.1459688	1.2158268
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0003490	0.2090398	0.0003490	0.2090398	0.0003490	0.2090398
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0004035	0.1908219	0.0004035	0.1908219	0.0004035	0.1908219
	$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0003918	0.1926062	0.000359	0.1998268	0.0003482	0.2071805

Table 4. Estimated MSE of the estimators for true and misspecified model when $\rho = 0.99$ and $\sigma = 1$

		$\boldsymbol{d = 0.1}$		$\boldsymbol{d = 0.5}$		$\boldsymbol{d = 0.9}$	
$\boldsymbol{k}$		$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$	$\boldsymbol{EMSE_T}$	$\boldsymbol{EMSE_M}$
0.1	$\tilde{\boldsymbol{\beta}}$	5.4933208	4.9729356	5.4933208	4.9729356	5.4933208	4.9729356
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	4.2007888	3.9736359	4.2007888	3.9736359	4.2007888	3.9736359
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	4.3210844	4.0679655	4.8221727	4.4577291	5.3551101	4.8674052
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0013960	0.2122746	0.0013960	0.2122746	0.0013960	0.2122746
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0013951	0.2118936	0.0013951	0.2118936	0.0013951	0.2118936
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0013952	0.2119316	0.0013956	0.2120840	0.0013959	0.2122365
0.5	$\tilde{\boldsymbol{\beta}}$	5.4933208	4.9729356	5.4933208	4.9729356	5.4933208	4.9729356
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	1.9721133	2.0517057	1.9721133	2.0517057	1.9721133	2.0517057
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	2.2368417	2.2799880	3.4899605	3.3349853	5.0538077	4.6169719
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0013960	0.2122746	0.0013960	0.2122746	0.0013960	0.2122746
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0013923	0.2103753	0.0013923	0.2103753	0.0013923	0.2103753
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0013926	0.2105648	0.0013940	0.2113239	0.0013956	0.2120843
0.9	$\tilde{\boldsymbol{\beta}}$	5.4933208	4.9729356	5.4933208	4.9729356	5.4933208	4.9729356
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	1.1725128	1.3109990	1.1725128	1.3109990	1.1725128	1.3109990
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	1.4507813	1.5580588	2.9056605	2.8110399	4.9074278	4.4876081
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0013960	0.2122746	0.0013960	0.2122746	0.0013960	0.2122746
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0013904	0.2088665	0.0013904	0.2088665	0.0013904	0.2088665
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0013907	0.2092061	0.0013926	0.2105671	0.0013952	0.2119325
1.5	$\tilde{\boldsymbol{\beta}}$	5.4933208	4.9729356	5.4933208	4.9729356	5.4933208	4.9729356
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.6581957	0.8249431	0.6581957	0.8249431	0.6581957	0.8249431
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.9188519	1.0601478	2.4567130	2.4000657	4.7869520	4.3785418
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0013960	0.2122746	0.0013960	0.2122746	0.0013960	0.2122746
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0013891	0.2066210	0.0013891	0.2066210	0.0013891	0.2066210
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0013893	0.2071829	0.0013910	0.2094382	0.0013947	0.2117058
5	$\tilde{\boldsymbol{\beta}}$	5.4933208	4.9729356	5.4933208	4.9729356	5.4933208	4.9729356
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k})$	0.1121679	0.2972680	0.1121679	0.2972680	0.1121679	0.2972680
	$\tilde{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	0.2789891	0.4501910	1.7713719	1.7610914	4.5839115	4.1907251
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$	0.0013960	0.2122746	0.0013960	0.2122746	0.0013960	0.2122746
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	0.0014216	0.1939342	0.0014216	0.1939342	0.0014216	0.1939342
	$\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$	0.0014128	0.1957307	0.0013915	0.2030001	0.0013923	0.2104030

Since ρ affects the structure of the design matrix, the estimated MSEs of $\tilde{\beta}$ and $\tilde{\beta}_r$ are the same for all values of k and d for a fixed σ in true and misspecified models. As expected, for higher value of σ, the estimated MSEs inflate for all the estimators in true and misspecified model as well. Similarly, when the collinearity among the regressors increases, the estimated MSEs of the estimators inflate in both the models.

As the theoretical results suggest, the MSE of the estimator may increase due to the omission of relevant regressors depending on the values of unknown parameters. When we compare the performances of the estimators in true and misspecified model for all choices of the parameters involved almost all the estimators have larger estimated MSE in the misspecified model than in the case where there is no misspecification.

While examining the variations in the estimated MSE of the estimators with respect to the variations in k and d from Tables 1-4, we observe that as the value of k increases, the values of the estimated MSEs decrease for all the estimators considered here where k is involved. However, $\tilde{\beta}_r(k,d)$ in true model exhibits a pattern of concave up function of k, that is the estimated MSE of $\tilde{\beta}_r(k,d)$ first decreases and then increases after attaining a minimum value of the MSE with the increase in the value of k. In our simulation, the minimum value of the MSE of the $r-(k,d)$ class estimator when $d = 0.1,0.5$ is attained for some value of k in between 0.9 to 1.5 and 1.5 to 5 for $\sigma = 0.5$ and $\sigma = 1$ respectively.

However, with the increase in the value of d, the estimated MSEs of $\tilde{\beta}(k,d)$ and $\tilde{\beta}_r(k,d)$ increase for the selected values of ρ and σ for both the models. The values of the estimated MSEs show that the $r-(k,d)$ class estimator performs better than the OLS, ORR, PCR, two-parameter class estimator for all chosen values of k, d, σ and ρ, although the dominance of the $r-(k,d)$ class estimator over the $r-k$ class estimator depends on the choices of k and d. In fact, the difference in the estimated MSE values of the $r-(k,d)$ class estimator and $r-k$ class estimator do not show much difference if seen up to the third or forth decimal places for small σ, however, if observed up to the sixth or seventh decimal places, the MSE of the $r-k$ class estimator is found to be less than that of the $r-(k,d)$ class estimator. For $\sigma = 1$, the $r-k$ class estimator shows dominance over the $r-(k,d)$ class estimator in the misspecified model, see Table 2 and 4, the reason being the condition of dominance of the $r-(k,d)$ class estimator over the $r-k$ class estimator (see Theorem 3.4) is not satisfied in this simulation.

5. Numerical example

In order to illustrate our theoretical results, in this section we now consider the data set on Total National Research and Development Expenditures as a Per cent of Gross National Product originally due to Gruber (1998), also analysed by Zhong and Yang (2007). It represents the relationship between the dependent variable Y, the percent spent by the U.S., and the four other independent variables X_1, X_2, X_3 and X_4. The variables X_1, X_2, X_3 and X_4, respectively represents the percent spent by France, West Germany, Japan and the former Soviet Union. The variables are standardized and the OLS estimator of $\beta = (\beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4)'$ is obtained as β= (0.6455,0.0896,0.1436,0.1526)'. We obtain the eigenvalues of $X'X$ as λ_1 = 302.9626, λ_2 = 0.7283, λ_3= 0.0446, and λ_4 = 0.0345, and the condition number is approximately 8,776.382. Hence, the design matrix is quite ill-conditioned.

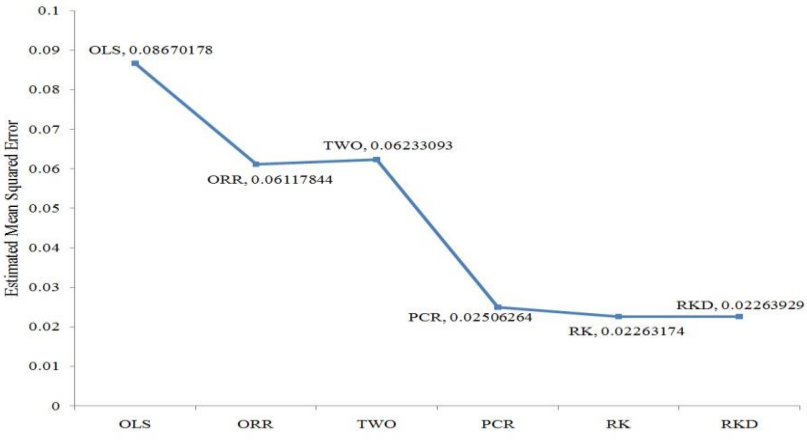
Now, let us consider that the investigator has omitted *Z* = [*X₄*] mistakenly, which results in the misspecified model (2.2) with *X* matrix having 3 variables X_1, X_2, and X_3. The eigenvalues of the *X* matrix in the misspecified model are 161.38584077, 0.10961836 and 0.04454088, and the condition number is 3623.32, which indicates an ill-conditioned design matrix in the misspecified model. The OLS estimators of β, γ and σ^2 in the model (2.2) is obtained as β= (0.80878236, 0.41402294, –0.09630492)', $\hat{\gamma} = \hat{\beta}_4$= 0.1526, $\hat{\sigma}^2$=0.002745 respectively and we chose *r*=2. The values of k and d are chosen as: $k_0 = \hat{\sigma}^2/\alpha^2_{\max}$=0.01178, where $\alpha_{\max}$ is the maximum element of *α=T'β,* which was suggested by Hoerl and Kennard (1970) and *d*=0.0557 is the positive solution of $\frac{\partial \text{ MSE}(\hat{\beta}_r(k_0,d)}{\partial d} = 0$. The MSEs of the estimators are estimated by replacing β with the PCR estimator, which is an unbiased estimator, and are presented in Table 5 along with the estimated values of regression coefficients for both true and misspecified model. Figure 2 represents the estimated MSEs of the estimators in the two models.

From Table 5, we can see the sign of β_3 has changed in the misspecified model from positive to negative, which gives an evidence of the well-established results that the omission affects the estimation of parameters. Further, the estimated MSEs increase in the misspecified model as compared to the true model. We observe that the *r*–(*k*,*d*) class estimator outperforms the OLS, ORR, two-parameter class and PCR estimators in MSE sense. However, the MSEs of the *r-k* class estimator and the *r-(k,d)* class estimator are almost equal and the difference can be only noticed at sixth decimal place. The dominance of the estimators can be easily seen in Figure 2.

On the other hand, from the results stated in Table 5 for the misspecified model, we see that the *r*–(*k*, *d*) class estimator is superior to the OLS, ORR, two-parameter class and PCR estimators, and does not perform better than the *r –k* class estimator under the MSE criterion. Moreover, the theoretical findings obtained in this study support the numerical results given in Table 5. Now, in order to verify the conditions of the dominance under MSE criterion, let us take Theorem 3.1, where we get $\sigma^2 - \lambda_i\alpha_i^2 + \eta_i^2/\lambda_i$= –57.0280, –0.0042, 0.0027, clearly condition (ii) of the theorem will be applied and the lower limits of *d* are –0.1592445–1.9733796, thus the *r*–(*k*, *d*) class estimator dominates the OLS estimator for all values of *d*, which is the result obtained in the numerical illustration. Next, let us take the condition of dominance of *r*–(*k*, *d*) class over the *r –k* class given in Theorem 3.4; $\lambda_i\alpha_i^2 - \alpha_i\eta_i$= 46.2405, 0.0037 for *i* =1,2 and the value of *d* is 0.0557, which satisfies the condition (i) in Theorem 3.4 as the values of $2(\lambda_i\alpha_i^2 - \alpha_i\eta_i)/(\sigma^2 + \lambda_i\alpha_i^2 + 2\alpha_i\eta_i - \frac{\eta_i^2}{\lambda_i})$ for *i*=1,2 are 1.0856, 0.1421. Further, the value of the lower bound of *k* in condition (i) of Theorem 3.4 comes out to be 63.872591. Evidently the condition is not satisfied, hence the *r*–(*k*, *d*) class estimator does not dominate the *r –k* class estimator in this numerical illustration. Similarly, other conditions can also be verified.

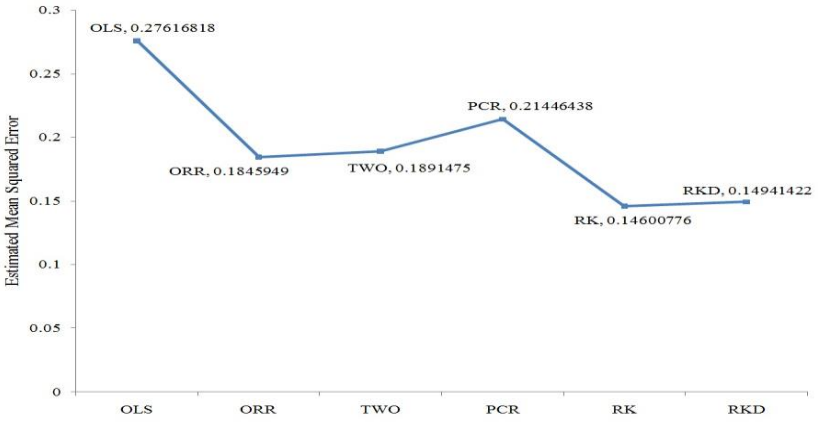
Table 5. Estimated values of regression coefficients and estimated MSEs for true and misspecified model.

	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}(\boldsymbol{k})$	$\widehat{\boldsymbol{\beta}}(\boldsymbol{k},\boldsymbol{d})$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}}$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k})$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}}(\boldsymbol{k},\boldsymbol{d})$
			True Model			
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$	0.645458	0.551069	0.556329	0.209956	0.209236	0.209276
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$	0.089588	0.115598	0.114148	0.240076	0.23948	0.239514
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}$	0.143557	0.180012	0.17798	0.304667	0.302885	0.302984
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}$	0.152618	0.163265	0.162671	0.186063	0.187951	0.187845
$\widehat{\mathbf{MSE}}$	**0.086702**	**0.061178**	**0.062331**	**0.025063**	**0.022632**	**0.022639**
			Misspecified Model			
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$	0.808782	0.71553	0.720727	0.409013	0.399393	0.399929
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$	0.414023	0.438716	0.437339	0.662707	0.635374	0.636897
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}$	-0.0963	-0.04272	-0.0457	-0.01613	0.020682	0.01863
$\widehat{\mathbf{MSE}}$	**0.276168**	**0.184595**	**0.189148**	**0.214464**	**0.146008**	**0.149414**



(a) In the case of no misspecification

Figure 2. Estimated MSE of the estimators



b) When there is misspecification

Figure 2. Estimated MSE of the estimators (cont.)

6. Conclusion

In this paper the effect of misspecification due to omission of relevant regressors in a linear regression model when the problem of multicollinearity exists, on the dominance of the *r*-(*k*,*d*) class estimator over the other competing estimators have been studied. The dominance conditions of the *r*-(*k*,*d*) class estimator over the OLS, ORR, PCR, *r*-*k* class and the two-parameter class estimators have been derived under scalar mean squared error criterion. It has been observed that the MSE of the estimators may increase or decrease due to misspecification depending on the values of the unknown parameters. Similarly, the ranges of dominance of the *r*-(*k*,*d*) class estimator over the others may shrink or widen in the misspecified model. To understand the effect of misspecification on dominance of the *r*-(*k*,*d*) class estimator over the others a Monte Carlo simulation and a numerical example have been given and it is observed that the MSE of the estimators increases in the misspecified model as compared to the model assumed to be true. The *r*-(*k*,*d*) class estimator performs better than the OLS, ORR, two-parameter class estimator and the PCR estimator in the misspecified model as well for all chosen values of the parameters. However, the *r*-(*k*,*d*) class estimator and the *r*-*k* class estimator do equally well when observed up to few decimal places in simulation, whereas in the numerical example the *r*-*k* class estimator is found to be the most suited as an alternative to the OLS estimator in the misspecified model with multicollinearity. Hence, the study stuggests that the r-k class estimator or the r-(k,d) class estimators are a better choice over the other estimators considered in this study in the case of the misspecified model with multicollinearity.

Acknowledgement

The authors are grateful to the editor, the associate editor, and the anonymous referees for their valuable comments and suggestions to improve this article.

REFERENCES

BAYE, M. R., PARKER, D. F., (1984). Combining ridge and principal component regression: A money demand illustration. Communications in Statistics- Theory and Methods, 13, pp. 197–205.

GIBBONS, D. G., (1981). A simulation study of some ridge estimators. Journal of the American Statistical Association, 76, pp. 131–139.

GRAHAM, M. H., (2003). Confronting multicollinearity in ecological multiple regression. Ecology, 84, pp. 2809–2815.

GRUBER, M., (1998). Improving efficiency by shrinkage: The James-Stein and ridge regression estimator. Marcel Dekker, Inc., New York.

HAMILTON, J. L., (1972). The demand for cigarettes: Advertising, the health scare, and the cigarette advertising ban. The Review of Economics and Statistics, 54, pp. 401–411.

HEIKKILA, E., (1988). Multicollinearity in regression models with multiple distance measures. Journal of Regional Science, 28, pp. 345–362.

HOERL, A. E., KENNARD, R. W., (1970). Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 12, pp. 55–67.

JOHNSON, R. A., WICHERN, D. W., (2007). Applied multivariate statistical analysis. Pearson-Prentice Hall, New Jersey.

KAÇIRANLAR, S., SAKALLIOĞLU, S., (2001). Combining the Liu estimator and the principal component regression estimator. Communications in Statistics – Theory and Methods, 30, pp. 2699–2705.

KADIYALA, K., (1986). Mixed regression estimator under misspecification. Economic Letters, 21, pp. 27–30.

KIBRIA, B., (2003). Performance of some new ridge regression estimators. Communications in Statistics – Theory and Methods, 32, pp. 419–435.

MAHAJAN, V., JAIN, A. K., BERGIER, M., (1977). Parameter estimation in marketing models in the presence of multicollinearity: An application of ridge regression. Journal of Marketing Research, 14, pp. 586–591.

MASSY, W. F., (1965). Principal components regression in exploratory statistical research. Journal of the American Statistical Association, 60, pp. 234–256.

MCDONALD, G. C., GALARNEAU, D. I., (1975). A Monte Carlo evaluation of some ridge-type estimators. Journal of the American Statistical Association, 70, pp.407–416.

NOMURA, M., OHKUBO, T., (1985). A note on combining ridge and principal component regression. Communications in Statistics-Theory and Methods, 14, pp. 2489–2493.

ÖZKALE, M. R., KAÇIRANLAR, S., (2007). The restricted and unrestricted two parameter estimators. Communications in Statistics- Theory and Methods, 36, pp. 2707–2725.

ÖZKALE, M. R., (2012). Combining the unrestricted estimators into a single estimator and a simulation study on the unrestricted estimators. Journal of Statistical Computation and Simulation, 82, pp. 653–688.

ÖZKALE, M., KAÇIRANLAR, S., (2008). Comparison of the $r-k$ class estimator to the ordinary least squares estimator under the Pitman's measure of closeness criterion. Statistical Papers, 49, pp. 503–512.

SARKAR, N., CHANDRA, S., (2015). Comparison of the r-(k,d) class estimator with some estimators for multicollinearity under the Mahalanobis loss function. Forthcoming paper in International Econometric Review, Vol. 7, issue 1, pp. 1–12.

SARKAR, N., (1989). Comparisons among some estimators in misspecified linear models with multicollinearity. Annals of Institute of Statistical Methods, 41, pp. 717–724.

SARKAR, N., (1996). Mean square error matrix comparison of some estimatorsin linear regressions with multicollinearity. Statistics & Probability Letters, 30, pp. 133–138.

TRENKLER, G., WIJEKOON, P., (1989). Mean squared error matrix superiority of the mixed regression estimator under misspecification. Statistica, 44, pp. 65–71.

WIJEKOON, P., TRENKLER, G., (1989). Mean squared error matrix superiority of estimators under linear restrictions and misspecification. Economics Letters, 30, pp. 141–149.

ZHONG, Z., YANG, H., (2007). Ridge estimation to the restricted linear model. Communications in Statistics- Theory and Methods, 36, pp. 2099–2115.

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 53–74

IMPROVED ESTIMATION OF THE SCALE PARAMETER FOR LOG-LOGISTIC DISTRIBUTION USING BALANCED RANKED SET SAMPLING

Housila P. Singh[1], **Vishal Mehta**[2]

ABSTRACT

In this article we have suggested some improved estimators of a scale parameter of log-logistic distribution (LLD) under a situation where the units in a sample can be ordered by judgement method without any error. We have also suggested some linear shrinkage estimator of a scale parameter of LLD. Efficiency comparisons are also made in this work.

Key words: minimum mean squared error estimator, shrinkage estimator, log-logistic distribution, best linear unbiased estimator, median ranked set sample.

AMS Subject Classification: 62G30; 62H12.

1. Introduction

Ranked set sampling (RSS) is a method of sampling that can be advantageous when quantification of all sampling units is costly but a small set of units can be easily ranked, according to the character under investigation, without actual quantification. The technique was first introduced by McIntyre (1952) for estimating means pasture and forage yields. The theory and application of ranked set sampling given by Chen *et al.* (2004).

A random variable X is said to have a log-logistic distribution with the scale parameter α and the shape parameter β if its cumulative distribution function (CDF) and probability density function (PDF) are respectively given as (see, Lesitha and Thomas (2012))

$$F(x;\alpha,\beta)=\frac{x^{\beta}}{\alpha^{\beta}+x^{\beta}},x>0,\alpha>0,\beta>1 \tag{1}$$

and

[1] School of Studies in Statistics, Vikram University, Ujjain - 456010, Madhya Pradesh, India. E-mail: hpsujn@gmail.com.

[2] Indian Statistical Institute (ISI), North-East Centre, Tezpur - 784028, Assam, India. E-mail: visdewas@gmail.com.

$$f(x;\alpha,\beta)=\frac{\beta\alpha^{\beta}x^{\beta-1}}{\left(\alpha^{\beta}+x^{\beta}\right)^{2}},x>0,\alpha>0,\beta>1. \tag{2}$$

Also, the *kth* moment of (2) exists only when $k<\beta$ and is given by

$$E\left(X^{k}\right)=\alpha^{k}B\left(1-\frac{k}{\beta},1+\frac{k}{\beta}\right), \tag{3}$$

where B denotes beta function.

The applications of log-logistic distribution are well known in a survival analysis of data sets such as survival times of cancer patients in which the hazard rate increases initially and decreases later (for example, see Bennett (1983)). In economic studies of distributions of wealth or income, it is known as Fisk distribution (see Fisk (1961)) and is considered as an equivalent alternative to a lognormal distribution. For further details on the importance and applications of a log-logistic distribution one may refer to Shoukri *et al.* (1988), Geskus (2001), Robson and Reed (1999) and Ahmad *et al.* (1988). For current reference in this context the reader is referred to Singh and Mehta (2013; 2014, a, b, 2015, 2016 a, b, c), Mehta and Singh (2014) and Mehta (2015).

If $X_{1:n},X_{2:n},...,X_{n:n}$ are the order statistics of a random sample of size n drawn from (1) then

$$Y_{r:n}=\frac{X_{r:n}}{\alpha},r=1,2,...,n, \tag{4}$$

are distributed as order statistics of the same sample size drawn from a $LLD(1,\beta)$with PDF given by

$$g(y,\beta)=\frac{\beta y^{\beta-1}}{\left(1+y^{\beta}\right)^{2}},y>0,\beta>1. \tag{5}$$

For a detailed description of various properties of order statistics arising from a $LLD(1,\beta)$ one may refer to Ragab and Green (1984). Balakrishnan and Malik (1987) have given some recurrence relations on the single and product moments of order statistics arising from a $LLD(1,\beta)$. Suppose

$$\gamma_{r:n}=E\left(Y_{r:n}\right),r=1,2,...,n, \tag{6}$$

$$\sigma_{r,s:n}=Cov\left(Y_{r:n},Y_{s:n}\right),1\le r<s\le n \tag{7}$$

and

$$\sigma_{r,r:n}=Var\left(Y_{r:n}\right),1\le r\le n. \tag{8}$$

By using (4) in (6)-(8) we have

$$E(X_{r:n}) = \alpha\gamma_{r:n}, 1 \le r \le n \tag{9}$$

$$Cov(X_{r:n}, X_{s:n}) = \alpha^2\sigma_{r,s:n}, 1 \le r < s \le n \tag{10}$$

$$Var(X_{r:n}) = \alpha^2\sigma_{r,r:n}, 1 \le r \le n\,. \tag{11}$$

Lesitha and Thomas (2012) have computed the values of $\gamma_{r:n}$ and $\sigma_{r,s:n}, 1 \le r,s \le n$ independently for $n = 2(1)8$ and for $\beta = 2.5(0.5)5.0$ using Mathcad software so as to use those values for the computation of BLUE of α based on order statistics. If $\boldsymbol{X} = (X_{1:n}, X_{2:n}, ..., X_{n:n})'$ then the mean vector $E(X)$ and dispersion matrix $D(X)$ of X are

$$E(\boldsymbol{X}) = \gamma\alpha$$

and

$$D(\boldsymbol{X}) = \alpha^2 G,$$

where $\gamma = (\gamma_{1:n}, \gamma_{2:n}, ..., \gamma_{n:n})'$ and $\boldsymbol{G} = ((\sigma_{r,s:n}))$.

Thus, by Gauss-Markov theorem Lesitha and Thomas (2012) gives the BLUE $\hat{\alpha}$ based on order statistics of a random sample of size n as:

$$t_1 = \hat{\alpha} = \left(\gamma' G^{-1}\gamma\right)^{-1}\gamma' G^{-1}\boldsymbol{X}$$

and

$$Var(t_1) = \left(\gamma' G^{-1}\gamma\right)^{-1}\alpha^2 = \alpha^2 V_1, \tag{12}$$

where $V_1 = \left(\gamma' G^{-1}\gamma\right)^{-1}$.

Lesitha and Thomas (2012) further estimate α based on the mean of unbiased estimators of α defined from each individual observations in the balanced ranked set sampling as:

$$t_2 = \alpha^* = \frac{1}{n}\sum_{r=1}^{n}\left[\frac{X_{(r:n)r}}{\gamma_{r:n}}\right],$$

with

$$Var(t_2) = \frac{1}{n^2}\sum_{r=1}^{n}\left[\frac{\sigma_{r,r:n}}{\gamma_{r:n}^2}\right]\alpha^2 = \alpha^2 V_2, \tag{13}$$

where $V_2 = \frac{1}{n^2}\sum_{r=1}^{n}\left[\frac{\sigma_{r,r:n}}{\gamma_{r:n}^2}\right]$.

Lesitha and Thomas (2012) also estimate α based on BLUE in the balanced ranked set sampling as:

$$t_3 = \alpha^{**} = \left(\gamma' G_1^{-1}\gamma\right)^{-1}\gamma' G_1^{-1} X_{rss}$$

and

$$Var(t_3) = \left(\gamma' G_1^{-1}\gamma\right)^{-1}\alpha^2 = \alpha^2 V_3\,. \tag{14}$$

where $\gamma = \left(\gamma_{1:n}, \gamma_{2:n}, ..., \gamma_{n:n}\right)'$, $G_1 = diag\left(\sigma_{1,1:n}, \sigma_{2,2:n}, ..., \sigma_{n,n:n}\right)$ and $V_3 = \left(\gamma' G_1^{-1}\gamma\right)^{-1}$.

When n is small the estimators α^* and α^{**} may not be acceptable for the expected level of precision. In such situations Lesitha and Thomas (2012) makes N cycles of RSS. For details see Chen *et al.* (2004). Suppose α_i^* and α_i^{**} denote the estimators of α corresponding to α^* and α^{**} respectively, based on the *ith* cycle. Then, estimators of α based on N cycles are given by:

$$\overline{\alpha^*} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i^*\,,$$

and

$$\overline{\alpha^{**}} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i^{**}$$

with

$$Var\left(\overline{\alpha^*}\right) = \frac{\alpha^2}{Nn^2}\sum_{r=1}^{N}\left[\frac{\sigma_{r,r:n}}{\gamma_{r:n}^2}\right],$$

and

$$Var\left(\overline{\alpha^{**}}\right) = \frac{\left(\gamma' G_1^{-1}\gamma\right)^{-1}\alpha^2}{N}\,.$$

Median ranked set sampling (MRSS) was first introduced by Muttlak (1997) to estimate the mean of a normal distribution. In general, MRSS is applied as a modification of RSS when one is interested in estimating a parameter associated with the central tendency of a distribution. The procedures of MRSS are given as: Select n independent samples each with n units as in the case of RSS. Then rank the units in each sample either by judgement method or by using some inexpensive means without having actual measurement on the unit. Lesitha and Thomas (2012) used MRSS method to estimate α as:

$$t_4 = \tilde{\alpha} = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{r=1}^{n}\left[\dfrac{X_{(m:n)r}}{\gamma_{m:n}}\right] & ;\, when\ m = \dfrac{n+1}{2} and\, n\ is\, odd \\ \left(\gamma_1' G_2^{-1}\gamma_1\right)^{-1}\gamma_1' G_2^{-1} X_{mrss} & ;when\ m = \dfrac{n}{2} and\, n\ is\, even \end{cases}$$

with

$$Var(t_4)=\begin{cases}\frac{1}{n}\left[\frac{\sigma_{m,m:n}}{\gamma^2_{m:n}}\right]\alpha^2 & ;\ when\ m=\frac{n+1}{2} and\, n\ is\, odd\\ \left(\gamma_1^{'}G_2^{-1}\gamma_1\right)^{-1}\alpha^2 & ;when\ m=\frac{n}{2} and\, n\ is\, even\end{cases}=\alpha^2V_4, \tag{15}$$

where

$$X_{mrss}=\left(X_{(m:n)1},X_{(m+1:n)2},X_{(m:n)3},X_{(m+1:n)4},...,X_{(m+1:n)n}\right)'$$

$$\gamma_1=\left(\gamma_{m:n},\gamma_{m+1:n},\gamma_{m:n},\gamma_{m+1:n},...,\gamma_{m+1:n}\right)'$$

$$G_2=diag\left(\sigma_{m,m:n},\sigma_{m+1,m+1:n},\sigma_{m,m:n},\sigma_{m+1,m+1:n},...,\sigma_{m+1,m+1:n}\right) \text{ and}$$

$$V_4=\begin{cases}\frac{1}{n}\left[\frac{\sigma_{m,m:n}}{\gamma^2_{m:n}}\right] & ;\ when\ m=\frac{n+1}{2} and\, n\ is\, odd\\ \left(\gamma_1^{'}G_2^{-1}\gamma_1\right)^{-1} & ;when\ m=\frac{n}{2} and\, n\ is\, even\end{cases}.$$

2. Improved estimation of the scale parameter α

Let $t_i, i=1,2,3,4$ be an unbiased estimator of the parameter α, then we define a class of estimators for α as

$$T_i=A_i t_i, i=1,2,3,4,$$

where $A_i's, i=1,2,3,4$ are suitably chosen constants such that mean squared error of the estimators $T_i's, i=1,2,3,4$ is minimum.

The biases and mean squared errors (MSEs) of $T_i, i=1,2,3,4$ are respectively given by

$$B(T_i)=\alpha(A_i-1),$$

and

$$MSE(T_i)=A_i^2Var(t_i)+(A_i-1)^2\alpha^2=\alpha^2\left[A_i^2(1+V_i)-2A_i+1\right].$$

The $MSE(T_i), i=1,2,3,4$ is minimized for

$$A_i=(1+V_i)^{-1}, i=1,2,3,4.$$

Thus, the resulting minimum MSE estimator of α is given by

$$T_{0i}=t_i(1+V_i)^{-1}, i=1,2,3,4.$$

The biases and MSEs of $T_{0i}, i = 1,2,3,4$ are respectively given as

$$B(T_{0i}) = -\alpha\left(\frac{V_i}{1+V_i}\right) , \tag{16}$$

and

$$MSE(T_{0i}) = \alpha^2\left(\frac{V_i}{1+V_i}\right). \tag{17}$$

We have from (12) - (15) and (17) that

$$Var(t_i) - MSE(T_{0i}) = \alpha^2 \frac{V_i^2}{(1+V_i)} > 0, i = 1,2,3,4 . \tag{18}$$

It follows from (18) that the proposed MMSE estimators $T_{0i}'s, i = 1,2,3,4$ are better than the corresponding usual unbiased estimators $t_i's, i = 1,2,3,4$.

3. Improved estimation of the scale parameter α with prior information

Let $t_i, i = 1,2,3,4$ be an unbiased estimator of the parameter α, then we define a class of estimators of α using the prior point estimate α_0 of α as

$$T_{1i} = \alpha_0 + B_i t_i , i = 1,2,3,4 , \tag{19}$$

where $B_i's, i = 1,2,3,4$ are suitably chosen constants such that mean squared error of the estimators $T_{1i}'s, i = 1,2,3,4$ are minimum.

The biases and mean squared errors (MSEs) of $T_{1i}, i = 1,2,3,4$ are respectively given by

$$B(T_{1i}) = \alpha(\phi + B_i) ,$$

and

$$MSE(T_{1i}) = \alpha^2\left[\phi^2 + B_i^2(1+V_i) + 2\phi B_i\right].$$

where $\phi = \left(\frac{\alpha_0}{\alpha} - 1\right) = (\lambda - 1)$ with $\lambda = \frac{\alpha_0}{\alpha}$.

The $MSE(T_{1i}), i = 1,2,3,4$ is minimized for

$$B_i = -\phi(1+V_i)^{-1}, i = 1,2,3,4 . \tag{20}$$

The value of $B_i, i = 1,2,3,4$ at (20) depends on the unknown parameter α, so an estimate of $B_i, i = 1,2,3,4$ based on sample data is given by

$$B_i^* = -\frac{\phi^*}{(1+V_i)} = -\frac{(\theta_{20} - t_i)}{t_i(1+V_i)}, i = 1,2,3,4 .$$

Putting $B_i^*, i = 1,2,3,4$ in (19), we get a shrinkage estimator of $t_i{'}s, i = 1,2,3,4$ as

$$T_{1i}^* = \alpha_0 - (1+V_i)^{-1}(\alpha_0 - t_i), i = 1,2,3,4 . \tag{21}$$

The biases and mean squared errors (MSEs) of the estimators $T_{1i}^*{'}s, i = 1,2,3,4$ are respectively given by

$$B(T_{1i}^*) = \alpha\phi\left(\frac{V_i}{1+V_i}\right), \tag{22}$$

and

$$MSE(T_{1i}^*) = \alpha^2 \frac{V_i(\phi^2 V_i + 1)}{(1+V_i)^2} . \tag{23}$$

Comparisons of the proposed shrinkage estimators $T_{1i}^*{'}s, i = 1,2,3,4$ with that of corresponding usual unbiased estimators $t_i{'}s, i = 1,2,3,4$ are given in the following Theorem 1.

Theorem 1: *The proposed shrinkage estimators* $T_{1i}^*{'}s, i = 1,2,3,4$ *are better than the corresponding usual unbiased estimators* $t_i{'}s, i = 1,2,3,4$ *if*

$$\lambda \in \left(0, \left(1 + \sqrt{(2+V_i)}\right)\right),$$

i.e. if $\alpha_0 \in \left(0, \alpha\left\{1 + \sqrt{(2+V_i)}\right\}\right)$,

i.e. if $\alpha \in \left(\frac{\alpha_0}{\left\{1 + \sqrt{(2+V_i)}\right\}}, \infty\right)$.

Proof: *From (12) - (15) and (23), we have that*

$MSE(T_{1i}^*) < Var(t_i), i = 1,2,3,4$ *if*

$$\alpha^2 \frac{V_i(\phi^2 V_i + 1)}{(1+V_i)^2} < \alpha^2 V_i,$$

i.e. if $\frac{(\phi^2 V_i + 1)}{(1+V_i)^2} < 1$, *i.e. if* $\phi^2 < \dfrac{\left((1+V_i)^2-1\right)}{V_i}$,

i.e. if $\phi^2 < \dfrac{V_i(2+V_i)}{V_i}$,

i.e. if $\phi^2 < 2+V_i$,

i.e. if $(\lambda-1)^2 < 2+V_i$,

$$\text{i.e. if } \left\{1-\sqrt{(2+V_i)}\right\} < \lambda < \left\{1+\sqrt{(2+V_i)}\right\}. \tag{24}$$

Since $\left\{1-\sqrt{(2+V_i)}\right\} < 0$ *and* $\lambda(=\alpha_0/\alpha)$ *cannot be negative therefore (24) reduces to*

$$0<\lambda<\left\{1+\sqrt{(2+V_i)}\right\},$$

or

$$0<\alpha_0<\alpha\left\{1+\sqrt{(2+V_i)}\right\},$$

or

$$\frac{\alpha_0}{\left\{1+\sqrt{(2+V_i)}\right\}}<\alpha<\infty.$$

Hence the theorem. ♦

Further, we have compared the proposed shrinkage estimators $T_{1i}^*{}'s, i=1,2,3,4$ with that of corresponding MMSE estimators $T_{0i}{}'s, i=1,2,3,4$ and the results are presented in Theorem 2.

Theorem 2: *The proposed shrinkage estimators* $T_{1i}^*{}'s, i=1,2,3,4$ *are better than the corresponding MMSE estimators* $T_{0i}{}'s, i=1,2,3,4$ *if*

$\lambda \in (0,2)$,

i.e. if $\alpha_0 \in (0,2\alpha)$,

i.e. if $\alpha \in \left(\dfrac{\alpha_0}{2},\infty\right)$.

Proof: *From (17) and (23) we have that*

$$MSE\left(T_{1i}^*\right) < MSE\left(T_{0i}\right), i=1,2,...,7 \text{ if}$$

$$\alpha^2\,\frac{V_i\left(\phi^2V_i+1\right)}{(1+V_i)^2}<\alpha^2\,\frac{V_i}{1+V_i},$$

$$\text{i.e. if} \quad \frac{\left(\phi^2 V_i + 1\right)}{\left(1 + V_i\right)} < 1,$$

$$\text{i.e. if} \quad \phi^2 < 1,$$

$$\text{i.e. if} \quad -1 < \phi < 1,$$

$$\text{i.e. if} \quad 0 < \lambda < 2,$$

$$\text{or} \quad 0 < \alpha_0 < 2\alpha,$$

$$\text{or} \quad \frac{\alpha_0}{2} < \alpha < \infty .$$

Hence the theorem. ♦

4. Relative efficiencies

We have computed the relative efficiencies of various suggested estimators to usual estimators by using the formulae:

$$e_1 = RE\left(T_{01}, t_1\right) = 1 + V_1;$$

$$e_2 = RE\left(T_{02}, t_2\right) = 1 + V_2;$$

$$e_3 = RE\left(T_{03}, t_3\right) = 1 + V_3;$$

$$e_4 = RE\left(T_{04}, t_4\right) = 1 + V_4;$$

$$e_5 = RE\left(T_{04}, T_{01}\right) = \frac{V_1\left(1 + V_4\right)}{V_4\left(1 + V_1\right)};$$

$$e_6 = RE\left(T_{04}, T_{02}\right) = \frac{V_2\left(1 + V_4\right)}{V_4\left(1 + V_2\right)};$$

$$e_7 = RE\left(T_{04}, T_{03}\right) = \frac{V_3\left(1 + V_4\right)}{V_4\left(1 + V_3\right)};$$

$$e_8 = RE\left(T_{11}^*, t_1\right) = \frac{\left(1 + V_1\right)^2}{\left(\phi^2 V_1 + 1\right)};$$

$$e_9 = RE\left(T_{11}^*, T_{01}\right) = \frac{\left(1 + V_1\right)}{\left(\phi^2 V_1 + 1\right)};$$

$$e_{10} = RE\left(T_{12}^*, t_2\right) = \frac{\left(1 + V_2\right)^2}{\left(\phi^2 V_2 + 1\right)};$$

$$e_{11} = RE\left(T_{12}^*, T_{02}\right) = \frac{\left(1 + V_2\right)}{\left(\phi^2 V_2 + 1\right)};$$

$$e_{12} = RE\left(T_{13}^*, t_3\right) = \frac{\left(1 + V_3\right)^2}{\left(\phi^2 V_3 + 1\right)};$$

$$e_{13} = RE\left(T_{13}^*, T_{03}\right) = \frac{\left(1 + V_3\right)}{\left(\phi^2 V_3 + 1\right)};$$

$$e_{14} = RE\left(T_{14}^*, t_4\right) = \frac{\left(1 + V_4\right)^2}{\left(\phi^2 V_4 + 1\right)};$$

and

$$e_{15} = RE\left(T_{14}^*, T_{04}\right) = \frac{\left(1 + V_4\right)}{\left(\phi^2 V_4 + 1\right)}.$$

- The values of $e_i, i = 1,2,\ldots,7$ are shown in Table 1 for $n = 2(1)8$ and $\beta = 2.5(0.5)5$.
- The values of $e_i, i = 8,9,\ldots,15$ are shown in Tables 2 to 5 for $n = 2(1)8;$ $\beta = 2.5(0.5)5$ and different values of $\lambda = \frac{\alpha_0}{\alpha}$.

Table 1. The values of $e_i's, i=1,2,...,7$.

n	*β*	e_1	e_2	e_3	e_4	e_5	e_6	e_7
2	2.5	1.3371	1.4025	1.2799	1.2799	1.1530	1.3124	1.0000
	3.0	1.2158	1.1978	1.1674	1.1674	1.2381	1.1516	1.0000
	3.5	1.1514	1.1254	1.1135	1.1135	1.2901	1.0932	1.0000
	4.0	1.1126	1.0885	1.0827	1.0827	1.3242	1.0643	1.0000
	4.5	1.0872	1.0665	1.0633	1.0633	1.3476	1.0474	1.0000
	5.0	1.0697	1.0521	1.0501	1.0501	1.3644	1.0368	1.0000
3	2.5	1.2011	1.1904	1.1180	1.0832	2.1798	2.0828	1.3740
	3.0	1.1308	1.0970	1.0752	1.0543	2.2446	1.7159	1.3572
	3.5	1.0937	1.0626	1.0526	1.0385	2.3093	1.5893	1.3475
	4.0	1.0706	1.0447	1.0391	1.0288	2.3527	1.5272	1.3422
	4.5	1.0552	1.0338	1.0303	1.0224	2.3815	1.4910	1.3382
	5.0	1.0443	1.0266	1.0242	1.0180	2.4033	1.4675	1.3356
4	2.5	1.1392	1.1118	1.0648	1.0477	2.6846	2.2089	1.3376
	3.0	1.0939	1.0582	1.0424	1.0317	2.7929	1.7903	1.3246
	3.5	1.0678	1.0380	1.0301	1.0227	2.8598	1.6495	1.3176
	4.0	1.0514	1.0273	1.0226	1.0171	2.9034	1.5803	1.3126
	4.5	1.0413	1.0208	1.0176	1.0134	3.0062	1.5408	1.3102
	5.0	1.0324	1.0164	1.0141	1.0107	2.9570	1.5158	1.3085
5	2.5	1.1078	1.0740	1.0409	1.0278	3.5999	2.5482	1.4547
	3.0	1.0733	1.0391	1.0272	1.0187	3.7126	2.0476	1.4397
	3.5	1.0531	1.0257	1.0195	1.0135	3.7819	1.8805	1.4317
	4.0	1.0403	1.0186	1.0147	1.0101	3.8704	1.8188	1.4421
	4.5	1.0316	1.0141	1.0115	1.0080	3.8489	1.7503	1.4223
	5.0	1.0256	1.0112	1.0092	1.0065	3.8785	1.7199	1.4196
6	2.5	1.0880	1.0528	1.0282	1.0196	4.2168	2.6150	1.4288
	3.0	1.0600	1.0282	1.0189	1.0133	4.3313	2.0988	1.4170
	3.5	1.0437	1.0187	1.0136	1.0096	4.3995	1.9265	1.4101
	4.0	1.0332	1.0135	1.0103	1.0073	4.4473	1.8430	1.4065
	4.5	1.0261	1.0103	1.0080	1.0057	4.4756	1.7943	1.4024
	5.0	1.0211	1.0082	1.0065	1.0046	4.5010	1.7638	1.4009
7	2.5	1.0735	1.0397	1.0206	1.0138	5.0286	2.8038	1.4800
	3.0	1.0509	1.0214	1.0139	1.0094	5.1991	2.2498	1.4690
	3.5	1.0372	1.0147	1.0106	1.0068	5.2926	2.1298	1.5419
	4.0	1.0282	1.0103	1.0076	1.0052	5.3054	1.9708	1.4542
	4.5	1.0222	1.0079	1.0060	1.0041	5.3447	1.9217	1.4556
	5.0	1.0179	1.0062	1.0048	1.0033	5.3582	1.8854	1.4494
8	2.5	1.0643	1.0310	1.0157	1.0107	5.7315	2.8525	1.4632
	3.0	1.0444	1.0168	1.0106	1.0073	5.8758	2.2888	1.4526
	3.5	1.0329	1.0112	1.0077	1.0053	6.0530	2.1067	1.4484
	4.0	1.0245	1.0081	1.0058	1.0040	5.9651	2.0116	1.4441
	4.5	1.0190	1.0062	1.0046	1.0032	5.9067	1.9593	1.4396
	5.0	1.0156	1.0049	1.0037	1.0025	6.0905	1.9479	1.4568

Table 2. The values of e_i for $i = 8 \; and \;\; 9$

n	β	$e_8 = RE\left(T_{11}^*, t_1\right)$						Range of λ in which T_{11}^* is efficient to t_1
		$\lambda = 1.00$	$\lambda = 1.25$ and $\lambda = 0.75$	$\lambda = 1.50$ and $\lambda = 0.50$	$\lambda = 1.75$ and $\lambda = 0.25$	$\lambda = 2.00$	$\lambda = 2.25$	
2	2.5	1.7878	1.7509	1.6489	1.5028	1.3371	1.1710	(0,2.53)
	3.0	1.4782	1.4586	1.4026	1.3182	1.2158	1.1054	(0,2.49)
	3.5	1.3257	1.3133	1.2773	1.2217	1.1514	1.0721	(0,2.47)
	4.0	1.2378	1.2292	1.2039	1.1641	1.1126	1.0527	(0,2.45)
	4.5	1.1820	1.1756	1.1568	1.1268	1.0872	1.0403	(0,2.44)
	5.0	1.1442	1.1392	1.1246	1.1010	1.0697	1.0319	(0,2.44)
3	2.5	1.4426	1.4247	1.3735	1.2960	1.2011	1.0977	(0,2.48)
	3.0	1.2786	1.2683	1.2382	1.1910	1.1308	1.0617	(0,2.46)
	3.5	1.1961	1.1892	1.1688	1.1363	1.0937	1.0434	(0,2.45)
	4.0	1.1461	1.1411	1.1263	1.1024	1.0706	1.0323	(0,2.44)
	4.5	1.1133	1.1095	1.0982	1.0798	1.0552	1.0250	(0,2.43)
	5.0	1.0906	1.0876	1.0787	1.0641	1.0443	1.0200	(0,2.43)
4	2.5	1.2977	1.2865	1.2541	1.2035	1.1392	1.0659	(0,2.46)
	3.0	1.1966	1.1896	1.1692	1.1366	1.0939	1.0435	(0,2.45)
	3.5	1.1402	1.1354	1.1212	1.0983	1.0678	1.0310	(0,2.44)
	4.0	1.1053	1.1018	1.0913	1.0743	1.0514	1.0232	(0,2.43)
	4.5	1.0843	1.0815	1.0732	1.0597	1.0413	1.0186	(0,2.43)
	5.0	1.0659	1.0638	1.0574	1.0468	1.0324	1.0145	(0,2.43)
5	2.5	1.2272	1.2190	1.1950	1.1570	1.1078	1.0503	(0,2.45)
	3.0	1.1519	1.1466	1.1312	1.1063	1.0733	1.0336	(0,2.44)
	3.5	1.1091	1.1054	1.0945	1.0769	1.0531	1.0241	(0,2.43)
	4.0	1.0823	1.0796	1.0715	1.0583	1.0403	1.0181	(0,2.43)
	4.5	1.0643	1.0622	1.0559	1.0457	1.0316	1.0141	(0,2.43)
	5.0	1.0518	1.0501	1.0451	1.0369	1.0256	1.0114	(0,2.42)
6	2.5	1.1837	1.1772	1.1582	1.1279	1.0880	1.0406	(0,2.44)
	3.0	1.1237	1.1195	1.1071	1.0870	1.0600	1.0273	(0,2.44)
	3.5	1.0892	1.0863	1.0775	1.0631	1.0437	1.0197	(0,2.43)
	4.0	1.0675	1.0653	1.0587	1.0479	1.0332	1.0149	(0,2.43)
	4.5	1.0529	1.0512	1.0461	1.0377	1.0261	1.0116	(0,2.42)
	5.0	1.0426	1.0413	1.0372	1.0304	1.0211	1.0094	(0,2.42)
7	2.5	1.1524	1.1471	1.1316	1.1066	1.0735	1.0337	(0,2.44)
	3.0	1.1043	1.1008	1.0905	1.0736	1.0509	1.0230	(0,2.43)
	3.5	1.0759	1.0734	1.0659	1.0538	1.0372	1.0167	(0,2.43)
	4.0	1.0572	1.0554	1.0498	1.0407	1.0282	1.0126	(0,2.42)
	4.5	1.0449	1.0434	1.0391	1.0320	1.0222	1.0099	(0,2.42)
	5.0	1.0362	1.0350	1.0316	1.0258	1.0179	1.0079	(0,2.42)
8	2.5	1.1327	1.1282	1.1148	1.0932	1.0643	1.0293	(0,2.44)
	3.0	1.0907	1.0877	1.0787	1.0641	1.0444	1.0200	(0,2.43)
	3.5	1.0669	1.0647	1.0582	1.0475	1.0329	1.0147	(0,2.43)
	4.0	1.0497	1.0481	1.0433	1.0354	1.0245	1.0109	(0,2.42)
	4.5	1.0384	1.0372	1.0335	1.0274	1.0190	1.0084	(0,2.42)
	5.0	1.0315	1.0305	1.0275	1.0225	1.0156	1.0069	(0,2.42)

Table 2. The values of e_i for $i = 8\ and\ 9$ (cont.)

n	β	$e_9 = RE\left(T^*_{11}, T_{01}\right)$					Range of λ in which T^*_{11} is efficient to T_{01}
		$\lambda = 1.00$	$\lambda = 1.25$ and $\lambda = 0.75$	$\lambda = 1.50$ and $\lambda = 0.50$	$\lambda = 1.75$ and $\lambda = 0.25$	$\lambda = 2.00$ and $\lambda = 0.00$	
2	2.5	1.3371	1.3095	1.2332	1.1240	1.0000	[0,2]
	3.0	1.2158	1.1996	1.1536	1.0842	1.0000	[0,2]
	3.5	1.1514	1.1406	1.1094	1.0610	1.0000	[0,2]
	4.0	1.1126	1.1048	1.0821	1.0463	1.0000	[0,2]
	4.5	1.0872	1.0813	1.0640	1.0364	1.0000	[0,2]
	5.0	1.0697	1.0650	1.0514	1.0293	1.0000	[0,2]
3	2.5	1.2011	1.1862	1.1436	1.0790	1.0000	[0,2]
	3.0	1.1308	1.1216	1.0950	1.0533	1.0000	[0,2]
	3.5	1.0937	1.0873	1.0687	1.0389	1.0000	[0,2]
	4.0	1.0706	1.0659	1.0520	1.0297	1.0000	[0,2]
	4.5	1.0552	1.0515	1.0408	1.0234	1.0000	[0,2]
	5.0	1.0443	1.0414	1.0329	1.0189	1.0000	[0,2]
4	2.5	1.1392	1.1294	1.1009	1.0565	1.0000	[0,2]
	3.0	1.0939	1.0875	1.0688	1.0390	1.0000	[0,2]
	3.5	1.0678	1.0633	1.0500	1.0286	1.0000	[0,2]
	4.0	1.0514	1.0480	1.0380	1.0218	1.0000	[0,2]
	4.5	1.0413	1.0386	1.0307	1.0177	1.0000	[0,2]
	5.0	1.0324	1.0304	1.0241	1.0139	1.0000	[0,2]
5	2.5	1.1078	1.1004	1.0787	1.0445	1.0000	[0,2]
	3.0	1.0733	1.0684	1.0540	1.0308	1.0000	[0,2]
	3.5	1.0531	1.0496	1.0393	1.0226	1.0000	[0,2]
	4.0	1.0403	1.0377	1.0300	1.0173	1.0000	[0,2]
	4.5	1.0316	1.0296	1.0235	1.0136	1.0000	[0,2]
	5.0	1.0256	1.0239	1.0191	1.0110	1.0000	[0,2]
6	2.5	1.0880	1.0820	1.0646	1.0367	1.0000	[0,2]
	3.0	1.0600	1.0561	1.0444	1.0254	1.0000	[0,2]
	3.5	1.0437	1.0408	1.0324	1.0186	1.0000	[0,2]
	4.0	1.0332	1.0311	1.0247	1.0143	1.0000	[0,2]
	4.5	1.0261	1.0244	1.0195	1.0113	1.0000	[0,2]
	5.0	1.0211	1.0197	1.0157	1.0091	1.0000	[0,2]
7	2.5	1.0735	1.0686	1.0541	1.0309	1.0000	[0,2]
	3.0	1.0509	1.0475	1.0377	1.0216	1.0000	[0,2]
	3.5	1.0372	1.0348	1.0277	1.0160	1.0000	[0,2]
	4.0	1.0282	1.0264	1.0210	1.0122	1.0000	[0,2]
	4.5	1.0222	1.0208	1.0166	1.0096	1.0000	[0,2]
	5.0	1.0179	1.0168	1.0134	1.0078	1.0000	[0,2]
8	2.5	1.0643	1.0600	1.0474	1.0271	1.0000	[0,2]
	3.0	1.0444	1.0415	1.0329	1.0189	1.0000	[0,2]
	3.5	1.0329	1.0308	1.0245	1.0141	1.0000	[0,2]
	4.0	1.0245	1.0230	1.0183	1.0106	1.0000	[0,2]
	4.5	1.0190	1.0178	1.0142	1.0082	1.0000	[0,2]
	5.0	1.0156	1.0146	1.0117	1.0068	1.0000	[0,2]

Table 3. The values of e_i for $i = 10 \; and \; 11$

n	β	$e_{10} = RE\left(T_{12}^{*}, t_2\right)$						Range of λ in which T_{12}^{*} is efficient to t_2
		$\lambda = 1.00$	$\lambda = 1.25$ and $\lambda = 0.75$	$\lambda = 1.50$ and $\lambda = 0.50$	$\lambda = 1.75$ and $\lambda = 0.25$	$\lambda = 2.00$	$\lambda = 2.25$	
2	2.5	1.9669	1.9186	1.7871	1.6038	1.4025	1.2075	(0,2.55)
	3.0	1.4347	1.4171	1.3671	1.2910	1.1978	1.0960	(0,2.48)
	3.5	1.2665	1.2566	1.2280	1.1830	1.1254	1.0590	(0,2.46)
	4.0	1.1849	1.1784	1.1592	1.1287	1.0885	1.0409	(0,2.45)
	4.5	1.1374	1.1327	1.1188	1.0964	1.0665	1.0304	(0,2.44)
	5.0	1.1069	1.1033	1.0926	1.0754	1.0521	1.0236	(0,2.43)
3	2.5	1.4171	1.4004	1.3527	1.2800	1.1904	1.0922	(0,2.48)
	3.0	1.2034	1.1961	1.1749	1.1411	1.0970	1.0450	(0,2.45)
	3.5	1.1292	1.1248	1.1118	1.0908	1.0626	1.0285	(0,2.44)
	4.0	1.0914	1.0884	1.0794	1.0646	1.0447	1.0202	(0,2.43)
	4.5	1.0688	1.0665	1.0598	1.0488	1.0338	1.0151	(0,2.43)
	5.0	1.0539	1.0522	1.0470	1.0384	1.0266	1.0119	(0,2.42)
4	2.5	1.2360	1.2274	1.2024	1.1629	1.1118	1.0523	(0,2.45)
	3.0	1.1199	1.1158	1.1038	1.0843	1.0582	1.0265	(0,2.43)
	3.5	1.0775	1.0749	1.0673	1.0549	1.0380	1.0171	(0,2.43)
	4.0	1.0554	1.0536	1.0482	1.0394	1.0273	1.0122	(0,2.42)
	4.5	1.0419	1.0406	1.0366	1.0299	1.0208	1.0092	(0,2.42)
	5.0	1.0330	1.0320	1.0288	1.0236	1.0164	1.0072	(0,2.42)
5	2.5	1.1534	1.1481	1.1325	1.1073	1.0740	1.0339	(0,2.44)
	3.0	1.0798	1.0771	1.0693	1.0565	1.0391	1.0176	(0,2.43)
	3.5	1.0521	1.0504	1.0454	1.0371	1.0257	1.0115	(0,2.42)
	4.0	1.0375	1.0363	1.0327	1.0267	1.0186	1.0082	(0,2.42)
	4.5	1.0285	1.0276	1.0249	1.0204	1.0141	1.0062	(0,2.42)
	5.0	1.0225	1.0218	1.0196	1.0161	1.0112	1.0049	(0,2.42)
6	2.5	1.1084	1.1047	1.0939	1.0764	1.0528	1.0239	(0,2.43)
	3.0	1.0572	1.0554	1.0498	1.0407	1.0282	1.0126	(0,2.42)
	3.5	1.0377	1.0365	1.0328	1.0269	1.0187	1.0083	(0,2.42)
	4.0	1.0272	1.0263	1.0237	1.0194	1.0135	1.0060	(0,2.42)
	4.5	1.0207	1.0201	1.0181	1.0148	1.0103	1.0045	(0,2.42)
	5.0	1.0164	1.0159	1.0143	1.0117	1.0082	1.0036	(0,2.42)
7	2.5	1.0809	1.0783	1.0703	1.0573	1.0397	1.0178	(0,2.43)
	3.0	1.0433	1.0419	1.0377	1.0308	1.0214	1.0095	(0,2.42)
	3.5	1.0295	1.0286	1.0258	1.0211	1.0147	1.0065	(0,2.42)
	4.0	1.0207	1.0200	1.0181	1.0148	1.0103	1.0045	(0,2.42)
	4.5	1.0158	1.0153	1.0138	1.0113	1.0079	1.0035	(0,2.42)
	5.0	1.0125	1.0121	1.0109	1.0090	1.0062	1.0027	(0,2.42)
8	2.5	1.0629	1.0609	1.0548	1.0447	1.0310	1.0138	(0,2.43)
	3.0	1.0339	1.0328	1.0296	1.0242	1.0168	1.0074	(0,2.42)
	3.5	1.0225	1.0218	1.0197	1.0161	1.0112	1.0049	(0,2.42)
	4.0	1.0163	1.0158	1.0143	1.0117	1.0081	1.0036	(0,2.42)
	4.5	1.0125	1.0121	1.0109	1.0090	1.0062	1.0027	(0,2.42)
	5.0	1.0099	1.0096	1.0087	1.0071	1.0049	1.0022	(0,2.42)

Table 3. The values of e_i for $i = 10\ and\ 11$ (cont.)

		$e_{11} = RE\left(T^*_{12}, T_{02}\right)$					Range of λ in which T^*_{12} is efficient to T_{02}
n	***β***	***λ = 1.00***	***λ = 1.25*** and ***λ = 0.75***	***λ = 1.50*** and ***λ = 0.50***	***λ = 1.75*** and ***λ = 0.25***	***λ = 2.00*** and ***λ = 0.00***	
2	2.5	1.4025	1.3680	1.2743	1.1436	1.0000	[0,2]
	3.0	1.1978	1.1831	1.1413	1.0779	1.0000	[0,2]
	3.5	1.1254	1.1166	1.0912	1.0512	1.0000	[0,2]
	4.0	1.0885	1.0825	1.0650	1.0369	1.0000	[0,2]
	4.5	1.0665	1.0621	1.0491	1.0280	1.0000	[0,2]
	5.0	1.0521	1.0487	1.0386	1.0221	1.0000	[0,2]
3	2.5	1.1904	1.1764	1.1363	1.0752	1.0000	[0,2]
	3.0	1.0970	1.0904	1.0710	1.0402	1.0000	[0,2]
	3.5	1.0626	1.0585	1.0463	1.0265	1.0000	[0,2]
	4.0	1.0447	1.0418	1.0332	1.0191	1.0000	[0,2]
	4.5	1.0338	1.0316	1.0252	1.0145	1.0000	[0,2]
	5.0	1.0266	1.0249	1.0198	1.0115	1.0000	[0,2]
4	2.5	1.1118	1.1040	1.0815	1.0460	1.0000	[0,2]
	3.0	1.0582	1.0544	1.0430	1.0247	1.0000	[0,2]
	3.5	1.0380	1.0356	1.0282	1.0163	1.0000	[0,2]
	4.0	1.0273	1.0256	1.0203	1.0118	1.0000	[0,2]
	4.5	1.0208	1.0194	1.0155	1.0090	1.0000	[0,2]
	5.0	1.0164	1.0153	1.0122	1.0071	1.0000	[0,2]
5	2.5	1.0740	1.0690	1.0545	1.0311	1.0000	[0,2]
	3.0	1.0391	1.0366	1.0291	1.0167	1.0000	[0,2]
	3.5	1.0257	1.0241	1.0192	1.0111	1.0000	[0,2]
	4.0	1.0186	1.0174	1.0139	1.0080	1.0000	[0,2]
	4.5	1.0141	1.0132	1.0106	1.0061	1.0000	[0,2]
	5.0	1.0112	1.0105	1.0084	1.0049	1.0000	[0,2]
6	2.5	1.0528	1.0493	1.0391	1.0224	1.0000	[0,2]
	3.0	1.0282	1.0264	1.0210	1.0122	1.0000	[0,2]
	3.5	1.0187	1.0175	1.0139	1.0081	1.0000	[0,2]
	4.0	1.0135	1.0126	1.0101	1.0059	1.0000	[0,2]
	4.5	1.0103	1.0097	1.0077	1.0045	1.0000	[0,2]
	5.0	1.0082	1.0076	1.0061	1.0036	1.0000	[0,2]
7	2.5	1.0397	1.0371	1.0295	1.0170	1.0000	[0,2]
	3.0	1.0214	1.0200	1.0160	1.0093	1.0000	[0,2]
	3.5	1.0147	1.0137	1.0110	1.0064	1.0000	[0,2]
	4.0	1.0103	1.0097	1.0077	1.0045	1.0000	[0,2]
	4.5	1.0079	1.0074	1.0059	1.0034	1.0000	[0,2]
	5.0	1.0062	1.0058	1.0047	1.0027	1.0000	[0,2]
8	2.5	1.0310	1.0290	1.0231	1.0133	1.0000	[0,2]
	3.0	1.0168	1.0158	1.0126	1.0073	1.0000	[0,2]
	3.5	1.0112	1.0105	1.0084	1.0049	1.0000	[0,2]
	4.0	1.0081	1.0076	1.0061	1.0035	1.0000	[0,2]
	4.5	1.0062	1.0058	1.0047	1.0027	1.0000	[0,2]
	5.0	1.0049	1.0046	1.0037	1.0022	1.0000	[0,2]

Table 4. The values of e_i for $i = 12 \ and \ 13$

n	β	$e_{12} = RE\left(T_{13}^{*}, t_3\right)$						Range of λ in which T_{13}^{*} is efficient to t_3
		$\lambda = 1.00$	$\lambda = 1.25$ and $\lambda = 0.75$	$\lambda = 1.50$ and $\lambda = 0.50$	$\lambda = 1.75$ and $\lambda = 0.25$	$\lambda = 2.00$	$\lambda = 2.25$	
2	2.5	1.6380	1.6099	1.5309	1.4152	1.2799	1.1397	(0,2.51)
	3.0	1.3628	1.3487	1.3080	1.2455	1.1674	1.0803	(0,2.47)
	3.5	1.2398	1.2311	1.2056	1.1654	1.1135	1.0531	(0,2.45)
	4.0	1.1723	1.1663	1.1485	1.1202	1.0827	1.0381	(0,2.44)
	4.5	1.1306	1.1262	1.1130	1.0917	1.0633	1.0288	(0,2.44)
	5.0	1.1028	1.0993	1.0891	1.0725	1.0501	1.0227	(0,2.43)
3	2.5	1.2499	1.2407	1.2141	1.1721	1.1180	1.0553	(0,2.46)
	3.0	1.1560	1.1506	1.1347	1.1091	1.0752	1.0345	(0,2.44)
	3.5	1.1080	1.1044	1.0936	1.0761	1.0526	1.0238	(0,2.43)
	4.0	1.0797	1.0771	1.0692	1.0565	1.0391	1.0176	(0,2.43)
	4.5	1.0614	1.0594	1.0535	1.0437	1.0303	1.0135	(0,2.42)
	5.0	1.0489	1.0473	1.0426	1.0348	1.0242	1.0107	(0,2.42)
4	2.5	1.1338	1.1293	1.1158	1.0940	1.0648	1.0296	(0,2.44)
	3.0	1.0867	1.0838	1.0753	1.0613	1.0424	1.0191	(0,2.43)
	3.5	1.0612	1.0592	1.0533	1.0435	1.0301	1.0135	(0,2.42)
	4.0	1.0457	1.0442	1.0398	1.0326	1.0226	1.0100	(0,2.42)
	4.5	1.0355	1.0344	1.0310	1.0253	1.0176	1.0078	(0,2.42)
	5.0	1.0284	1.0275	1.0248	1.0203	1.0141	1.0062	(0,2.42)
5	2.5	1.0835	1.0808	1.0726	1.0592	1.0409	1.0184	(0,2.43)
	3.0	1.0551	1.0533	1.0480	1.0392	1.0272	1.0121	(0,2.42)
	3.5	1.0393	1.0381	1.0343	1.0281	1.0195	1.0086	(0,2.42)
	4.0	1.0295	1.0286	1.0258	1.0211	1.0147	1.0065	(0,2.42)
	4.5	1.0231	1.0223	1.0201	1.0165	1.0115	1.0051	(0,2.42)
	5.0	1.0185	1.0179	1.0162	1.0133	1.0092	1.0041	(0,2.42)
6	2.5	1.0571	1.0553	1.0497	1.0406	1.0282	1.0126	(0,2.42)
	3.0	1.0381	1.0369	1.0332	1.0272	1.0189	1.0084	(0,2.42)
	3.5	1.0274	1.0265	1.0239	1.0196	1.0136	1.0060	(0,2.42)
	4.0	1.0206	1.0200	1.0180	1.0148	1.0103	1.0045	(0,2.42)
	4.5	1.0161	1.0156	1.0141	1.0116	1.0080	1.0035	(0,2.42)
	5.0	1.0130	1.0126	1.0113	1.0093	1.0065	1.0028	(0,2.42)
7	2.5	1.0415	1.0402	1.0362	1.0296	1.0206	1.0091	(0,2.42)
	3.0	1.0279	1.0270	1.0244	1.0200	1.0139	1.0061	(0,2.42)
	3.5	1.0213	1.0206	1.0186	1.0152	1.0106	1.0047	(0,2.42)
	4.0	1.0152	1.0147	1.0133	1.0109	1.0076	1.0033	(0,2.42)
	4.5	1.0119	1.0116	1.0104	1.0086	1.0060	1.0026	(0,2.42)
	5.0	1.0096	1.0093	1.0084	1.0069	1.0048	1.0021	(0,2.42)
8	2.5	1.0316	1.0306	1.0275	1.0226	1.0157	1.0069	(0,2.42)
	3.0	1.0213	1.0207	1.0186	1.0153	1.0106	1.0047	(0,2.42)
	3.5	1.0154	1.0149	1.0135	1.0111	1.0077	1.0034	(0,2.42)
	4.0	1.0117	1.0113	1.0102	1.0084	1.0058	1.0026	(0,2.42)
	4.5	1.0092	1.0089	1.0080	1.0066	1.0046	1.0020	(0,2.42)
	5.0	1.0074	1.0072	1.0065	1.0053	1.0037	1.0016	(0,2.42)

Table 4. The values of e_i for $i = 12\ and\ 13$ (cont.)

n	***β***	$e_{13} = RE\left(T_{13}^{*}, T_{03}\right)$					Range of λ in which T_{13}^{*} is efficient to T_{03}
		λ = 1.00	***λ = 1.25*** and ***λ = 0.75***	***λ = 1.50*** and ***λ = 0.50***	***λ = 1.75*** and ***λ = 0.25***	***λ = 2.00*** and ***λ = 0.00***	
2	2.5	1.2799	1.2578	1.1962	1.1058	1.0000	[0,2]
	3.0	1.1674	1.1553	1.1205	1.0669	1.0000	[0,2]
	3.5	1.1135	1.1056	1.0828	1.0467	1.0000	[0,2]
	4.0	1.0827	1.0772	1.0608	1.0346	1.0000	[0,2]
	4.5	1.0633	1.0591	1.0467	1.0267	1.0000	[0,2]
	5.0	1.0501	1.0469	1.0371	1.0213	1.0000	[0,2]
3	2.5	1.1180	1.1098	1.0859	1.0484	1.0000	[0,2]
	3.0	1.0752	1.0702	1.0553	1.0316	1.0000	[0,2]
	3.5	1.0526	1.0492	1.0389	1.0224	1.0000	[0,2]
	4.0	1.0391	1.0365	1.0290	1.0167	1.0000	[0,2]
	4.5	1.0303	1.0283	1.0225	1.0130	1.0000	[0,2]
	5.0	1.0242	1.0226	1.0180	1.0104	1.0000	[0,2]
4	2.5	1.0648	1.0605	1.0478	1.0274	1.0000	[0,2]
	3.0	1.0424	1.0397	1.0315	1.0181	1.0000	[0,2]
	3.5	1.0301	1.0282	1.0224	1.0130	1.0000	[0,2]
	4.0	1.0226	1.0211	1.0168	1.0098	1.0000	[0,2]
	4.5	1.0176	1.0165	1.0131	1.0076	1.0000	[0,2]
	5.0	1.0141	1.0132	1.0105	1.0061	1.0000	[0,2]
5	2.5	1.0409	1.0383	1.0304	1.0175	1.0000	[0,2]
	3.0	1.0272	1.0254	1.0203	1.0117	1.0000	[0,2]
	3.5	1.0195	1.0182	1.0145	1.0084	1.0000	[0,2]
	4.0	1.0147	1.0137	1.0110	1.0064	1.0000	[0,2]
	4.5	1.0115	1.0107	1.0086	1.0050	1.0000	[0,2]
	5.0	1.0092	1.0086	1.0069	1.0040	1.0000	[0,2]
6	2.5	1.0282	1.0264	1.0210	1.0121	1.0000	[0,2]
	3.0	1.0189	1.0177	1.0141	1.0082	1.0000	[0,2]
	3.5	1.0136	1.0127	1.0102	1.0059	1.0000	[0,2]
	4.0	1.0103	1.0096	1.0077	1.0045	1.0000	[0,2]
	4.5	1.0080	1.0075	1.0060	1.0035	1.0000	[0,2]
	5.0	1.0065	1.0061	1.0048	1.0028	1.0000	[0,2]
7	2.5	1.0206	1.0193	1.0153	1.0089	1.0000	[0,2]
	3.0	1.0139	1.0130	1.0104	1.0060	1.0000	[0,2]
	3.5	1.0106	1.0099	1.0079	1.0046	1.0000	[0,2]
	4.0	1.0076	1.0071	1.0057	1.0033	1.0000	[0,2]
	4.5	1.0060	1.0056	1.0045	1.0026	1.0000	[0,2]
	5.0	1.0048	1.0045	1.0036	1.0021	1.0000	[0,2]
8	2.5	1.0157	1.0147	1.0117	1.0068	1.0000	[0,2]
	3.0	1.0106	1.0099	1.0079	1.0046	1.0000	[0,2]
	3.5	1.0077	1.0072	1.0057	1.0033	1.0000	[0,2]
	4.0	1.0058	1.0055	1.0044	1.0025	1.0000	[0,2]
	4.5	1.0046	1.0043	1.0034	1.0020	1.0000	[0,2]
	5.0	1.0037	1.0035	1.0028	1.0016	1.0000	[0,2]

Table 5. The values of e_i for $i = 14 \; and \; 15$

n	β	$e_{14} = RE\left(T^{*}_{14}, t_4\right)$						Range of λ in which T^{*}_{14} is efficient to t_4
		$\lambda = 1.00$	$\lambda = 1.25$ and $\lambda = 0.75$	$\lambda = 1.50$ and $\lambda = 0.50$	$\lambda = 1.75$ and $\lambda = 0.25$	$\lambda = 2.00$	$\lambda = 2.25$	
2	2.5	1.6380	1.6099	1.5309	1.4152	1.2799	1.1397	(0,2.51)
	3.0	1.3628	1.3487	1.3080	1.2455	1.1674	1.0803	(0,2.47)
	3.5	1.2398	1.2311	1.2056	1.1654	1.1135	1.0531	(0,2.45)
	4.0	1.1723	1.1663	1.1485	1.1202	1.0827	1.0381	(0,2.44)
	4.5	1.1306	1.1262	1.1130	1.0917	1.0633	1.0288	(0,2.44)
	5.0	1.1028	1.0993	1.0891	1.0725	1.0501	1.0227	(0,2.43)
3	2.5	1.1733	1.1672	1.1494	1.1209	1.0832	1.0383	(0,2.44)
	3.0	1.1116	1.1078	1.0967	1.0786	1.0543	1.0246	(0,2.43)
	3.5	1.0785	1.0759	1.0682	1.0557	1.0385	1.0173	(0,2.43)
	4.0	1.0585	1.0566	1.0509	1.0416	1.0288	1.0129	(0,2.42)
	4.5	1.0454	1.0439	1.0396	1.0324	1.0224	1.0100	(0,2.42)
	5.0	1.0363	1.0351	1.0316	1.0259	1.0180	1.0080	(0,2.42)
4	2.5	1.0976	1.0944	1.0847	1.0690	1.0477	1.0215	(0,2.43)
	3.0	1.0644	1.0623	1.0561	1.0458	1.0317	1.0142	(0,2.43)
	3.5	1.0459	1.0445	1.0400	1.0327	1.0227	1.0101	(0,2.42)
	4.0	1.0345	1.0334	1.0301	1.0247	1.0171	1.0076	(0,2.42)
	4.5	1.0269	1.0261	1.0235	1.0193	1.0134	1.0059	(0,2.42)
	5.0	1.0216	1.0209	1.0189	1.0155	1.0107	1.0047	(0,2.42)
5	2.5	1.0563	1.0545	1.0490	1.0401	1.0278	1.0124	(0,2.42)
	3.0	1.0378	1.0366	1.0330	1.0270	1.0187	1.0083	(0,2.42)
	3.5	1.0272	1.0264	1.0238	1.0195	1.0135	1.0060	(0,2.42)
	4.0	1.0203	1.0197	1.0178	1.0146	1.0101	1.0045	(0,2.42)
	4.5	1.0161	1.0156	1.0141	1.0116	1.0080	1.0035	(0,2.42)
	5.0	1.0130	1.0126	1.0113	1.0093	1.0065	1.0028	(0,2.42)
6	2.5	1.0395	1.0382	1.0344	1.0282	1.0196	1.0087	(0,2.42)
	3.0	1.0267	1.0258	1.0233	1.0191	1.0133	1.0059	(0,2.42)
	3.5	1.0193	1.0187	1.0169	1.0138	1.0096	1.0042	(0,2.42)
	4.0	1.0146	1.0142	1.0128	1.0105	1.0073	1.0032	(0,2.42)
	4.5	1.0115	1.0111	1.0100	1.0082	1.0057	1.0025	(0,2.42)
	5.0	1.0092	1.0090	1.0081	1.0066	1.0046	1.0020	(0,2.42)
7	2.5	1.0278	1.0269	1.0243	1.0199	1.0138	1.0061	(0,2.42)
	3.0	1.0189	1.0183	1.0165	1.0135	1.0094	1.0041	(0,2.42)
	3.5	1.0137	1.0133	1.0120	1.0098	1.0068	1.0030	(0,2.42)
	4.0	1.0104	1.0101	1.0091	1.0075	1.0052	1.0023	(0,2.42)
	4.5	1.0082	1.0079	1.0071	1.0059	1.0041	1.0018	(0,2.42)
	5.0	1.0066	1.0064	1.0058	1.0047	1.0033	1.0014	(0,2.42)
8	2.5	1.0214	1.0207	1.0187	1.0153	1.0107	1.0047	(0,2.42)
	3.0	1.0146	1.0142	1.0128	1.0105	1.0073	1.0032	(0,2.42)
	3.5	1.0106	1.0103	1.0093	1.0076	1.0053	1.0023	(0,2.42)
	4.0	1.0081	1.0078	1.0071	1.0058	1.0040	1.0018	(0,2.42)
	4.5	1.0064	1.0062	1.0056	1.0046	1.0032	1.0014	(0,2.42)
	5.0	1.0051	1.0049	1.0044	1.0036	1.0025	1.0011	(0,2.42)

Table 5. The values of e_i for $i = 14 \text{ and } 15$ (cont.)

		$e_{15} = RE\left(T^*_{14}, T_{04}\right)$					Range of λ in which T^*_{14} is efficient to T_{04}
n	***β***	***λ* = 1.00**	***λ* = 1.25** and ***λ* = 0.75**	***λ* = 1.50** and ***λ* = 0.50**	***λ* = 1.75** and ***λ* = 0.25**	***λ* = 2.00** and ***λ* = 0.00**	
2	2.5	1.2799	1.2578	1.1962	1.1058	1.0000	[0,2]
	3.0	1.1674	1.1553	1.1205	1.0669	1.0000	[0,2]
	3.5	1.1135	1.1056	1.0828	1.0467	1.0000	[0,2]
	4.0	1.0827	1.0772	1.0608	1.0346	1.0000	[0,2]
	4.5	1.0633	1.0591	1.0467	1.0267	1.0000	[0,2]
	5.0	1.0501	1.0469	1.0371	1.0213	1.0000	[0,2]
3	2.5	1.0832	1.0776	1.0611	1.0348	1.0000	[0,2]
	3.0	1.0543	1.0508	1.0402	1.0231	1.0000	[0,2]
	3.5	1.0385	1.0360	1.0286	1.0165	1.0000	[0,2]
	4.0	1.0288	1.0270	1.0215	1.0124	1.0000	[0,2]
	4.5	1.0224	1.0210	1.0167	1.0097	1.0000	[0,2]
	5.0	1.0180	1.0168	1.0134	1.0078	1.0000	[0,2]
4	2.5	1.0477	1.0446	1.0353	1.0203	1.0000	[0,2]
	3.0	1.0317	1.0297	1.0236	1.0136	1.0000	[0,2]
	3.5	1.0227	1.0213	1.0169	1.0098	1.0000	[0,2]
	4.0	1.0171	1.0160	1.0128	1.0074	1.0000	[0,2]
	4.5	1.0134	1.0125	1.0100	1.0058	1.0000	[0,2]
	5.0	1.0107	1.0101	1.0080	1.0047	1.0000	[0,2]
5	2.5	1.0278	1.0260	1.0207	1.0120	1.0000	[0,2]
	3.0	1.0187	1.0175	1.0140	1.0081	1.0000	[0,2]
	3.5	1.0135	1.0127	1.0101	1.0059	1.0000	[0,2]
	4.0	1.0101	1.0095	1.0076	1.0044	1.0000	[0,2]
	4.5	1.0080	1.0075	1.0060	1.0035	1.0000	[0,2]
	5.0	1.0065	1.0061	1.0048	1.0028	1.0000	[0,2]
6	2.5	1.0196	1.0183	1.0146	1.0085	1.0000	[0,2]
	3.0	1.0133	1.0124	1.0099	1.0058	1.0000	[0,2]
	3.5	1.0096	1.0090	1.0072	1.0042	1.0000	[0,2]
	4.0	1.0073	1.0068	1.0055	1.0032	1.0000	[0,2]
	4.5	1.0057	1.0054	1.0043	1.0025	1.0000	[0,2]
	5.0	1.0046	1.0043	1.0035	1.0020	1.0000	[0,2]
7	2.5	1.0138	1.0129	1.0103	1.0060	1.0000	[0,2]
	3.0	1.0094	1.0088	1.0070	1.0041	1.0000	[0,2]
	3.5	1.0068	1.0064	1.0051	1.0030	1.0000	[0,2]
	4.0	1.0052	1.0049	1.0039	1.0023	1.0000	[0,2]
	4.5	1.0041	1.0038	1.0031	1.0018	1.0000	[0,2]
	5.0	1.0033	1.0031	1.0025	1.0014	1.0000	[0,2]
8	2.5	1.0107	1.0100	1.0080	1.0046	1.0000	[0,2]
	3.0	1.0073	1.0068	1.0055	1.0032	1.0000	[0,2]
	3.5	1.0053	1.0050	1.0040	1.0023	1.0000	[0,2]
	4.0	1.0040	1.0038	1.0030	1.0018	1.0000	[0,2]
	4.5	1.0032	1.0030	1.0024	1.0014	1.0000	[0,2]
	5.0	1.0025	1.0024	1.0019	1.0011	1.0000	[0,2]

5. Conclusion

It is observed from Table 1 that the values of the relative efficiencies $e_i's, i = 1,2,...,7$ of the proposed minimum mean squared error (MMSE) estimators $T_{0i}, i = 1,2,3,4$ with respect to Lesitha and Thomas (2012) estimators $t_i's, i = 1,2,3,4$ respectively are greater than 'unity'. Thus, the proposed estimators $T_{0i}, i = 1,2,3,4$ are more efficient than the corresponding usual estimators $t_i's, i = 1,2,3,4$ respectively.

It is further observed that the values of the relative efficiency e_5 of T_{04} with respect to T_{01} are the largest among $e_i's, i = 1,2,...,7$, from which follows that the proposed MMSE estimator T_{04} is the best estimator among Lesitha and Thomas (2012) estimators $t_i's, i = 1,2,3,4$ and MMSE estimators $T_{0i}, i = 1,2,3,4$.

Tables 2 to 5 demonstrate that for fixed (n, β) the values of relative efficiencies $e_i's, i = 8,9,...,15$ increase as λ increases up to 1, while it decreases if λ goes beyond 'unity'. When the value of λ is unity (i.e. the guessed value α_0 coincides with the true value α) a higher gain in efficiency is seen. For fixed values of (n, λ) the values of $e_i's, i = 8,9,...,15$ decrease as β increases. When (β, λ) are fixed the values of $e_i's, i = 8,9,...,15$ also decrease as sample size n increases. A higher gain in efficiency is obtained when the sample size n is small. In general, the estimators $T_{1i}^*, i = 1,2,3,4$ are more efficient than Lesitha and Thomas (2012) estimators $t_i's, i = 1,2,3,4$ and MMSE estimators $T_{0i}, i = 1,2,3,4$ respectively when $\lambda \in (0, 2.42)$. It is further observed that $T_{1i}^*, i = 1,2,3,4$ are respectively better than MMSE estimators $T_{0i}, i = 1,2,3,4$ when $\lambda \in [0,2]$.

Acknowledgement

The authors are highly grateful to the referees for their constructive comments/ suggestions that helped in the improvement of the revised version of the paper.

REFERENCES

AHMAD, M. I., SINCLAIR, C. D., WERRITTY, A., (1988). Log-logistic flood frequency analysis. J Hydrol, 98, pp. 205–224.

BALAKRISHNAN, N., MALIK, H. J., (1987). Best linear unbiased estimation of location and scale parameter of the log-logistic distribution. Commun Stat Theory Methods, 16, pp. 3477–3495.

BENNETT, S., (1983). Log-logistic regression models for survival data. J R Stat Soc, Ser C 32, pp. 165–171.

CHEN, Z., BAI, Z., SINHA, B. K., (2004). Ranked set sampling, theory and applications. Lecture Notes in Statistics, Springer, New York.

FISK, P. R., (1961). The graduation of income distributions. Econometrica, 29, pp. 171–185.

GESKUS, R. B., (2001). Methods for estimating the AIDS incubation time distribution when data of seroconversion is censored. Stat Med, 20, 795–812.

LESITHA, G., THOMAS, P. Y., (2012). Estimation of the scale parameter of A LOG-LOGISTICS DISTRIBUTION. METRIKA, DOI 10.1007/S00184-012-0397-5.

MCINTYRE, G. A., (1952). A method for unbiased selective sampling using ranked sets. Aust J Agric Res, 3, pp. 385–390.

MEHTA, V., (2015). Estimation in Morgenstern Type Bivariate Exponential Distribution with Known Coefficient of Variation by Ranked Set Sampling. Proceeding of the "30th M. P. Young Scientist Congress" (MPYSC-2015), M. P. Council of Science and Technology, Vigyan Bhawan, Nehru Nagar, Bhopal -462 003, Madhya Pradesh, India.

MEHTA, V., SINGH, H. P., (2014). Shrinkage Estimators of Parameters of Morgenstern Type Bivariate Logistic Distribution Using Ranked Set Sampling. Journal of Basic and Applied Engineering Research (JBAER), 1 (13), pp. 1–6.

MUTTLAK, H. A., (1997). Median ranked set sampling. J Appl Stat Sci. 6, pp. 245–255.

RAGAB, A., GREEN, J., (1984). On order statistics from the log-logistic distribution and their properties. Commun Stat Theory Methods, 13, pp. 2713–2724.

ROBSON, A., REED, D., (1999). Flood estimation handbook, 3. Statistical procedures for flood frequency estimation. Institute of Hydrology, Wallingford, UK.

SHOUKRI, M. M., MIAN, I. U. M., TRACY, D., (1988). Sampling properties of estimators of log-logistic distribution with application to Canadian precipitation data. Can J Stat, 16, pp. 223–236.

SINGH, H. P., MEHTA, V., (2013). An Improved Estimation of Parameters of Morgenstern Type Bivariate Logistic Distribution Using Ranked Set Sampling. STATISTICA, 73 (4), pp. 437–461.

SINGH, H. P., MEHTA, V., (2016a). Improved Estimation of Scale Parameter of Morgenstern Type Bivariate Uniform Distribution Using Ranked Set Sampling. Communications in Statistics – Theory and Methods, 45 (5), pp. 1466–1476.

SINGH, H. P., MEHTA, V., (2014a). Linear shrinkage estimator of scale parameter of Morgenstern type bivariate logistic distribution using ranked set sampling. Model Assisted Statistics and Applications (MASA), 9, pp. 295–307.

SINGH, H. P., MEHTA, V., (2014b). An Alternative Estimation of the Scale Parameter for Morgenstern Type Bivariate Log-Logistic Distribution Using Ranked Set Sampling. Journal of Reliability and Statistical Studies, 7 (1), pp. 19–29.

SINGH, H. P., MEHTA, V., (2015). Estimation of Scale Parameter of a Morgenstern Type Bivariate Uniform Distribution Using Censored Ranked Set Samples. Model Assisted Statistics and Applications (MASA), 10, pp. 139–153.

SINGH, H. P., MEHTA, V., (2016b). Some Classes of Shrinkage Estimators in the Morgenstern Type Bivariate Exponential Distribution Using Ranked Set Sampling. Hacettepe Journal of Mathematics and Statistics, 45 (2), pp. 575–591.

SINGH, H. P., MEHTA, V., (2016c). A Class of Shrinkage Estimators of Scale Parameter of Uniform Distribution Based on K- Record Values. National Academy Science Letters, 39, pp. 221–227.

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 75–90

THE APPLICATION OF MULTIVARIATE STATISTICAL ANALYSIS TO THE VALUATION OF DURABLE GOODS BRANDS

Marta Dziechciarz-Duda [1], Anna Król [2]

ABSTRACT

Nowadays, due to changes in the market and new trends in consumer behaviours, intangible assets, such as brand, have gained fundamental importance. The more frequent conviction that a product with a well-known name is better than other products contributes to the case of replacing the price of a product by its brand name as the predominant factor in the purchase decision process. Thus, for many companies the strengthening of brand equity has become one of the key elements of marketing strategy.

The main aim of this study is an attempt to improve the process of analysing the position and value of brands using selected multivariate statistical analysis methods (hedonic regression, multidimensional scaling, classification and linear ordination methods). In the conducted research the direct approach to the evaluation of the position of the brands for a selected ICT good – smartphones – have been applied. The measurement was performed on two levels: the product level, in which the prices of branded products were compared, and the consumer level, where the perception and attitudes of consumers towards the brands were studied. The analyses have been carried out on two sets of data, which enabled fuller and more comprehensible understanding of decision rules that guide consumers in choosing the brand.

Key words: brand valuation, multivariate statistical analysis, durable goods.

1. Brand and its value

In an era of fast technological development, when durable consumer goods are endowed in number of complex features, become more and more advanced and undergo rapid improvements, the consumers more frequently face the problem of a difficult choice between many variants of complicated commodities. This issue is particularly severe on the highly advanced durable goods markets (such as, for example, ICT goods), which on the one hand are violently expanding thus posing promising prospects for the future, and on the other hand are extremely changeable

[1] Wrocław University of Economics. E-mail: marta.dziechciarz@ue.wroc.pl
[2] Wrocław University of Economics. E-mail: anna.krol@ue.wroc.pl

and, furthermore, constantly experience shortening of the life cycles of products. In such unstable conditions one of the ways to win over the confused customers, induce their purchase decision and gain their trust and future loyalty is to create and strengthen the product brand. Increasing the brand value is particularly important on the markets where it is extremely difficult to fully assess the quality of the product before the purchase, and moreover the frequency of the purchases is low in comparison to the rapid pace of technological development, so that it is very hard for the consumers to build on their own experience of the past. Therefore, the well build brand might become a kind of safety buffer for the customers. It is known that brand names can convey information about various aspects of the product, such as reputation, reliability, quality, and also are synonyms of certain prestige and even social status and identity. Thus, they might provide the means for the customers to reduce risk level involved in the buying process and increase information efficiency of the purchase (see Keller and Lehmann (2006)).

The contemporary notion of brand name is understood as a much broader concept than originally comprehended. According to the classic definition formulated by the American Marketing Association (1960) brand name is „(...) a name, term, symbol, or design, or logo or a combination of them used to identify and differentiate a product or service from the competitor in the marketplace". Thus, the main role of brand name in this early approach was to inform the potential customers about the product existence and discriminate it from other similar commodities. Nowadays the brand name notion encompasses much more. The brand represents broadly understood trust, as well as certain connection between the company and the consumer, and is extended to include representing certain quality level, introducing the image of prestige and social status, as well as building customers identity, among other things. Detailed discussion on modern brand definitions may be found, among others, in papers by de Chernatony and Dall'Olmo Riley (1998) and Maurya and Mishra (2012). Maurya and Mishra list 12 different aspects in which the brand have been regarded in the literature: brand as a logo, brand as a legal instrument, brand as a company, brand as a shorthand, brand as a risk reducer, brand as an identity system, brand as an image in consumer's mind, brand as a value system, brand as a personality, brand as relationship, brand as adding value and brand as an evolving entity. This multitude of various approaches and issues accounted for in brand name analyses shows the complexity of the brand name evaluation problem.

The presented approaches to brand definition can be related to the question of the brand equity assessment by assigning them to two major trends present in the literature: cognitive psychology and informational economics (cf. Aaker (1991), Erdem and Swait (1998), Erdem et al. (1999), Baltas and Saridakis (2010)).

The authors of the cognitive psychology concept define brand equity as the effect that the knowledge of the brand and its features has on the response and revealed preferences of the consumer (Aaker (1991); Keller (1993)). Crucial elements in the process of brand valuation include brand awareness (the degree to which consumer precisely recognizes the brand and associates it with the specific product), brand associations (the extent to which a particular brand calls to mind the attributes of a general product category), brand loyalty (the extent of the faithfulness of consumer to a particular brand), brand perceived quality (consumer's opinion of a brand's ability to fulfill expectations) and other proprietary brand assets (patents, trademarks, and channel relationships that prevent the competitors from eroding a customer base). The stronger and more positive those factors, the higher brand price premium. Informational economics approach somewhat differently presents the issue of brand equity (Erdem and Swait (1998)). This concept emphasizes the incompleteness of signals coming from the market, which forces the consumer to make decisions in a situation of partial misinformation. In this context, the brand is understood as a source of information on the quality of goods, as well as a way to reduce purchase risks and costs of searching for the right product on the market. Brand price premium reflect the return on the investment in creating and strengthening the brand made by the company.

The value of the brand name and its strength affect the mechanisms that occur during the process of purchasing goods by the consumer. Cobb-Walgren et al. (1995) showed a significant positive correlation between brand equity and consumer's brand preference. Moreover, an increase in brand value translates into higher purchase intentions and significantly influences the final purchase decision made by the customer. The more frequent conviction that a product with a well-known name is better than other products contributes to the case of replacing the price of a product by its brand name as the predominant factor in the purchase decision process. Thus, for many companies the strengthening of brand equity has become one of the key elements of marketing strategy. The relationship and interactions between the brand and the actual purchase is schematically shown in Figure 1.

Brand is especially important on durable goods market, because of the specificity of the market and durable goods features. Durables typically cost substantially more than nondurable products and thus entail greater financial risk for consumers. Even though consumers are in the market for a short period and after purchase stay away for a long time, they undoubtedly spend a substantial amount of money in that period. The individual consumer is present on the market intermittently so it is difficult to evaluate the quality of the product before the purchase. Thus, customers

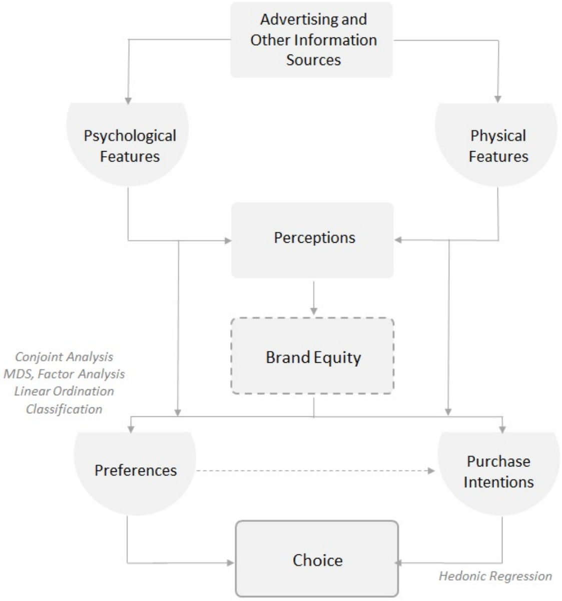


Figure 1: Brand equity in the purchase process (Source: based on Cobb-Walgren et al. (1995))

often rely on the brand name reputation in order to reduce the perceived risk and search costs.

Since in many cases brand is one of the most valuable intangible assets of the company, there is a necessity of developing accurate measures of brand equity. Brand equity measurement approaches include three main areas - one financial, and two more closely related to the marketing concepts (Keller and Lehmann (2006)):

- financial-based approach, which focuses on the monetary or financial value of the brand in the marketplace;
- consumer-based brand equity, which involves the value added to a product or service by consumers' associations and perceptions of a brand name;
- product-market-based approach.

It is worth noticing that brand in financial terms, as an asset of the company, is dependent on such factors as brand loyalty, brand awareness, perception of brand quality, brand associations (cf. Aaker (1991), which in fact are closely linked to the area of marketing in the company. Thus, financial and marketing approaches to brand valuation are closely related and the measurement of brand value is a complex multi-level issue. This paper focuses mainly on methods of brand valuation and brand equity measurement from marketing perspectives, and thus consumer-based and product-market-based measurement approaches are analysed. Measuring brand as a financial resource - analyses of the company's assets and all related aspects in the accounting areas of the company's operations are not the subject of this study. Further aspects of the presented research concerning consumer preferences may be found in a paper by Dziechciarz-Duda and Król (2016).

2. Marketing approaches to brand equity measurement

From the marketing point of view both the product-market-based and consumer-based approaches are applied to brand valuation. In the product-market-based approach a strong brand is one that increases the effectiveness of advertising, differentiates the product from the competition and facilitates prospective expansion into new market segments. In this context, the brand value can be understood as an additional bonus due to the ownership of a particular brand, which is absent in the case of an equivalent, comparable product that does not have a strong brand. A common tool for brand valuation from the product-market perspective is the analysis of the price premium as an indicator of brand equity (Erdem et al. (1999); Netemeyer et al. (2004)). The price premium is defined as the amount of money the consumer is willing to pay for the product of preferred brand in comparison with other products having similar characteristics but different brand names (Kamakura and Russell (1993)). The price premium is considered as a valuable and comprehensive tool for brand equity measurement, which is in addition relatively easy to calculate and straightforward in interpretation. However, there are some critical views in the literature (Ailawadi et al. (2002)) indicating that the models used to calculate the price premium (e.g. hedonic models) do not directly capture the essential elements of the marketing mix (e.g. advertising). Nevertheless, it is assumed that the forces of supply and demand, as well as other market mechanisms take into account all these aspects indirectly. From that point of view hedonic modeling is a useful statistical tool in the analysis of brand value at the product level.

Numerous studies devoted to issues of brand valuation focus on the consumer-based brand equity. This approach to brand value analysis is considered especially valuable because it directly examines the consumer behavior, which provides the basis for formulating marketing strategies (Keller (1993)). According to this approach the subject of analysis is the type of associations with the brand that comes to the mind of the customer. Moreover, it is assumed that the unique, positive and strong associations influence purchasing decisions of potential customers. Brand value at the consumer level is connected with the knowledge of the brand and types of reactions toward the brand, which can be expressed in five areas (Keller and Lehmann (2006)): awareness (ranging from recognition to recall), associations (encompassing tangible and intangible product considerations), attitude (ranging from acceptability to attraction), attachment (ranging from loyalty to addiction) and activity (including purchase and consumption frequency).

As mentioned before, the purchase process is a complex and multi-faceted operation (Figure 1), and the preferences of consumers (including the preference towards the brand), purchase intentions and the actual purchasing decision form composite relations. This raises the question whether declarative behaviours of the consumers always translate directly into actual purchasing decisions. As numerous research on the process of buying suggests, the declared preference towards the brand is highly correlated with customers' choices, but there are markets where this relationship is weaker. Undoubtedly, consumer durables market falls into that category. On durable goods markets, only few percent of the declared intention of buying is actually implemented (cf. Dziechciarz (2008), Cobb-Walgren et al. (1995), Morwitz and Schmittlein (1992)), which pose a particular challenge for analyses of consumer future reactions to the elements of marketing mix.

Given the above concerns, it seems reasonable to confront the results of studies assessing the brand at the consumer level with the objective study of the current market offer (product level approach). The combined application of these approaches to brand valuation and juxtaposition of the results of the brand analyses from the consumer and the market perspectives provides added value in the form of additional information and areas open to interpretation. On the basis of the consumer preferences research one can obtain information about subjective associations with the brand. Such a study is extremely valuable because it provides insights into individual attitudes of potential customers. The obvious disadvantage of this approach lies in its subjectivity. In addition, there are many signs saying that declarative behaviours do not always translate into actual purchasing decisions. Therefore, the authors believe that there is a need to confront the subjective brand valuations provided by the consumers with the information from the market (i.e. the calculated price premium based on existing market offer).

In the following part of the paper an empirical example illustrating the possible applications of multivariate statistical methods for the brand equity measurement for selected durable good - smartphone - will be presented. The estimated parameters from an econometric model based on the data showing the market offer will be compared with assessments of brands from a study of consumer preferences towards brands. Thus, on the product-market level of brand valuation, hedonic regression model will be used. In turn, on the consumer-base level multidimensional scaling, linear ordering and hierarchical classification methods will be applied.

3. Data sets

The presented analyses have been carried out on two separate sets of data. On the product level the prices and significant characteristics of smartphones (including brands) were collected from price lists available on an Internet website of one of the biggest in Poland price comparison service provider, whereas on the consumer level the perception and attitudes of consumers towards the smartphone brands were measured using a specifically designed on-line survey. This two-sided approach enabled fuller and more comprehensible understanding of decision rules that guide consumers in choosing the brand of a smartphone.

3.1. Data set for hedonic analysis

Database used in this part of the study have been created using a tool for data collection from web pages created by the authors. The data originate from Polish price comparison service providers. The data set comprises 910 smartphones of 27 different brands offered in Internet shops in Poland in February 2015. Each offer is described by price (PRICE [PLN]) and the following smartphones' characteristics: SCREEN – screen size [inch], STORAGE – internal storage [GB] and CAMERA – camera resolution [Mpix]. Moreover, the following dummy variables (take value 1 if the feature is present and 0 otherwise) were used: LTE, GPS, ANDROID, as well as dummies representing the brands.

3.2. On–line survey data set

The data from on-line survey was gathered in February 2015 among the students of Wrocław University of Economics. The questionnaire was focused on measuring consumers' preferences towards smartphone characteristics and possible applications of the device. The sample consisted of 451 respondents selected based on their accessibility and proximity (convenience sampling).

The respondents were expected to asses popular brands of a smartphone, its important characteristics, as well as the common usage patterns of the device. Thus, in order to evaluate the analyzed criteria, each respondent have created his individual rankings of the brand names, the criteria, that would be taken into account while purchasing a smartphone and the common usage patterns of a smartphone. Moreover, the respondents have assessed the brand names of smartphones by assigning the rate (on 5-point scale: 5 - highest rate, 1 - lowest rate) to five brand attributes: prestige, design, modernity, support and reliability.

4. Product level approach

4.1. Hedonic modelling

The foundations of hedonic methods are formed by the so-called hedonic hypothesis, which states that heterogeneous commodities are characterized by a set of relatively homogeneous attributes (characteristics) relevant both from the point of view of the customer and the producer (Brachinger (2002); Dziechciarz (2004)). The relationship between the price of commodity ($PRICE$) and the set of its characteristics (X) described by certain function f is called hedonic regression and may be described in the following general notation:

$$PRICE = f\left(X;\beta;\varepsilon\right), \qquad (1)$$

where ε is the error term of the model. The estimate of the vector of parameters, obtained by estimation of the correctly specified hedonic regression model using data set, allows to calculate the prices of individual characteristics of the given good (so-called hedonic prices or implicit prices). It is assumed that the consumers derive utility from goods attributes, and therefore the hedonic prices reflect the willingness to pay for certain levels of attributes. In that context the hedonic model may be used to measure the brand price premium.

4.2. Estimation results

The results of estimation of the hedonic model for smartphone prices are presented in Table 1. The best functional form turned out to be the model with dependent variable transformed to logarithm (lnPRICE), and some of the independent variables in logarithmic transformation (SCREEN, CAMERA), and quadratic transformation (STORAGE). Due to heteroskedasticity of the error term weighted least squares method proposed by White (1980) was applied for model estimation.

Out of 27 brand names present in the dataset 18 were statistically significant (in the parentheses number of models representing given brand is given): ACER (11), ALCATEL (24), ALIGATOR (6), APPLE (40), ARCHOS (7), ASUS (6), BLACKBERRY (16), GIGABYTE (22), HTC (62), HUAWEI (37), LG (91), MOTOROLA (17), NOKIA (94), PRESTIGIO (41), SAMSUNG (217), SONY (87), ZOPO (15), ZTE (7). The remaining 9 brands formed the reference group: BE (9), GOCLEVER (25), KRUGER&MATZ (15), MANTA (6), MEDIA-TECH (10), MYPHONE (14), OVERMAX (16), TELEFUNKEN (6), WIKO (9). Almost all variables in the model are highly statistically significant (on the significance level lower than 0.01). Variables ARCHOS and PRESTIGIO are significant on the level 0.05. The signs of obtained parameters estimates are in accordance with expectations. The goodness-of-fit of the model measured by adjusted R^2 statistic is on the satisfactory level 90.85%.

Table 1. Hedonic model estimation results

	Coefficient	p-value		Coefficient	p-value
constant	3.789830	0.0000	ASSUS	0.528194	0.0027
lnSCREEN	1.071200	0.0000	BLACKBERRY	0.806932	0.0000
lnCAMERA	0.340625	0.0000	GIGABYTE	0.347101	0.0000
STORAGE	0.031558	0.0000	HTC	0.707189	0.0000
STORAGE^2	−0.000348	0.0000	HUAWEI	0.383858	0.0000
ANDROID	−0.129963	0.0020	LG	0.354884	0.0000
GPS	0.120036	0.0015	MOTOROLA	0.422555	0.0000
LTE	0.149875	0.0000	NOKIA	0.288791	0.0000
ACER	0.407796	0.0000	PRESTIGIO	0.087313	0.0194
ALCATEL	0.281955	0.0000	SAMSUNG	0.512381	0.0000
ALIGATOR	0.419322	0.0017	SONY	0.500371	0.0000
APPLE	1.145070	0.0000	ZOPO	0.170541	0.0024
ARCHOS	0.136089	0.0408	ZTE	0.431813	0.0000

The estimated parameters for various brands in the hedonic model can be interpreted as brand premiums - the surplus amounts the consumers are willing to pay just because the smartphone is of a certain brand. Table 2 presents the brand premiums for the brand names which were assessed by the respondents in the second part of the study. For example, the most valued brand is Apple. The smartphones from this producer are on average about 215% more expensive in comparison to the smartphones with brands from reference group, ceteris paribus.

Table 2. Smartphone brand premiums

Brand name	Brand premium	Brand name	Brand premium
APPLE	214.27%	MOTOROLA	52.59%
BLACKBERRY	124.10%	HUAWEI	46.79%
HTC	102.83%	LG	42.60%
SAMSUNG	69.59%	NOKIA	33.48%
SONY	64.93%	GOCLEVER	-

5. Consumer level approach

For the analysis of the position of the brand at the consumer level three methods of multivariate statistical analysis have been applied. The first method, unfolding

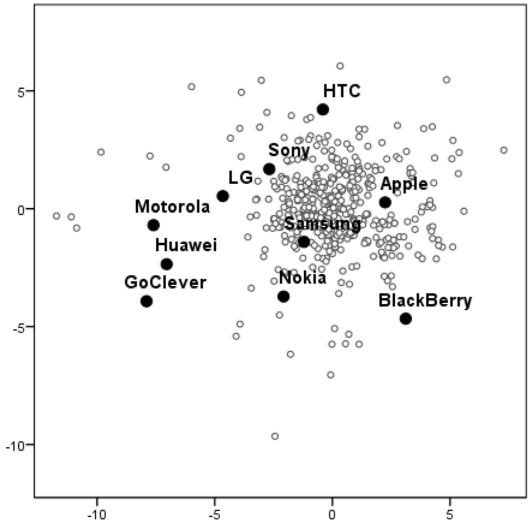


Figure 2: Preference map for smartphone brand names

analysis, belongs to a large group of multidimensional scaling methods, all of which result in the preference map creation. Multidimensional scaling is widely used in marketing research because it allows for intuitive interpretation of the results, including for example the evaluation of preference for brands (cf. Walesiak and Gatnar (2004)). The analysis was supplemented by the classification in which relatively homogeneous groups of brands were created. The similarity criterion in the clusters was the rating of selected brands of smartphones provided by the respondents. Both the perception map and the dendrogram obtained by the classification allows one to determine the groups of competing brands and provide guidance as to which brands are perceived by consumers as substitutes and which are considered exceptional. Additional analysis of consumers' preference towards brands included brands attributes such as: reliability, modernity, design, support, prestige and general brand image. On the basis of the respondents' evaluation of those criteria of brand quality, linear ordination was used, resulting in arrangement of brands from the most to the least preferred in view of the respondents.

5.1. Multidimensional scaling

In the presented empirical example, the PREFSCAL procedure of unfolding analysis was applied. The method allows for the presentation of objects and respondents in a joint two-dimensional space, which provides information on their co-existence.

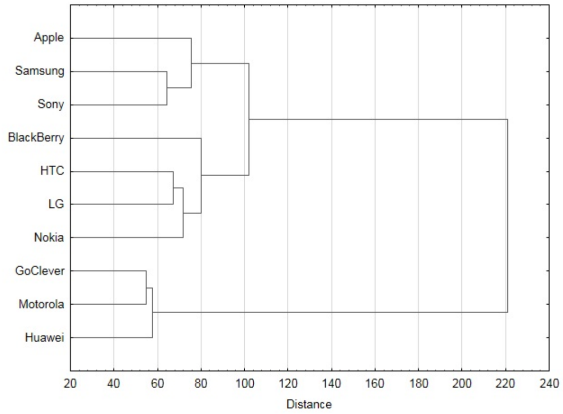


Figure 3: Dendrogram for smartphone brand names

The map of preferences (Figure 2) shows three groups of brands that are perceived by consumers in a similar manner. This observed result will be verified in the following example using hierarchical classification. By far the most preferred are the two leading brands in the market, Samsung and Apple, and slightly lower ranked is the Sony brand. Clearly, the Motorola, Huawei and GoClever brands are the least preferred by the respondents. In addition, it can be said that these brands (Motorola, Huawei and GoClever) are seen as substitutes, due to the small distance between these brands on the perception map. All other brands (Nokia, HTC, LG and BlackBerry) are valued by respondents with specific expectations and preferences.

5.2. Classification of brands

In the classification procedure the evaluations of respondents' preferences towards brands were again used. The clusters were created using Ward method, which assures high homogeneity of obtained groups (cf. Walesiak and Gatnar (2004)). As a result of the procedure, three classes of brands were created, which on the one hand led to the creation of brands groups perceived in a similar manner, and on the other hand provided the initial ranking of brand preference.

The analysis of dendrogram confirmed previous findings regarding smartphone brands preferences (Figure 3). As before, the most valued brands are Apple, Samsung and Sony. Similarly, the least valued brands (Motorola, Huawei and GoClever) are arranged in a clearly separate group. In addition, obtained classification gives the possibility to create a preliminary ranking of brand preferences according to the characteristics of established groups. Top rated are Apple, Samsung, and Sony, fol-

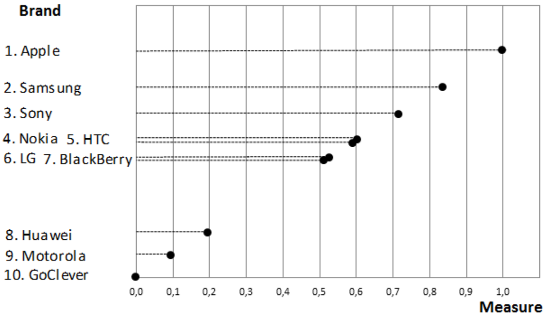


Figure 4: Measures of development for smartphone brands

lowed by another group, namely the brands BlackBerry, HTC, LG and Nokia. The brands GoClever, Motorola and Huawei are at the end of the ranking. This initial ranking is later verified using the classical method used to order objects according to multiple criteria at the same time - linear ordination.

5.3. Linear ordination

In the final step of customer-based brands analysis the ranking of smartphones' brands from the most to least preferred was created. The applied tool was one of the commonly used method of linear ordination - the index of the development method first introduced by Hellwig (1968). The method has been repeatedly discussed in the literature, detailed procedure is described in Dziechciarz et al. (1986). The principal idea is based on the concept that the best possible brand should have the shortest taxonomic distance from the pattern of development, and the longest distance from the anti-pattern of development. The pattern of development represents the abstract object for which all characteristic have the most desirable values, in the respondents' opinions. The anti-pattern of development is the reverse - the abstract brand with the worst possible attributes. The multi-attribute ranking was created using six criteria describing the brands (reliability, modernity, design, support, prestige and general brand image), all of which were stimulants. In the study weights were not applied because of the assumption of equal participation of each variable in the creation of the synthetic measure. The higher the values of development measure, the better the place of brand preferences in the ranking. Thus, the procedure of linear ordination method allowed to order smartphones' brands from the most to least preferred taking into account all mentioned criteria.

The illustration of the obtained result is presented in Figure 4. The undisputed leader in the ranking of the most preferred brands is Apple. The following positions are occupied by Samsung and Sony, respectively. On subsequent places Nokia and HTC, as well as LG and BlackBerry brands are located. Among the least appreciated brands are Huawei, Motorola and GoClever.

6. Summary and conclusions

The juxtaposition of the results from two approaches - product level and customer level - provides additional information on consumer preferences for brands. The main idea is to compare and interpret the results obtained from the estimated hedonic model and summarized in Table 2 with the results from multivariate methods presented in Figures 2, 3 and 4. In the case of some brands (e.g. Apple, Samsung, Sony, GoClever) both approaches give similar results suggesting compatibility of respondents' preferences with market brand valuation from hedonic analysis. It is worth noticing that the strongest convergence of consumer preferences occurs for the most preferred brand (Apple) and the least respected brand - GoClever. In the case of other brands (e.g. BlackBerry, Motorola) the respondents tend to appreciate the brands less regardless of their higher valuation on the product level. Finally, some brands (e.g. Nokia) are highly preferred despite significantly lower influence of the brand on the product price.

The presented empirical example confirmed that the selected multivariate statistical methods used to analyze consumer preferences are especially valuable because of the possible applications in the field of brand valuation and brand equity measurement. Both the multidimensional scaling and hierarchical classification allowed us to determine the groups of brands most and least preferred. In addition, it was possible to determine the position of each brand against the competition, and an indication of complementary brands. In order to confirm the pre-developed linear ranking linear ordination was used, which ultimately helped to identify the most and least preferred brands. It should be emphasized that the results of analyzes carried out by various methods of multivariate statistical analysis presented in the section on consumer lever approach were very consistent. However, as previously mentioned, the results are burdened with the subjectivity of respondents, and also with the possibility that not all statements will be reflected in future activities of the consumers. For this reason it is advisable to supplement the analysis of consumers brand evaluation with objective analysis of the price premiums. The confrontation of the results of consumer-based research with product-market based hedonic modelling allowed us to identify similar as well as divergent areas in the conducted empirical example. Such approach allows one, for example, to indicate those brands which are overvalued by the market and those that could be priced higher. Moreover, the analysis facilitates the distinction of segments within the market, and identification of brands that might be appreciated by the consumer with specific preferences.

Acknowledgements

The study was partly conducted in the framework of the research project entitled *The Application of Hedonic Methods in Quality-Adjusted Price Indices* (*Zastosowanie metod hedonicznych do uwzględniania różnic jakości dóbr we wskaźnikach dynamiki cen*). The project has been financed by the National Science Centre on the basis of decision no. DEC-2013/09/N/HS4/03645.

REFERENCES

AAKER, D. A., (1991). *Managing Brand Equity*, New York: The Free Press.

AILAWADI, K. L., LEHMANN, D. R., NESLIN, S., (2002). A Product-Market-Based Measure of Brand Equity. *Marketing Science Institute Working Paper*, 02-102.

AMERICAN MARKETING ASSOCIATION, (1960). *Marketing Definitions: A Glossary of Marketing Terms*, Chicago: American Marketing Association.

BALTAS, G., SARIDAKIS, C., (2010). Measuring Brand Equity in the Car Market: a Hedonic Price Analysis. *Journal of the Operational Research Society*, 61(2), 284–293.

BRACHIGNER, H. W., (2002). Statistical Theory of Hedonic Price Indices. *DQE Working Papers*, 1, Switzerland: Department of Quantitative Economics, University of Freiburg/Fribourg.

COBB-WALGREN, C. J., RUBLE C., DONTHU, N., (1995). Brand Equity, Brand Preference, and Purchase Intent. *Journal of Advertising*, 24(3), 25–40

DE CHERNATONY, L., DALL'OLMO RILEY, F., (1998). Defining a „Brand": Beyond the Literature with Experts' Interpretations. *Journal of Marketing Management*, 14(5), 417–443.

DZIECHCIARZ, J., et al., (1986). *Ekonometria z elementami programowania matematycznego i analizy porównawczej*, Bartosiewicz S. (ed.); Wrocław: Wyd. AE.

DZIECHCIARZ, J., (2004). Regresja hedoniczna. Próba wskazania obszarów stosowalności. In *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych.* A. Zeliaś (ed.), Kraków: Wyd. AE, 163–175.

DZIECHCIARZ, M., (2008). Deklaracje intencji zakupu w analizie wyposażenia gospodarstw domowych w innowacyjne dobra trwałego użytku. In *Zastosowania badań marketingowych w procesie tworzenia nowych produktów (cena, opakowanie, znak towarowy)*, S. Kaczmarczyk, M. Schulz (eds.), Toruń: TNOIK, 157–166.

DZIECHCIARZ-DUDA, M., KRÓL, A., (2016). The Analysis of Consumers' Preferences with the Application of Multivariate Models: Hedonic Regression and Multidimensional Scaling. *Archives of Data Science*, under review.

ERDEM, T., SWAIT, J., (1998). Brand Equity as a Signalling Phenomenon. *Journal of Consumer Psychology*, 7(2), 131–157.

ERDEM, T., SWAIT, J., BRONIARCZYK, S., CHAKRAVARTI, D., KAPFERER, J. N., KEANE, M., ROBERTS, J., STEENKAMP, J-B. E. M., ZETTELMEYER, F., (1999). Brand Equity, Consumer Learning and Choice. *Marketing Letters*, 10(3), 301–318.

HELLWIG, Z., (1968), Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr, *Przegląd Statystyczny*, 4, 307–326.

KAMAKURA, W. A., Russell, G. J., (1993). Measuring Brand Value with Scanner Data. *International Journal of Research in Marketing*, 10(1), 9–22.

KELLER, K. L., (1993). Conceptualizing, Measuring, and Managing Customer-based Brand Equity. *Journal of Marketing*, 57(1); 1–22.

KELLER, K. L., LEHMANN, D. R., (2006). Brands and Branding: Research Findings and Future Priorities. *Marketing Science*, 25(6), 740–759.

MAURYA, U. K., MISHRA, P., (2012). What Is a Brand? A Perspective on Brand Meaning. *European Journal of Business and Management*, 4(3), 122–134.

MORWITZ, V., SCHMITTLEIN, D., (1992). Using segmentation to improve sales forecasts based on purchase intent: which „intenders" actually buy? *Journal of Marketing Research*, 29(4), 391–405.

NETEMEYER, R. G., KRISHNAN, B., PULLING, C., WANG, G., YAGCI, M., DEAN, D., RICKS, J., WIRTH, F., (2004). Developing and Validating Measures of Facets of Customer-based Brand Equity. *Journal of Business Research*, 57 (2), 209–224.

WALESIAK, M., GATNAR, E. (2004). *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wrocław: Wyd. AE.

WHITE, H., (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4), 817–838.

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 91–114

TRADE PATTERN ON WARSAW STOCK EXCHANGE AND PREDICTION OF NUMBER OF TRADES

Henryk Gurgul[1], **Artur Machno**[2]

ABSTRACT

The main goal of this paper is to present the method for describing and predicting trade intensity on the Warsaw Stock Exchange. The approach is based on generalized linear models, the variable selection is performed using shrinkage methods such as the Lasso or Ridge regression. The variable under investigation is the number of trades of a particular stock 5-minute interval.

The main conclusion is that the number of trades during short intervals is predictable in the sense that the prediction, even based on relatively simple models, is with respect to statistical properties better than the prediction based on the random walk, which is used as a benchmark model.

Key words: high frequency data, daily trade pattern, Warsaw Stock Exchange, market microstructure.

JEL classification: C53, G17.

1. Introduction

Trade intensity in a high frequency setting is an interesting and important topic. Standard models for time series are invalid in a high frequency world for many reasons. For example, returns on stocks, in classic time series models, assume a continuous distribution. This assumption is clearly not met, because of the tick size, possible changes in price and multiple tick sizes. In standard time series analysis, e.g. daily or weekly data, this simplification is acceptable. However, for example for 1-minute returns, the probability of a return being equal to zero is very high, close to one, which makes this assumption materially incorrect. Another issue is the exact time of a transaction, for 1-minute data it

[1] Department of Applications of Mathematics in Economics, Faculty of Management, AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Krakow, Poland. E-mail: henryk.gurgul@gmail.com.

[2] Department of Applications of Mathematics in Economics, Faculty of Management, AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Krakow, Poland. E-mail: artur.machno@gmail.com.

might be significant if times of trades are reported with a margin of error of one or two seconds.

One approach to investigating the trade intensity is to analyze the time between trades. The most popular model based on this approach is the Autoregressive Conditional Duration (ACD) model. The ACD model is conceptually neat. Generally, the basic ACD model assumes a conditionally exponential distribution of the time to the next transaction. The assumption is not met for the trade of stocks. The involvement of more complex distributions makes an ACD estimation computationally heavy. Additionally, the U-shape of the daily trade pattern makes the estimation much more difficult. It is relatively easy to overcome this problem using a proper data transformation, although the results for transformed data are more difficult to interpret.

Those concerns have motivated us to analyze trade intensity in a different way. The proposed method is conceptually relatively easy, although most of the inference possible for ACD is achievable. The idea is to regress the number of trades of a particular stock during a 5-minute interval, the choice of the length of interval is arbitrary and chosen as an example. From the data analyst’s point of view, this is a prediction problem with the independent variable being a count variable. Dependent variables in this situation can be chosen freely. In this paper we use the characteristics from the previous interval as regressors. They are the number of trades, the number of trades with a higher price, the number of trades with a lower price and the value of all transactions.

Additionally, the trade pattern is analyzed. This is not the core of this research, thus it is not discussed fully. The trade pattern is visible and not considering it in the prediction would be incorrect. We use statistical methods to cluster the intervals in such a way that in every cluster the trade intensity is statistically similar.

Prediction models are based on the linear and generalized linear model (GLM). The authors are aware that the results can be improved in terms of prediction power by using more sophisticated machine-learning techniques. However, the goal of this analysis is to show that the prediction is manageable and the authors try to present the relation between the number of trades and other factors. This inference would be difficult for machine-learning techniques, which are usually “black boxes”.

2. Literature overview

The progress of computational techniques and trading methods has recently had a massive impact on the most important topics in financial research reflected in the financial literature. One of the most important directions of both theoretical and empirical research is the market microstructure. On the basis of intraday data numerous researchers try to describe the price, volatility and trading volume and the process of their formation. One of the first streams of research is represented in a paper by De Jong and Rindi (2009). Similarly to other researchers, the authors try to analyze the market structure and create market designs. The authors aim to find an effect of these factors on the intraday price formation. In the last decade researchers obtained access to tick-by-tick data and thus other intraday high frequency data. The availability of high frequency data has made it possible to prove the theories of market microstructure empirically.

One of the first topics analyzed empirically in the framework of market microstructure was intraday price dynamics. From a theoretical (statistical) point of view, the calendar-time distribution of stock price dynamics on small scales of time depends on both the distribution of price changes and the distribution of duration. These empirical studies are strongly interrelated with the observation of trading volume given by tick-by-tick data. The intraday behavior of stock prices was tested by Engle (2000). He noticed that the largest rise in the volume of transactions can be observed at the open and at the close of the market. On the basis of existing theories and this observation, Engle explained the U-shaped pattern of volatility over the course of the day. The research based on the intraday data supplied evidence that the size of the time intervals is also essential at long-time scales, which contradicts claims made using traditional stock price models. Numerous papers on market microstructure including Diamond and Verrechia (1987), Easley and O'Hara (1992), Engle and Russell (1998), Engle (2000), Dufour and Engle (2000), Manganelli (2005) and Cartea and Meyer-Brandis (2010) have confirmed that the duration at high frequencies between trades supplies constitutes important information about the intraday dynamics of tick-by-tick trades. This information reflects the behavior of the market, the differences in the market activity of uninformed or informed market participants, the volatility of price changes and the implied volatility from the option markets.

Numerous empirical papers confirm that duration, as a random variable, can be considered as one of the most significant factors determining stock price behavior, especially with respect to short periods of time. As we just mentioned above, in the past this random variable was not usually taken into account in most asset pricing models with horizons of more than one day. Those who research stock market behavior on the basis of daily (or more aggregated) data assumed that any effect of durations is dissipated immediately. Nowadays, many trades are made within algorithmic trading processing on a tick-by-tick level. Thus, the duration between trades is included in the most important explanatory variables in most recent research. It delivers important information about the stock market behavior over short-time intervals.

The main goal of most trading strategies is to benefit from the price patterns and behavior of prices, volatility and trading volume over ever shrinking scales of time. Empirical studies suggest that the amount of time required to complete a trade has decreased in the last couple of years by a digit. Nowadays, most trades are conducted very quickly over short periods of time. This extremely fast trading has become one of the most popular kinds of trading (especially algorithmic trading) among market participants. The important aim of both theoretical and empirical research was the identification of the main factors which enhance the fast expansion of algorithmic trading. The impulse for speeding up the trades comes from essential changes in the market structure, observed especially in the last decade. In addition, in recent years the capacities of computers have significantly increased. However, the cost of even powerful computers has significantly decreased. These factors have increased not only the number of market participants but also raised the speed of trades on all stock markets.

Most researchers who write about duration just one paper of Engle and Russell (1998). Their autoregressive conditional duration (ACD) model maps the time of arrival of financial data. This model is a starting point for numerous authors who aim to generalize the ACD model in different ways. Probably the best known and most frequently applied examples are the logarithmic model of Bauwens and Giot (2000) and extended class of models by Fernandes and Grammig (2005). Some other generalizations refer to regime-switching and mixture ACD models. They are referenced in Maheu and McCurdy (2000), Zhang et al. (2001), Meitz and Terasvirta (2006), Hujer et al. (2002). The structural model for durations between events and associated marks is presented in a paper by Renault et al. (2012). An extensive review of different duration models which reflect duration can be found in Bauwens and Hautsch (2009).

Considering the literature as a whole, high frequency trade pattern is one of the most important topics from both the theoretical and empirical points of view. The trade pattern is also a very interesting topic, especially in the case of the stock markets in countries in transition like the Warsaw Stock Exchange (WSE). The Polish economy (and stock market) is the largest among the economies in transition from CEE. However, in relation to bigger economies such as the German economy it is small. According to recent empirical studies, all stock markets react to news from the US stock market in terms of price, volatility and trading volume performance. Trade on NYSE starts at 3 p.m. CET (Polish time), therefore in addition to the usual U-shape of the trade pattern, one can observe a fluctuation at 3 p.m. Trade pattern analysis is used later in this paper for a description of time in the model used for the predictions.

The goal of this article is to analyze the number of trades in short periods on WSE. The prediction is based on GLM and shrinkage methods. Recent research on GLM is presented by Friedman et al. (2010). The shrinkage methods which are used in this article are presented by Tibshirani (1996), a recent contribution on shrinkage methods can be found in Zou and Hastie (2005). An introduction to GLM models and shrinkage methods in a broader context can be found in Hastie et al. (2008). The analysis consists of two main parts. Firstly, a visualization of the trade pattern in terms of the number of trades is presented. Secondly, the task of predicting the number of trades during small intervals is undertaken. The analysis is conducted for 19 of the 20 biggest Polish listed companies in the period from January 1, 2014 to September 22, 2014. The only company that has been omitted was Orange Polska, which was preceded by Telekomunikacja Polska. The change of ownership and the name caused a technical problem and unreliable results, therefore the company has been removed from the analysis. The time interval that has been chosen for the analysis (frequency) is 5 minutes.

3. Data description and its daily pattern

3.1. Data description

The electronic system of the WSE has been changed since August 8, 2013; the most important change from the data analyst's point of view is the increased preciseness of the trade time from seconds to microseconds. This precision is very convenient from the researcher's point of view since in the period where microseconds are present no cases of two transactions taking place at the same time are reported. The trade hours on the WSE are 9:00 a.m. to 5:05 p.m. However, there is a break from 4:50 p.m. to 5:00 p.m. and during the last 5 minutes trade is not conducted as it is during normal hours, so transactions which took place after 4:50 p.m. are excluded from the analysis. This leaves us with 470 minutes of trade each day.

All transactions on the WSE are quoted in Polish zloty (PLN). At the time this publication is being prepared 1USD=3.95PLN and 1EUR=4.38PLN. In the text we present all values in PLN.

The stocks which form the main Polish equity index, WIG20, have been chosen for the analysis during the period January 1, 2014 to September 22, 2014. Table 1 summarizes the composition of WIG20; note that the analysis is performed for 19 out of 20 stocks from WIG20.

Table 1. The composition of the WIG20 index

Company's name	Abbr. Name	Prime line of business	Index weightening (%)
Alior Bank	ALIOR	Finance	1.67
Asseco Poland	ASSECOPOL	Software	1.94
Bank Pekao	PEKAO	Finance	12.21
Bank Zachodni WBK	BZWBK	Finance	5.62
Eurocash	EUR	FMCG	1.15
Grupa LOTOS	LOTOS	Oil and Natural Gas	0.84
Jastrzębska Spółka Węglowa	JSW	Mining	0.64
Kernel Holding	KERNEL	Food	0.59
KGHM Polska Miedź	KGHM	Mining	8.78
LPP	LPP	Trade	6.08
Lubelski Węgiel „Bogdanka" SA	BOGDANKA	Mining	1.91
mBank	MBANK	Finance	3.19
Orange Polska	ORANGEPL	Telecommunication	3.29
PKN Orlen	PKNORLEN	Fuels	6.85
PKO Bank Polski	PKOBP	Finance	14.44
Polska Grupa Energetyczna	PGE	Energy	6.72
Polskie Górnictwo Naftowe i Gazownictwo	PGNIG	Oil and Natural Gas	4.09
PZU SA	PZU	Insurance	14.09
Synthos	SYNTHOS	Chemistry	1.05
Tauron	TAURONPE	Energy	2.82

Note: The Orange Polska has not been used in the analysis.

The variable being studied is the number of trades in a 5-minute interval for a particular stock. Therefore, the analysis consists of 19 regression problems. The raw data has been transformed into 5-minute data with four variables for each of 19 assets:

- *Number*- total number of trades
- *Plus*- total number of trades with a higher price than the preceding trade
- *Minus*- total number of trades with a lower price than the preceding trade
- *Volume*- total number of shares traded

In addition, the categorical variable *time*, which indicates the time interval, is added and there are 92 levels of this variable. A more in-depth analysis of the time variable is given later in this paper. Additionally, the data set has been divided into training and testing datasets. The training sample is the sample which is used for estimation. The training set is used for the out-of-sample performance analysis and it is not touched during the estimation.

Table 2. Partition of the dataset

Dataset	No. of days	Beginning	End
Total	183	January 1, 2014	September 22, 2014
Training	120	January 1, 2014	June 25, 2014
Testing	63	June 26, 2014	September 22, 2014

Later in this paper, the results presented are derived from the training set if not stated otherwise. Table 3 consists of descriptive statistics for the number of trades for all assets considered. Figure 1 shows the distribution of all four numerical variables considered per asset. Note the logarithmic scale on the graph. The differences in the number of trades and the number of shared traded are considerable across the assets.

Table 3. Descriptive statistics of the number of trades in 5-minute intervals

Company	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
PKOBP	0	14	27	37.71	49	520
PZU	0	9	18	25.94	34	459
PEKAO	0	8	16	22.29	29	420
KGHM	0	15	28	40.65	50	1075
PKNORLEN	0	5	12	18.68	24	432
PGE	0	11	22	29.09	38	472
LPP	0	0	1	4.60	5	227
BZWBK	0	3	9	14.20	19	551
PGNIG	0	5	11	17.67	22	483
MBANK	0	1	5	8.79	11	219
TAURONPE	0	3	7	10.98	14	246

Table 3. Descriptive statistics of the number of trades in 5-minute intervals (cont.)

Company	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ASSECOPOL	0	1	4	5.98	8	165
BOGDANKA	0	0	2	4.89	6	189
ALIOR	0	1	3	6.49	8	356
EUROCASH	0	2	7	12.27	16	489
SYNTHOS	0	1	4	7.28	9	273
LOTOS	0	1	4	7.52	9	371
JSW	0	4	8	11.68	15	325
KERNEL	0	2	6	11.38	14	374

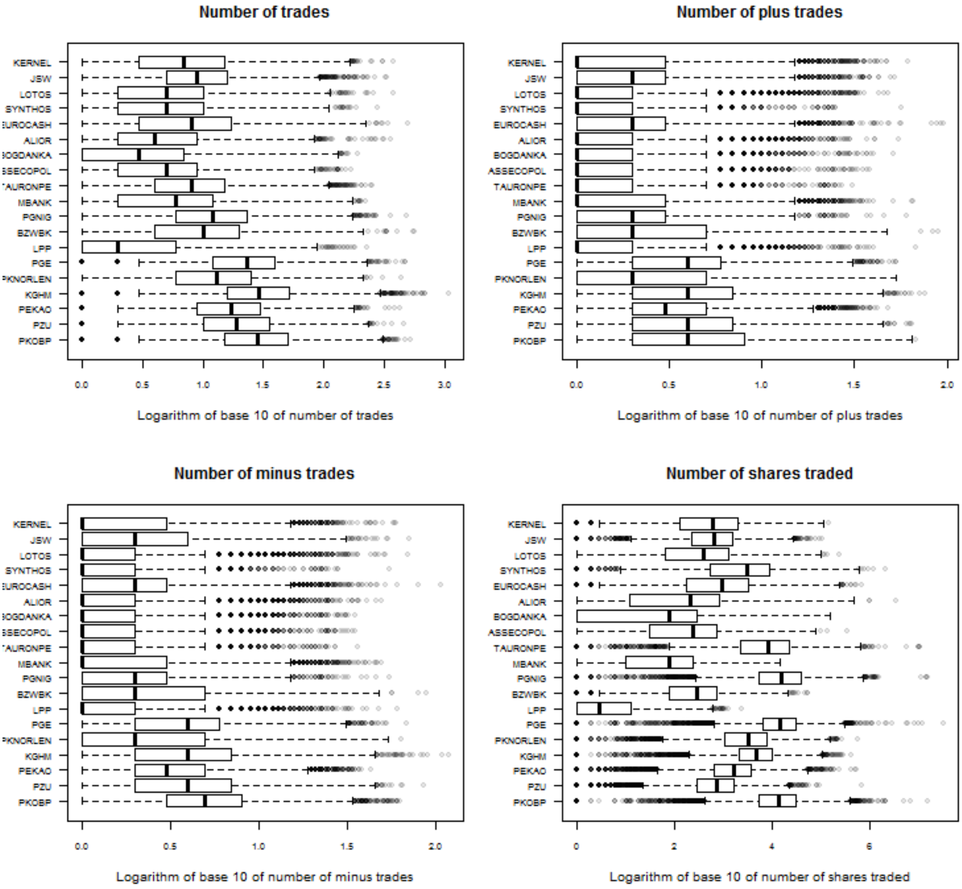


Figure 1. Boxplots of the four analyzed numerical variables for all assets.

Note: "Plus trade"- trade with a higher price than the preceding trade, "minus trade"- lower.

Logarithms are taken of raw values enlarged by one. Outliers are plotted with 90% transparency if the area is black, 10 or more points are plotted in it.

3.2.Daily pattern on WSE

The number of transactions displays a U-shape, which is characteristic of the daily pattern. Just after opening, trade intensity is very high, it drops significantly after a couple of minutes and then a slow downward trend is observed until midday. After that the trend changes to a slow upward one and starts to increase quickly around 3 p.m. The last five minutes are again significantly more intensive. A similar pattern is observable for the other variables considered (plus, minus and volume).

Figures 2 and 3 display the trade pattern for two sample companies, the biggest one - PKOBP and the smallest one - KERNEL. The PKOBP possesses a clear U-shape for all variables, although the KERNEL does not have visibly higher values in the morning (excluding the first interval).

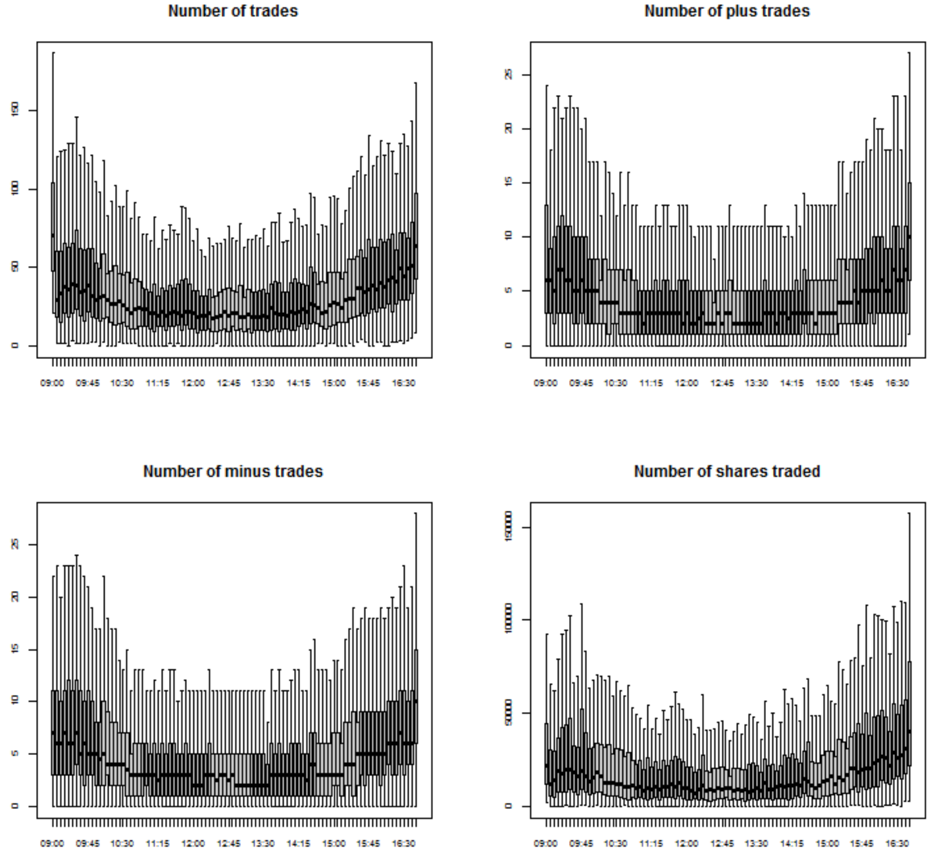


Figure 2. Boxplots of the four considered numerical variables by time for PKOBP.

Note: “Plus trade”- trade with a higher price than the preceding trade, “minus trade”- lower. The outliers have not been plotted.

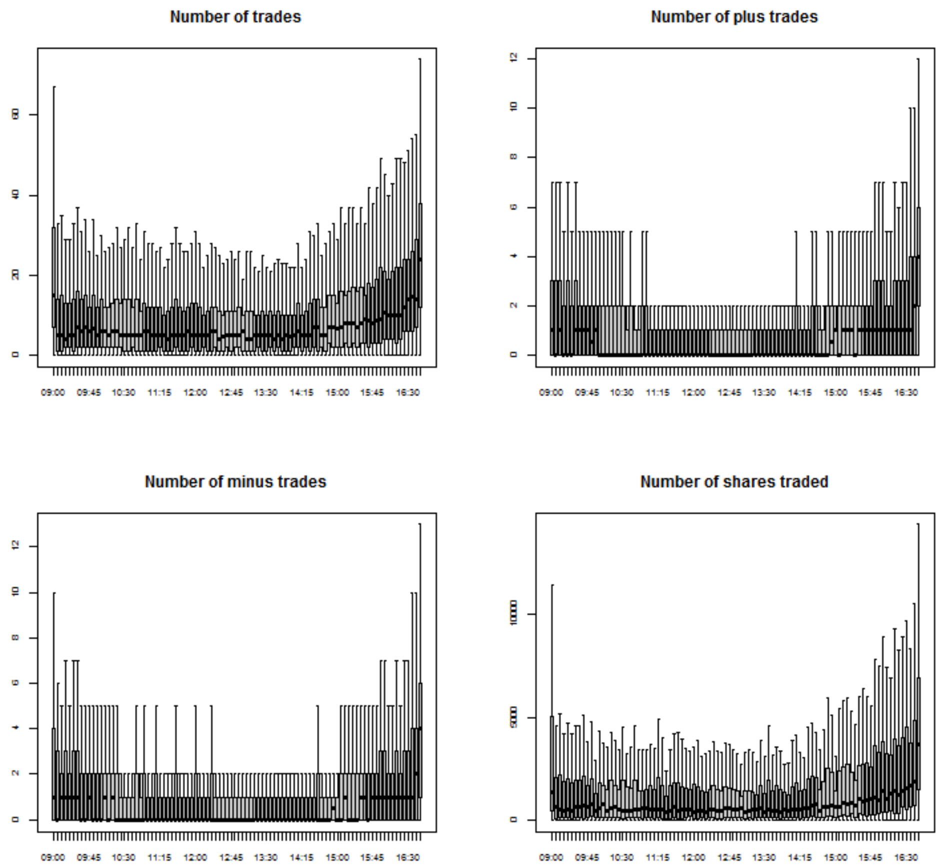


Figure 3. Boxplots of the four considered numerical variables by time for KERNEL.

Note: “Plus trade”- trade with a higher price than the preceding trade, “minus trade”- lower. The outliers have not been plotted.

The main goal of this section is to construct a factor time variable. A factor variable with 92 levels is not practical since if one wanted to use it in a regression model it would have 92 dummy variables. It seems that trade intensity is not significantly different, e.g. during intervals 10:00-10:05 and 10:05-10:10. However, we need a technique to group intervals.

In order to capture the daily pattern differences across names and across variables we use the following procedure:

I. Standardize 76 numerical variables (with zero mean and unit variance); logarithms of number, plus, minus and volume for 19 stocks.

II. Perform Principal Component Analysis (PCA) on standardized variables.

III. Take the first component as a new variable, named PC1, of the PCA.

IV. Perform hierarchical clustering of time intervals (92 levels of the time factor variable) using the PC1 as an explanatory variable.

V. Cluster the intervals by PC1 using K-means clustering.

VI. Divide a cluster if it does not contain a joint set of intervals.

Standardization is a common preliminary step in the PCA, it is done in order to avoid the PCA being driven by variables which have the highest absolute values. The data is skewed and thus the logarithms of variables are taken. PCA is performed on all variables in order to capture potential specific features of particular stocks or variables. Step IV is mainly used in order to visualize the clustering and choose the number of clusters. The complete linkage and average linkage are used and both suggest 4 clusters at most. K-means clustering results in the following clusters:

A. 9:00-9:05 and 16:45-16:50, 2 intervals.

B. 9:05-11:00 and 14:50-15:50, 35 intervals.

C. 11:00-14:50,46 intervals.

D. 15:50-16:45, 11 intervals.

Finally, clusters A and B are divided in order to construct clusters in the form of joint intervals.

Figure 4 summarizes the clustering of time variables. The result is very intuitive and expected. Both extreme intervals constitute separate factors. There are two moderately intensive periods, 9:05-11:00 and 14:50-15:50. The period of the lowest intensity for one consistent cluster is 11:00-14:50. The period 15:50-16:45 is characterized by a significant number of outliers in comparison with others.

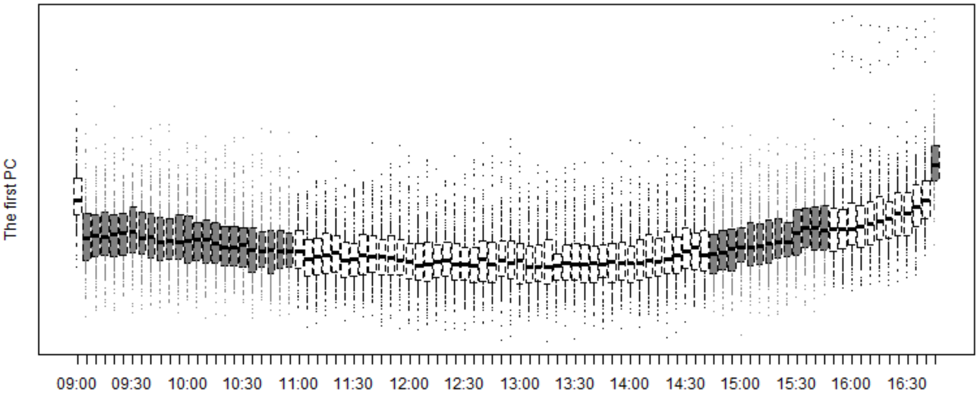


Figure 4. Boxplots of the PC1 variable by time.

Note: The clustering into 6 factor variables is presented by changes in color. Boxplots do not have whiskers, all observations outside quartile ranges are plotted as dots.

4. Forecasts by selected models

The main goal of this article is to verify whether the number of trades during short intervals is predictable. However, we do not aim to find the best algorithm for prediction. The models chosen are based on the GLM. We use a relatively large number of independent variables. In order to avoid overfitting we use two shrinkage methods: ridge regression and lasso regression. The biggest advantage of the linear model is its simplicity and ease of interpretation.

We test four models and compare the performance to the simple random walk. Let $\{N_{it}\}_{t=1}^{T}$ be the series of the number of trades in the *t*-th 5-minute interval for the *i*-th company; analogously $\{P_{it}\}_{t=1}^{T}$-the number of trades with a higher price than the preceding trade; $\{M_{it}\}_{t=1}^{T}$- the number of trades with a lower price than the preceding trade; $\{V_{it}\}_{t=1}^{T}$- volume (number of shares traded).

For the sake of simplicity, let $\left\{X_{jt}\right\}_{t=1}^{T}$ be the set of all observed variables, $j = 1, \dots, 92$.

4.1. Random Walk

The random walk strategy takes the most recent number of trades as the best prediction of future change. The model is defined by:

$$N_{it+1} = N_{it} + a_{it}; \tag{4.1}$$

whereN_t is the actual number of trades at period t and a_{it} is white noise.

Therefore

$$\widehat{N}_{it+1} = N_{it}; \tag{4.2}$$

where$\widehat{N}_{t+1}$ is the forecast of the number of trades for the following period.

4.2. Ridge linear regression

The linear regression prediction is defined by:

$$\widehat{N}_{it+1} = \beta_0^i + \sum_{j=1}^{92} \beta_j^i X_{jt} + \sum_{k=2}^{6} \tau_k^i T_{kt+1}\,; \tag{4.3}$$

where $\beta_0^i, \dots, \beta_{92}^i$ are coefficients for the numeric variables; T_{kt} is a dummy variable which equals one if the *t*-th observation belongs to the *k*-th interval defined in section 3.2 and visualized on Figure 4, and zero otherwise; $\tau_2^i, \dots \tau_6^i$ are intercept changes in comparison with the first interval (9:00-9:05).

The ridge regression is an estimation method which was originally derived in order to solve the problem of colinearity in the multiple linear regression model. This type of regression is similar to the ordinary least squares (OLS), except that the coefficients are estimated by minimizing a different quantity. The ridge regression coefficient estimates are the values that minimize:

$$RSS_i + \lambda \sum_{j=1}^{92} {\beta_j^i}^2 ; \tag{4.4}$$

where

$$RSS_i := \sum_{t=1}^{T-1} \left(N_{it+1} - \beta_0^i - \sum_{j=1}^{92} \beta_j^i X_{jt} - \sum_{k=2}^{6} \tau_k^i T_{kt+1} \right)^2 ; \tag{4.5}$$

whereλ is a tuning parameter. The tuning parameter is chosen in order to minimize in-sample mean squared error (MSE).

4.3. Lasso linear regression

The prediction for lasso linear regression is the same as for ridge linear regression, see (4.3). The difference is in the estimation method. In fact, the difference is in the shrinkage penalty. The estimates of the lasso regression coefficients are the values that minimize:

$$RSS_i + \lambda \sum_{j=1}^{92} \left|\beta_j^i\right|. \tag{4.6}$$

From a practical point of view, the main difference between the ridge and lasso penalty functions is that the latter works as a variable selection method. When lambda increases, the number of coefficients which equal zero also increases. In the case of the ridge regression, when lambda increases, the absolute value of the coefficients decreases to zero, but is not equal to zero. We refer to Tibshirani (1996) for more details on this method.

4.4. Ridge Poisson regression

The number of trades is an integer variable, thus a linear regression approach is not the natural one. Firstly, the fitted values might be negative. Secondly, the fitted values in linear regression are continuous, meaning that the prediction is not an integer. In fact, using linear regression, one assumes normal distribution of the forecast with certain parameters, specifically with the mean given by (4.3).

The distribution of the dependent variable under the Poisson regression is given by:

$$P\left(\widehat{N}_{it+1} = k\right) = \frac{\Lambda^k}{k!} e^{-\Lambda}; \tag{4.7}$$

where

$$log(\Lambda) = \beta_0^i + \sum_{j=1}^{92} \beta_j^i X_{jt} + \sum_{k=2}^{6} \tau_k^i T_{kt+1}. \tag{4.8}$$

The Poisson regression is simply GLM with the Poisson distribution and the exponential link function. The standard estimation is performed using the maximum likelihood (ML) method. Combining the ML with the ridge shrinkage method, the coefficients are the values which minimize:

$$LogLik_i + \lambda \sum_{j=1}^{92} {\beta_j^i}^2 ; \tag{4.9}$$

where

$$LogLik_i \coloneqq \sum_{t=1}^{T-1} \left(N_{it+1}\Theta_i - e^{\Theta_i} - \log(N_{it+1}!) \right); \tag{4.10}$$

and

$$\Theta_i \coloneqq \beta_0^i + \sum_{j=1}^{92} \beta_j^i X_{jt} + \sum_{k=2}^{6} \tau_k^i T_{kt+1}. \tag{4.11}$$

This setting is mathematically quite complex, although it is very intuitive. Formula (4.7) tells us that the number of trades is conditionally a Poisson distribution. Formula (4.8) defines the link function; the logarithm of the parameter in the Poisson distribution is assumed to be linearly linked to the independent variables. Formula (4.9) is similar to (4.4), except that the OLS is not a valid estimation method in the case of GLM, thus the log-likelihood function is used instead. Formulas (4.10) and (4.11) rewrite (4.7) and (4.8) in terms of estimates for GLM.

4.5. Lasso Poisson regression

The prediction for the lasso Poisson regression is the same as for the ridge Poisson regression, see (4.7). The difference is in the estimation method. The lasso Poisson regression coefficients estimates are the values that minimize:

$$LogLik_i + \lambda \sum_{j=1}^{92} \left|\beta_j^i\right|. \tag{4.12}$$

We compare the methods presented in terms of prediction power in the following section.

We refer to Hastie et al. (2008) for an extensive introduction to the methods presented, to Friedman et al. (2010) for recent research on GLM and to Zou and Hastie (2005) for a recent paper on shrinkage methods.

5. Empirical results

We estimated four models described in section 4 for 19 stocks. The presentation and interpretation of 76 models was challenging. Most attention was paid to the prediction ability of those models. Four models were compared with respect to prediction quality. Additionally, an analysis of the estimates was undertaken; the number of parameters combining all models was very high. The linear regression with lasso penalty function is the easiest to analyze estimates from. Thus, the estimates for this model, for all 19 stocks, are presented in Tables 6-9 in the Appendix.

5.1. Prediction accuracy

The quality of prediction is a very broad topic. Most importantly, the in-sample and out-of-sample predictions should be distinguished. It is of course important for the model to work well on the estimation sample (in-sample prediction), although the goal is usually the prediction on the basis of data for which the outcome is unknown (out-of-sample prediction).

Technically, the verification of the prediction abilities of the model focuses on the analysis of residuals, the difference between the prediction and the actual value. The most obvious goal for the model is to produce small residuals in terms of the absolute value. In this article, we only analyze this part of the prediction abilities. Examples of more detailed analyses, which go beyond the scope of this article, are sensitivity analysis or analysis of the distribution of residuals.

Table 4 Prediction quality

	Root Mean Squared Error					**Mean Average Error**				
	Random walk	Ridge LM	Ridge PR	Lasso LM	Lasso PR	Random walk	Ridge LM	Ridge PR	Lasso LM	Lasso PR
PKOBP	38.2	33.6	**32.6**	33.6	32.7	25.8	21.8	**20.8**	21.8	20.8
PZU	23.6	21.3	**20.6**	21.4	20.6	15.6	14.6	13.9	14.5	**13.9**
PEKAO	22.4	20.1	**19.6**	20.2	19.6	15.2	13.4	**13.1**	13.5	13.1
KGHM	34.1	31.6	30.2	31.5	**30.2**	22.6	21.7	**19.7**	21.5	19.7
PKNORLEN	22.0	19.8	**19.3**	19.8	19.3	14.5	12.2	**11.9**	12.2	11.9
PGE	26.8	25.1	**24.4**	25.1	24.5	17.0	16.8	**16.4**	16.8	16.5
LPP	14.1	14.1	13.5	14.1	**13.5**	8.8	6.9	6.7	6.9	**6.7**
BZWBK	21.7	20.0	**18.9**	20.0	18.9	11.8	10.6	**10.4**	10.6	10.4
PGNIG	19.6	18.3	17.4	18.3	17.4	11.7	11.3	**10.8**	11.3	10.8
MBANK	14.8	13.8	13.1	13.7	13.1	8.6	7.4	**7.4**	7.4	7.4
TAURONPE	14.0	12.9	**12.4**	13.0	12.3	8.4	7.7	**7.5**	7.6	7.5
ASSECOPOL	8.8	8.0	**7.6**	8.0	7.6	5.3	4.6	**4.6**	4.7	4.6
BOGDANKA	8.6	7.9	**7.1**	7.9	7.1	4.5	4.0	**3.9**	4.0	3.9
ALIOR	14.6	13.5	**12.3**	13.5	12.3	7.1	5.9	**5.8**	5.9	5.8
EUROCASH	16.5	15.5	15.1	15.5	**15.1**	9.6	9.3	9.1	9.3	**9.1**
SYNTHOS	11.2	10.0	**9.4**	10.0	9.4	6.1	5.5	**5.4**	5.5	5.4
LOTOS	13.7	12.7	12.2	12.7	**12.2**	6.8	5.8	**5.7**	5.8	5.7
JSW	17.1	15.5	**14.7**	15.5	14.7	9.9	8.4	8.2	8.4	**8.2**
KERNEL	12.4	11.7	**10.5**	11.7	10.5	**6.8**	7.8	7.0	7.7	7.0

Note: LM stands for linear model, PR for Poisson Regression. Shaded cells indicate the minimum value per measure and asset out.

Table 4 summarizes prediction power of the proposed models. Firstly, of the chosen models Poisson regression with the ridge penalty function seems to outperform others, although Poisson regression with the lasso penalty function gives very similar results. Secondly, the performance of the prediction is better for more liquid stocks. Comparing MAEs and MSEs in Table 4 to descriptive statistics, especially the mean and median, in Table 3, relative MAEs and relative MSEs are lower for more liquid stocks. Similarly, in the case of less liquid stocks, MAEs and MSEs are not much lower for the predictions of the models in comparison with the naïve strategy. The MAE for KERNEL is even lower using random walk as a predictor. Note that the performance was validated out-of-sample, thus even small differences are meaningful, although they might not be statistically significant. It is impossible for the model to perform worse than the naïve strategy in-sample.

5.2. Estimation results

The second goal of the analysis performed is an evaluation of the predictive power of each of the chosen variables. There are 4 models for each of 19 analyzed stocks with 82 parameters each; this results in 6232 parameters. We present estimates for the linear model with the lasso penalty function. It does not possess the best properties with respect to quality of forecast, although it is easiest to infer from. Estimates in linear model have very natural and clear interpretation: a value of the variable one unit higher has been observed with, on average, the independent variables changed by the value of the coefficient. For the sake of comparability and validity of shrinkage methods, we estimated models using standardized variables, with zero mean and unit variance. Lasso penalty function works as a variable selection, thus it even simplifies the interpretation. The zero coefficients mean that the respective predetermined variables are insignificant.

Table 5. Estimates for time variable.

	09:05-10:55	11:00-14:45	14:50-15:45	15:50-16:40	16:45
PKOBP	-3.65	-9.48	0.24	6.49	42.39
PZU	-2.79	-6.51	-1.66	6.31	28.25
PEKAO	-0.69	-4.4	0.49	4.17	3.7
KGHM	-3.8	-9.79	-5.22	2.11	147.7
PKNORLEN	-4.7	-6.26	-0.31	6.28	11.9
PGE	-2.49	-7	-0.85	7.41	14.18
LPP	-0.7	-0.71	0.13	0.49	2.05
BZWBK	-1.08	-1.71	-0.51	2.4	0.37
PGNIG	-0.2	-4	-1.74	1.57	48.86
MBANK	-2.31	-2.4	0.47	1.9	3.82
TAURONPE	-2.75	-3.44	-2.18	3.62	25.33
ASSECOPOL	-1.02	-2.28	-0.17	2	7.85
BOGDANKA	-1.08	-1.08	0.18	0.71	3.49
ALIOR	-1.27	-1.79	-0.49	1.66	8.91
EUROCASH	-2.3	-2.39	-0.41	2.29	8.58
SYNTHOS	-2.53	-2.51	-1.46	2.55	18.01
LOTOS	-2.26	-2.5	-0.35	2.01	12.28
JSW	-2.46	-3.48	-2.38	1.96	45.16
KERNEL	-3.52	-4.04	-1.27	4.42	14.76

Table 5 summarizes the time pattern in the model. The coefficients are interpreted as the difference between the number of trades during 5-minute intervals in the corresponding interval in comparison with the first interval, 9:00-9:05. It can be seen that for all stocks trade intensity is lower during 9:05-11:00 than during 9:00-9:05. In the interval 14:50-15:50 trade intensity is comparable to the first 5 minutes and increases between15:50 and 16:45. In the last 5 minutes it spikes for every stock. It is important to note that the coefficients during the periods 9:05-11:00 and 14:50-15:50, and partially in 14:50-15:50, are negative. It means that intercepts for those periods are lower than the ones for 9:00-9:05. This is surprising because the average number of trades is higher in the first interval. The interpretation is that more information is stored in other variables than in the intercept.

All estimates for this model are presented in the Appendix, here only summary and interpretations are presented. There is a strong autoregression observed in the data, all estimates for the number of the corresponding stock are positive and relatively very high. The cross-sectional autoregression is also visible and mostly positive, meaning that high intensity in one period is observed before high intensity in the next period for other stocks on average. This property is mostly seen for big and liquid companies, and for some less liquid ones like BOGDANKA or ASSECO we see a negative coefficient.

The plus variable (the number of trades with a higher price than the preceding one) has a positive coefficient for the corresponding stocks (or zero in 3 cases). It means that plus-trades precede a higher intensity in trade, although the coefficients are much smaller than in the case of the number variable. The cross-sectional dependence between the plus variable and the number of trades is not clear, some coefficient are negative, some positive and a large number equal zero.

The minus variable (the number of trades with a lower price than the preceding one) seems less influential than the others, and there is no clear pattern recognized. There are many zero-coefficient ones and the majority of non-zero ones are positive. This, however, might be caused by the correlation between minus, plus and number variables.

There is a very interesting relationship observed for the volume variable. All except one (which is zero) coefficient are negative for the volume in the model for the corresponding stock. The immediate interpretation might be that a higher volume in one period is observed with a lower intensity in the next period. However, the number and the volume variables are highly correlated, thus the volume variable works as an off-set. After a period with many trades, there is a period which also has many trades (positive autocorrelation), although if those trades are relatively large (many stocks traded per trade) this relationship is lower. The cross-sectional dependence between the volume variable and the number of trades is low; most of the coefficients are zero and there is no visible pattern for the rest.

6. Conclusions

The forecasting of the number of trades in a given period is an alternative to the ACD model. The results obtained show a significant forecasting ability of GLM. Shrinkage methods such as lasso and ridge penalty functions are valuable tools in the model selection problem. For most of the stocks analyzed, the Poisson regression with the ridge penalty function is shown to be the best model in terms of MSE and MAE for the out-of-sample forecast. The Poisson regression with lasso penalty function gave almost the same results for all the stocks analyzed. The forecasting ability of the models proposed seems to work better for more liquid assets, with a higher average number of trades.

The analysis of the daily pattern of trade intensity showed 6 periods during a trading day. During each period, trade intensity is statistically similar. For a 5-minute interval during different periods, the expected number of trades is different. Those periods are 09:00-9:05, 9:05-11:00, 11:00-14:50, 14:50-15:50, 15:50-16:45 and 16:45-50. The trade pattern seems marginally different for more liquid stocks than for less illiquid ones. Liquid assets show the expected U-shape in the daily pattern. Trade is relatively intense in the first period (09:00-9:05), in the second (9:05-11:00) it is lower and it is lowest in the third period (11:00-14:50). The intensity in the fourth (14:50-15:50) period is comparable to the intensity in the second one (9:05-11:00). The intensity in the fifth period (15:50-16:45) is higher than in preceding ones, relatively very high values (outliers) are observed more often in this period. Trade intensity is absolutely highest in the last 5-minute interval. However, for relatively illiquid assets, trade during 9:05-15:50 is similarly intense and becomes more intense only in the last hour of trade.

There is a significant positive serial correlation observed in the high frequency data. Most of the data shows positive cross-sectional autoregression. This means that in most cases high intensity in one period is observed before high intensity in the next period for other stocks. This property is mostly detected for big and liquid companies. For some less liquid firms like BOGDANKA or ASSECO, we see negative coefficients.

The cross-sectional dependence between the plus variable (the number of trades with a higher price than the preceding one) and the number of trades is not clear, some coefficients are negative, some positive and a large number equal zero. In the case of the minus variable (the number of trades with a lower price than the preceding one) no clear pattern can be recognized. There are many zero-coefficients and the majority of non-zero ones are positive. This, however, might be caused by correlations among regressors.

The volume variable (the value of trades for a corresponding stock in a given interval) shows an interesting feature. All coefficients for the volume except one (which is zero) are negative for the corresponding stock. A possible interpretation might be that a higher volume in one period is associated with a lower intensity in the next period. However, the number and the volume variables are highly correlated, thus the volume variable works as an off-set. After a period with many trades, there is a period which also has many trades (positive autocorrelation), although if the volume of transactions is high (many stocks traded per trade) this relationship becomes weaker. The cross-sectional dependence between the volume variable and the number of trades is low, most of the coefficients are zero and there is no visible pattern in the rest.

There are at least three potential directions for further studies. Firstly, the forecast abilities of other models are needed. Machine-learning techniques like neural network often possess better forecasting abilities than linear or GLM. Other approaches based on linear models are also possible, for instance, the time variable or input to the model can be defined in a different way.

Secondly, different regressors might be used in the regression. The four (except the time) variables chosen are relatively easy to interpret. However, they may not possess the best prediction power. Moreover, they are all strongly correlated, thus the results might not be stable. One of the possible regressors which might be used in the further analysis is the percentage of trades with a higher price. In addition, more lags may be taken into account. In this paper, the authors have used only variables obtained from the trades in one preceding interval.

Thirdly, the analysis may be repeated for longer and shorter intervals. In this paper, the authors show the analysis for the number of trades during 5-minute intervals. It is possible that some conclusions would be different for shorter or longer ones.

REFERENCES

BAUWENS, L.,GIOT., P., (2000). The logarithmic ACD model: An application to the bid–ask quote process of three NYSE stocks, “Annales D’economie Et De Statistique”, 60, pp. 117–149.

BAUWENS, L.,HAUTSCH, N., (2009). Modelling financial high frequency data using point processes, pp. 953–979, Berlin: Springer.

CARTEA, Á., JAIMUNGAL, S., (2013). Modelling Asset Prices for Algorithmic and High-Frequency Trading, “Applied Mathematical Finance”, Vol. 20, No. 6, pp. 512–547.

CARTEA, Á., MEYER-BRANDIS, T., (2010). How duration between trades of underlying securities affects option prices, “Review of Finance”, 14 (4), pp. 749–785.

de JONG, F., RINDI, B., (2009). The microstructure of financial markets (1st ed.), Cambridge: Cambridge University Press.

DIAMOND, D. W.,VERRECHIA, R. E., (1987). Constraints on short-selling and asset price adjustment to private information, “Journal of Financial Economics”, 18, pp. 277–311.

DUFOUR, A., ENGLE, R. F., (2000). Time and the price impact of a trade, “The Journal of Finance”, LV (6), pp. 2467–2498.

EASLEY, D., O’HARA, M., (1992). Time and the process of security price adjustment, “The Journal of Finance”, XLVII (2), pp. 577–605.

ENGLE, R. F., (2000). The econometrics of ultra-high-frequency data, ”Econometrica”, 68 (1), pp. 1–22.

ENGLE, R. F., RUSSELL, J. R., (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data,” Econometrica”, 66 (5), pp. 1127–1162.

FERNANDES, M., GRAMMIG, J., (2005). Nonparametric specification tests for conditional duration models, “Journal of Econometrics”, 127 (1), pp. 35–68.

FRIEDMAN, J., TIBSHIRANI, R., HASTIE, T., (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent, “Journal of Statistical Software”, Vol. 33, No. 1. http://www.jstatsoft.org/v33/i01.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J., (2008).The Elements of Statistical Learning, 2nd edition, Springer, New York.

HUJER, R., VULETIC, S., KOKOT, S., (2002). The Markov switching ACD model, SSRN eLibrary, Retrieved from http://ssrn.com/abstract=332381.

MAHEU, J. M., MCCURDY, T. H., (2000). Volatility dynamics under duration-dependent mixing, “Journal of Empirical Finance”, 7 (3–4), pp. 345–372.

MANGANELLI, S., (2005). Duration, volume and volatility impact of trades, “Journal of Financial Markets”, 8 (4), pp. 377–399.

MEITZ, M., TERASVIRTA, T., (2006). Evaluating models of autoregressive conditional duration, “Journal of Business & Economic Statistics”, 24, pp. 104–124.

RENAULT, E., VAN DER HEIJDEN, T.,WERKER, B. J. M., (2012). The dynamic mixed hitting-time model for multiple transaction prices and times, Working Paper, Retrieved fromhttp://dx.doi.org/10.2139/ssrn.2146220.

TIBSHIRANI, R., (1996). Regression Shrinkage and Selection via the Lasso, “Journal of the Royal Statistical Society, Series B”, Vol. 58, No. 1, pp. 267–288.

ZHANG, M. Y., RUSSELL, J. R., TSAY, R. S., (2001). A nonlinear autoregressive conditional duration model with applications to financial transaction data, “Journal of Econometrics”, 104 (1), pp. 179–207.

ZOU, H., HASTIE, T., (2005). Regularization and Variable Selection via the Elastic Net, “Journal of the Royal Statistical Society, Series B”, Vol. 67, No. 2, pp. 301–320.

APPENDIX

Table 6. Estimates for number variable

	PKOBP Number	PZU Number	PEKAO Number	KGHM Number	PKNORLEN Number	PGE Number	LPP Number	BZWBK Number	PGNIG Number	MBANK Number	TAURONPE Number	ASSECOPOL Number	BOGDANKA Number	ALIOR Number	EUROCASH Number	SYNTHOS Number	LOTOS Number	JSW Number	KERNEL Number
PKOBP	**15.92**	**1.17**	**2.47**	0	0	*-0.1*	0	0	**2.55**	0	0	0	*-0.74*	*-0.37*	0	0	0	**0.27**	0
PZU	**1.87**	**11.1**	**0.34**	0	**0.41**	**0.6**	0	0	**0.43**	0	0	*-0.59*	0	0	**0.74**	0	0	**0.34**	*-0.57*
PEKAO	**1.48**	0	**9.01**	**0.06**	**0.28**	**0.26**	0	**0.3**	**0.01**	0	**0.09**	*-0.07*	*-0.06*	*-0.18*	0	0	0	0	0
KGHM	*-0.37*	**1.2**	0	**25.61**	0	0	**1.62**	0	**1.02**	*-0.72*	**0.08**	*-0.31*	0	0	0	*-0.09*	0	0	0
PKN.	**0.2**	**0.12**	0	**1.04**	**8.31**	0	*-0.38*	0	**0.64**	0	**0.7**	0	*-0.37*	0	**0.33**	**1.09**	*-0.03*	0	*-0.81*
PGE	**0.28**	0	**0.17**	**0.24**	0	**16**	**0.46**	**0.2**	**0.63**	0	**0.73**	*-0.1*	*-0.16*	**0.01**	0	0	0	*-0.18*	*-0.11*
LPP	0	0	0	0	0	**0.24**	**1.11**	**0.13**	0	0	0	0	0	**0.24**	0	0	0	0	**0.08**
BZWBK	**0.05**	**0.81**	**0.11**	**0.01**	**0.42**	**0.39**	**0.24**	**6.06**	**0.08**	0	0	0	0	**0.24**	0	0	**0.16**	**0.2**	*-0.06*
PGNIG	0	**0.04**	0	**1.46**	**0.07**	**0.31**	**0.5**	**0.19**	**9**	0	**0.33**	*-0.14*	0	**0.07**	0	0	0	0	0
MBANK	**0.55**	**0.05**	**0.2**	**0.24**	**0.03**	**0.25**	**0.43**	0	*-0.28*	**2.83**	**0.02**	0	0	**0.03**	0	**0.19**	0	**0.33**	0
TAU.	0	**0.22**	**0.12**	**0.93**	**0.31**	**1.53**	0	0	**0.42**	0	**6.37**	0	*-0.53*	0	*-0.29*	**0.67**	*-0.23*	*-0.28*	*-0.35*
ASSEC.	**0.27**	0	**0.4**	**0.1**	**0.15**	**0.12**	**0.16**	0	0	0	**0.07**	**3.75**	0	*-0.02*	**0.01**	0	0	**0.54**	**0.02**
BOGD.	0	**0.11**	*-0.08*	0	0	**0.05**	**0.15**	**0.15**	**0.08**	0	0	**0.08**	**2.32**	**0.06**	0	0	0	0	0
ALIOR	0	0	**0.1**	0	0	0	**0.09**	**0.07**	*-0.07*	0	0	0	0	**1.95**	0	**0.19**	0	0	0
EURO.	0	**0.07**	0	0	**0.32**	**0.27**	0	0	**0.01**	0	0	*-0.61*	0	0	**5.84**	*-0.22*	0	0	0
SYNT.	0	0	0	**0.04**	**0.4**	**0.2**	0	0	**0.25**	0	**0.45**	0	*-0.23*	0	*-0.11*	**3.99**	0	**0.05**	0
LOTOS	0	0	*-0.05*	0	0	**0.18**	*-0.01*	0	**0.1**	0	**0.07**	0	*-0.04*	0	0	0	*-0.58*	0	0
JSW	**0.07**	0	**0.05**	0	0	**0.13**	0	0	0	0	0	0	0	0	0	0	0	*-1.29*	0
KERNEL	*-0.14*	*-0.17*	**0.02**	**0.3**	*-0.17*	*-0.46*	*-0.08*	**0.06**	0	*-0.04*	**0.05**	**0.09**	**0.31**	*-0.22*	0	0	**0.1**	**0.29**	*-2.12*

Table 7. Estimates for plus variable

	PKOBP Number	PZU Number	PEKAO Number	KGHM Number	PKNORLEN Number	PGE Number	LPP Number	BZWBK Number	PGNIG Number	MBANK Number	TAURONPE Number	ASSECOPOL Number	BOGDANKA Number	ALIOR Number	EUROCASH Number	SYNTHOS Number	LOTOS Number	JSW Number	KERNEL Number
PKOBP	**0.51**	0	**0.21**	**0.11**	**0.63**	**0.62**	*-0.17*	**1.44**	**0.32**	**1.51**	*-0.35*	0	0	0	*-0.41*	*-0.39*	*-0.99*	*-0.07*	0
PZU	**0.34**	**0.11**	**0.06**	**0.43**	**0.28**	**0.07**	0	**1.14**	0	**0.05**	**0.64**	*-0.37*	0	0	0	**0.82**	0	**0.86**	0
PEKAO	**0.37**	**0.3**	**0.98**	**1.29**	0	**0.51**	0	0	0	0	0	0	0	0	*-0.2*	0	*-0.32*	**0.61**	0
KGHM	0	0	**0.43**	**2.22**	*-0.11*	*-0.8*	*-1.07*	0	**0.43**	0	**0.59**	*-1.16*	*-2.19*	0	**0.78**	**1.49**	**0.94**	**0.92**	**1.28**
PKN.	**0.24**	**0.23**	0	*-0.43*	0	**0.17**	**0.24**	**0.49**	0	0	**0.56**	0	**0.3**	**0.04**	**0.22**	**0.43**	0	**0.28**	0
PGE	0	**0.63**	**1**	0	0	**1.01**	**0.15**	0	**0.1**	*-0.41*	**0.69**	0	*-0.22*	0	0	0	*-0.71*	0	*-0.16*
LPP	*-0.03*	0	0	**0.04**	*-0.05*	0	**0.08**	0	*-0.13*	0	0	*-0.09*	**0.08**	0	**0.06**	**0.15**	**0.02**	**0.01**	**0.01**
BZWBK	0	0	**0.15**	0	*-0.27*	0	**0.12**	**0.22**	0	**0.22**	0	0	0	**0.31**	0	**0.26**	0	0	*-0.22*
PGNIG	0	0	**0.08**	0	0	0	0	0	**0.71**	0	0	*-0.38*	0	**0.01**	**0.26**	0	**0.06**	0	**0.09**
MBANK	0	**0.17**	0	*-0.23*	**0.04**	**0.09**	**0.07**	**0.16**	**0.09**	**0.25**	0	0	**0.17**	**0.33**	**0.22**	*-0.08*	**0.18**	0	0
TAU.	0	*-0.11*	*-0.17*	*-0.14*	*-0.26*	*-0.15*	*-0.09*	*-0.09*	*-0.03*	0	*-0.19*	**0.27**	**0.35**	**0.02**	**0.28**	0	0	**1.15**	**0.11**
ASSEC.	0	*-0.17*	*-0.21*	*-0.33*	0	**0.06**	*-0.49*	0	**0.03**	**0.05**	**0.16**	0	0	**0.1**	*-0.13*	**0.1**	*-0.13*	0	0
BOGD.	0	0	0	0	0	0	0	0	*-0.07*	*-0.01*	0	0	**0.65**	**0.07**	**0.12**	**0.17**	0	**0.17**	**0.03**
ALIOR	0	0	0	**0.04**	*-0.12*	*-0.33*	**0.44**	0	*-0.21*	**0.37**	0	0	**0.52**	**1.16**	**0.28**	**0.1**	**0.03**	*-0.19*	**0.06**
EURO.	0	0	**0.16**	**0.05**	0	0	0	0	0	**0.32**	*-0.29*	0	**0.19**	*-0.11*	0	0	0	**0.53**	**0.29**
SYNT.	**0.02**	0	0	0	0	0	0	0	0	0	0	**0.04**	0	0	0	**0.06**	**0.15**	0	0
LOTOS	0	0	0	0	0	0	0	0	0	*-0.25*	*-0.18*	0	**0.14**	0	**0.19**	**0.28**	**0.68**	*-0.14*	**0.08**
JSW	**0.09**	0	*-0.15*	0	0	0	**0.12**	**0.29**	**0.07**	0	0	**0.11**	*-0.18*	**0.07**	0	**0.62**	0	**1.2**	*-0.13*
KERNEL	*-0.47*	**0.77**	*-0.63*	0	**0.03**	*-0.61*	**0.51**	*-0.03*	**0.39**	0	**0.65**	**0.54**	**0.65**	**0.64**	**0.07**	**1.8**	**0.52**	*-0.57*	**1.31**

Table 8. Estimates for minus variable

	PKOBP Number	PZU Number	PEKAO Number	KGHM Number	PKNORLEN Number	PGE Number	LPP Number	BZWBK Number	PGNIG Number	MBANK Number	TAURONPE Number	ASSECOPOL Number	BOGDANKA Number	ALIOR Number	EUROCASH Number	SYNTHOS Number	LOTOS Number	JSW Number	KERNEL Number
PKOBP	0	**0.64**	**0.59**	**0.03**	**0.17**	0	*-0.73*	0	0	0	0	**0.81**	*-0.33*	0	0	0	**0.78**	**0.71**	0
PZU	0	**0.5**	**0.18**	**0.45**	**0.59**	0	*-0.9*	0	*-0.34*	0	0	0	0	*-0.28*	0	*-0.06*	**0.74**	**0.31**	**0.5**
PEKAO	0	**0.4**	0	**0.04**	0	0	0	**0.03**	**0.33**	0	**0.67**	0	*-0.74*	0	0	0	0	0	*-0.09*
KGHM	0	**0.16**	0	**3.78**	*-0.07*	0	*-0.69*	**0.17**	*-0.06*	**0.38**	0	*-1.15*	**0.49**	0	**0.21**	**1.72**	*-0.21*	**0.38**	**0.74**
PKN.	**0.12**	**0.77**	**0.58**	0	0	**0.46**	*-0.97*	**0.04**	**0.45**	**0.1**	**0.23**	*-1.26*	*-0.32*	0	*-0.67*	**0.49**	**0.46**	**0.86**	**1.03**
PGE	0	0	0	**0.6**	0	*-3.72*	0	0	0	0	**0.21**	0	*-0.05*	**0.19**	0	0	**0.16**	0	**0.44**
LPP	0	*-0.11*	0	0	0	*-0.07*	**0.39**	*-0.11*	*-0.01*	**0.09**	0	0	**0.06**	**0.11**	0	*-0.08*	**0.02**	0	**0.04**
BZWBK	0	0	*-0.07*	**0.03**	0	0	0	**0.05**	0	*-0.35*	*-0.09*	0	*-0.18*	0	**0.29**	**0.01**	0	**0.13**	*-0.08*
PGNIG	**0.23**	**0.05**	0	0	**0.41**	0	0	0	**1.08**	0	0	0	0	0	0	**0.15**	**0.41**	**0.3**	**1.14**
MBANK	*-0.02*	0	*-0.32*	0	*-0.02*	0	*-0.66*	**0.18**	0	**1.21**	**0.48**	*-0.14*	**0.09**	**0.17**	*-0.04*	0	**0.39**	*-0.13*	*-0.07*
TAU.	**0.12**	0	0	*-0.05*	**0.26**	*-0.87*	*-0.45*	0	0	0	**0.11**	**0.17**	0	**0.13**	*-0.06*	**0.08**	**0.21**	0	**0.21**
ASSEC.	0	0	*-0.29*	0	*-0.02*	*-0.01*	*-0.34*	*-0.03*	*-0.09*	**0.35**	**0.21**	*-0.27*	**0.12**	**0.21**	**0.02**	0	0	*-0.47*	**0.04**
BOGD.	0	**0.04**	0	**0.16**	*-0.01*	0	*-0.14*	0	0	0	**0.13**	0	**0.68**	0	**0.11**	0	**0.23**	0	0
ALIOR	*-0.07*	*-0.03*	*-0.32*	**0.05**	*-0.02*	*-0.11*	0	0	0	0	0	**0.05**	0	**1.39**	0	0	**0.27**	0	0
EURO.	**0.46**	0	0	**0.79**	**0.08**	**0.06**	*-0.11*	*-0.15*	0	*-0.18*	0	0	**0.09**	**0.13**	**0.86**	*-0.28*	**0.25**	**0.15**	**0.22**
SYNT.	0	**0.04**	0	0	0	*-0.1*	0	**0.17**	0	0	0	*-0.26*	0	**0.24**	0	0	0	**0.16**	**0.22**
LOTOS	0	**0.06**	0	**0.25**	0	*-0.1*	0	0	0	0	0	0	**0.08**	**0.28**	**0.1**	**0.09**	**1.38**	0	**0.18**
JSW	0	**0.36**	**0.07**	**0.65**	0	0	*-0.34*	0	0	*-0.03*	**0.08**	0	0	**0.48**	0	**0.36**	**0.03**	**0.18**	0
KERNEL	**0.1**	*-0.35*	*-0.42*	*-0.37*	**0.39**	*-1.16*	0	**0.72**	**0.66**	**0.45**	**0.64**	*-0.01*	*-0.25*	**0.31**	0	**0.52**	**0.83**	0	**1.06**

Table 9. Estimates for volume variable

	PKOBP Volume	PZU Volume	PEKAO Volume	KGHM Volume	PKNORLEN Volume	PGE Volume	LPP Volume	BZWBK Volume	PGNIG Volume	MBANK Volume	TAURONPE Volume	ASSECOPOL Volume	BOGDANKA Volume	ALIOR Volume	EUROCASH Volume	SYNTHOS Volume	LOTOS Volume	JSW Volume	KERNEL Volume
PKOBP	*-1.21*	0	*-0.64*	**0.24**	0	0	0	*-0.24*	*-0.68*	0	*-0.11*	*-0.04*	0	*-0.11*	0	*-0.07*	0	0	*-0.12*
PZU	**0.09**	*-2.29*	**0.05**	**0.44**	0	0	*-0.3*	*-0.08*	0	**0.09**	0	0	0	0	0	**0.03**	0	0	0
PEKAO	0	**0.02**	*-1.88*	0	**0.01**	0	0	0	**0.06**	0	0	0	0	*-0.06*	**0.06**	0	0	0	0
KGHM	**0.06**	*-0.53*	0	*-5.22*	*-0.04*	**0.17**	*-0.97*	**0.14**	**0.19**	0	0	0	*-0.02*	**0.13**	*-0.01*	*-0.15*	**0.1**	**0.27**	*-0.18*
PKN.	**0.36**	*-0.14*	0	*-0.11*	*-1.58*	**0.09**	0	0	*-0.12*	0	*-0.16*	0	*-0.09*	**0.09**	0	*-0.2*	0	0	0
PGE	0	0	0	**0.45**	0	*-2.98*	0	**0.08**	**0.13**	0	0	0	0	0	**0.04**	*-0.1*	**0.03**	0	*-0.16*
LPP	0	0	**0.01**	0	0	**0.02**	0	**0.02**	**0.06**	0	0	*-0.03*	0	0	**0.01**	0	*-0.01*	0	**0.04**
BZWBK	**0.34**	0	0	**0.1**	**0.05**	**0.09**	0	*-1.41*	**0.08**	**0.15**	**0.01**	*-0.02*	0	0	**0.04**	*-0.07*	*-0.02*	0	0
PGNIG	0	0	0	0	0	0	0	0	*-1.27*	0	0	0	0	0	0	0	**0.01**	**0.17**	0
MBANK	*-0.13*	**0.12**	*-0.02*	**0.01**	**0.16**	0	0	**0.02**	**0.16**	*-0.82*	**0.01**	*-0.05*	0	0	*-0.07*	0	*-0.05*	0	0
TAU.	0	0	0	*-0.04*	**0.01**	0	0	**0.04**	**0.02**	0	*-1.13*	*-0.08*	**0.1**	**0.01**	**0.12**	*-0.18*	0	0	**0.03**
ASSEC.	0	**0.06**	*-0.02*	**0.11**	*-0.04*	0	0	0	**0.02**	*-0.05*	*-0.02*	*-0.69*	*-0.02*	*-0.02*	0	0	0	*-0.06*	0
BOGD.	0	0	0	0	0	0	0	0	**0.04**	**0.02**	**0.04**	0	*-0.44*	0	0	*-0.04*	*-0.01*	0	**0.01**
ALIOR	0	0	0	0	0	**0.38**	**0.12**	0	**0.18**	*-0.06*	0	*-0.01*	*-0.05*	*-0.26*	*-0.05*	0	*-0.02*	**0.05**	**0.03**
EURO.	0	0	*-0.13*	**0.22**	0	**0.08**	0	**0.07**	0	*-0.06*	**0.01**	0	0	0	*-0.92*	0	0	0	0
SYNT.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*-0.63*	0	0	0
LOTOS	0	0	*-0.05*	0	0	**0.18**	*-0.01*	0	**0.1**	0	**0.07**	0	*-0.04*	0	0	0	*-0.58*	0	0
JSW	**0.07**	0	**0.05**	0	0	**0.13**	0	0	0	0	0	0	0	0	0	0	0	*-1.29*	0
KERNEL	*-0.14*	*-0.17*	**0.02**	**0.3**	*-0.17*	*-0.46*	*-0.08*	**0.06**	0	*-0.04*	**0.05**	**0.09**	**0.31**	*-0.22*	0	0	**0.1**	**0.29**	*-2.12*

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 115–138

THE CHALLENGES OF THE POPULATION CENSUS ROUND OF 2020. OUTLINE OF THE METHODS OF QUALITY ASSESSMENT OF POPULATION CENSUS DATA[1]

Jan Kordos[2]

ABSTRACT

The author begins with the challenges of the population census round of 2020 and gives an outline of the methods of quality assessment of population census data. Next, the synthesis of two Eurostat documents (2007, 2009) and UN Statistics Division (2010) monograph relating to census data quality assessment are considered. Special attention is paid to three methods of census quality assessment: (i) demographic method, (ii) post-enumeration survey, and (iii) comparison with existing household surveys. Census quality assessment methods are discussed in the context of the international recommendations and experience of selected countries. Some Polish experience in these fields is also discussed, and some suggestions to the 2021 Polish Census of Population and Housing preparations are given.

Key words: census of population, demographic analysis, quality assessment, post enumeration survey, household survey.

1. Introduction

Conference of European Statisticians Recommendations for the 2020 Censuses of Population and Housing is published[3]. The main objectives of this Recommendations are:

(a) to provide guidance and assistance to countries in the planning and execution of their population and housing censuses; and

(b) to facilitate and improve the comparability of a census at the UN regional level through the identification of a core set of census topics and the harmonization of concepts, definitions and classifications.

[1] This is an updated and extended version of the author's paper presented at the Plenary Session of the Committee on Demographic Studies, Polish Academy of Sciences, on 15 November 2010.

[2] Warsaw Management University.

[3] http://unstats.un.org/unsd/demographic/sources/census/census3.htm.

The Recommendations are also expected to be used as the general framework for the European Union programme for the 2021 round of population and housing censuses, from which data are to be provided to Eurostat by Member States under the provisions of Regulation (EC) 763/2008.

The author would like to focus his attention on the quality of census of population and housing presented in different documents. His first selection are two Eurostat publications: *Handbook on Data Quality Assessment Methods and Tools* (Eurostat, 2007) and *Handbook on Quality Reports* (Eurostat, 2009), which relate to the quality assessment in statistics, and particularly to the quality assessment of censuses of population. However, his special concern is the publication of the UN Statistics Division on *Post-enumeration Surveys* (2010), which is extremely important for population census results evaluation.

These Eurostat publications underline that the production of high quality statistics depends on the assessment of data quality (Eurostat, 2007, p.5): "*Without a systematic assessment of data quality, the statistical office will risk to lose control of the various statistical processes such as data collection, editing or weighting. Doing without data quality assessment would result in assuming that the processes can not be further improved and that problems will always be detected without systematic analysis. At the same time, data quality assessment is a precondition for informing the users about the possible uses of the data, or which results could be published with or without a warning. Certainly, without good approaches for data quality assessment statistical institutes are working in the blind and can make no justified claim of being professional and of delivering quality in the first place. Assessing data quality is therefore one of the core aspects of a statistical institute's work*".

The quality of population and housing census data is very important for many reasons, building public trust and understanding in the national statistical system. The purpose of census evaluation is to provide users with a level of confidence when utilizing the data, and to explain errors in the census results. It is therefore important to choose an appropriate way of sending out these messages to the right group of people.

2. Errors in population censuses and methods of their assessment

It is universally accepted that a population census is not perfect and that errors can and do occur at all stages of the census operation. Errors in the census results are classified into two general categories:

- coverage errors, and
- content errors.

Coverage errors are the errors that arise due to omissions or duplications of persons or housing units in the census enumeration.

Content errors are errors that arise in the incorrect reporting or recording of the characteristics of persons, households and housing units enumerated in the census.

Many countries have recognized the need to evaluate the overall quality of their census results and have employed various methods for evaluating census coverage as well as certain types of content error. Numerous methods are available to estimate the coverage and content error of censuses. These include:

- simple techniques of quality assurance such as internal consistency checks;
- comparisons of results with other data sources including previous censuses, current household surveys and/or administrative records are also useful techniques[4].

However, for evaluating census data there are two very important methods:

- ***demographic analysis*** and
- ***post enumeration surveys.***

Nevertheless, countries have frequently encountered problems in implementing evaluation methods, and especially in designing and implementation post enumeration surveys (UN Statistics Division, 2010). For this reasons, the author decided, after presenting basic methods of census quality evaluation and some experience of selected countries in this field, to focus his attention on ***post enumeration survey designing and implementation.***

2.1. Demographic analysis of census results

By undertaking a demographic analysis, results from a census may be compared with data from other demographic systems such as vital registration of births and deaths including net migration if such data are available. The cohort component method of demographic analysis uses data from successive censuses as well as life-table survival rates, age-specific fertility rates and estimated levels of international migration between censuses. The population is projected forward to the reference date of the second census based on estimated levels and age schedules of fertility, mortality and migration, and the expected population is compared with enumerated population in the second census.

Another method of analysis involves comparing age distributions of successive censuses. This method is widely used because it requires little data. Yet another method in use is the cohort survival regression method, which uses population counts by age from two censuses and deaths by age during the inter-census period to estimate the coverage rate.

[4] See: UN Economic Commission for Europe (2011, 2015).

Record checks

In this type of analysis, census records are matched with a sample of records from the vital registration or other identification systems, where the relevant respondents to the census questionnaire are traced to the time synchronized with the census. Such sources include previous censuses, birth registrations, school enrolment registries, voter registration list, health and social security records, immigration registers, national or citizen registration cards, etc. Both coverage and content errors could be measured through such comparisons. For coverage evaluation purposes, the following pre-conditions are necessary:

(i) a large proportion of the census target population should be covered in the record system;

(ii) the census and record system should be independent from each other; and

(iii) there should be sufficient information in the records so that accurate matching is possible.

For content evaluation purposes, the record system should contain some relevant items covered in the census such as age, sex, education, income, etc. It is important to ensure that the definitions of items are the same.

2.2. Post Enumeration Surveys

According to the Conference of European Statisticians (2007): *Principles and Recommendations for Population and Housing Censuses*, the post enumeration survey (PES) is a complete enumeration of a representative sample of a census population followed by matching each individual enumerated in the PES with information from the census enumeration. The results of the comparison are mainly used to measure coverage and content error in the context of the census. Some countries only confine the PES to evaluating coverage error (ONS, 2005). Coverage error refers to housing units and people missed in the census or those erroneously included. On the other hand, content errors evaluate the response quality of selected items. In general, an evaluation of the magnitude and direction of errors in a census is necessary in order to present to users the extent of reliability and accuracy of some characteristics reported. The evaluation, therefore, allows for a better interpretation of census results by presenting limitations to users by quantitatively evaluating the accuracy of census results with respect to coverage or/and quality of responses to questions on selected variables. For some countries the results of the PES are used to adjust census results if, for instance, there is evidence of major coverage errors (Hogan, 1992, 2001; Whitford & Banda, 2001).

2.3. Comparison with existing household surveys

Theoretically, any probability sample of households or individuals can be used to evaluate coverage and content errors in a census if they have some identical items with the same concepts and definitions. However, the post enumeration survey discussed above is specifically designed and most ideal to do so. In the absence of a post enumeration survey, other households survey results can be used to evaluate census results provided the principles of independence from the census and closeness to the census date are upheld. In addition, there should be sufficient identical information to perform accurate matching. For content evaluation, it is essential that several of the same data items are collected.

In Poland, comparisons with existing household sample surveys, such as the Household Budget Survey, the Labour Force Survey, and the Statistics of Income and Living Conditions survey (EU-SILC), to evaluate coverage and content errors in a census have not been used. Some indirect attempts have been tried for the census of agriculture (Bartosińska, 2006). The author thinks that the household sample surveys conducted in Poland may fulfil most of the above prerequisites, and be used to evaluate coverage and content errors. However, some research in this field is required.

2.4. Interpenetration studies on current census

This method involves drawing subsamples, selected in an identical manner, from the census frame, with each subsample capable of providing a valid estimate of the population parameter (UN Statistics Division, 2010). The assignment of census personnel (enumerators, coders, data entry staff, etc.) is also done randomly. This method helps to provide an appraisal of the quality of the census information, as the interpenetrating subsamples can, for example, be used to secure information on content error. In censuses and surveys, nonsampling errors, for instance, arise from differential interviewer bias, different methods of eliciting information, etc. After the subsamples have been enumerated by different groups of interviewers and processed by different teams of workers at the tabulation stage, comparison of estimates based on the subsamples provides a broad check on the quality of the census results. The results, from such studies, could be useful in improving the operations of future censuses and large-scale sample surveys. Until now, the interpenetrating studies have not been used in Poland (Kordos, 1987, 1988, 2007).

3. Some experience of other countries in census quality assessments

The author would like to refer to some experience in the census quality assessments of the U.S. Census Bureau, the Statistics Canada, the UK Office for National Statistics (ONS), and the Nordic Countries experience in register-based statistics and census of population, and comment of some Polish experience in this field.

3.1. Experience of the U.S. Census Bureau in census quality assessment[5]

The U.S. Census Bureau has used post-enumeration surveys with dual system estimation to measure coverage in the Decennial Censuses of Population and Housing since 1980. This approach involves case-by-case matching of persons in an independent survey with persons in the census to determine who was missed or counted in error. The post-enumeration survey-based coverage measurement program associated with the 1980 Census was called the Post-Enumeration Program (PEP); in the 1990 Census it was called the Post-Enumeration Survey (PES); in the Census 2000 it was called the Accuracy and Coverage Evaluation (A.C.E.); and for the 2010 Census it is called Census Coverage Measurement (CCM).

Coverage Measurement in the 2010 Census[6]. The primary goal of the 2010 CCM program was to measure coverage error in the 2010 Census such that this information could be used to improve the coverage of future censuses. As a result, the scope of coverage measurement was broader and the emphasis was different than it had been in the past. Specifically, the 2010 CCM goals were to:

1) produce measures of coverage error, including its components of omissions and erroneous enumerations;
2) produce these measures of coverage error not only for demographic groups and geographic areas, but also for key census operations; and
3) continue to provide measures of net coverage error.

National Research Council (2010), *Envisioning the 2020 Census*[7]: Planning for the 2020 census is already beginning. This book from the National Research Council examines several aspects of census planning, including questionnaire design, address updating, non-response follow-up, coverage follow-up, de-duplication of housing units and residents, editing and imputation procedures, and several other census operations. The book recommends that the Census Bureau overhaul its approach to research and development. The report urges the Bureau to set cost and quality goals for the 2020 and future censuses, improving efficiency by taking advantage of new technologies.

3.2. Experience of the Statistics Canada in Census Quality Assessment [8]

Conducted every five years, the Canadian Census of Population is a major undertaking whose planning and implementation spans a period of over eight years. Statistics Canada generally works on two and even three censuses at any given point in time. Before the final results of one census are out, planning and systems development are already well under way for the next one.

[5] https://www.census.gov/coverage_measurement/post-enumeration_surveys/.

[6] http://www.nap.edu/catalog/12524/coverage-measurement-in-the-2010-census.

[7] http://www.nap.edu/catalog/12865/envisioning-the-2020-census.

[8] http://unstats.un.org/unsd/dnss/docs-nqaf/Canada-12-539-x2009001-eng.pdf.

Consultations on the 2011 Census content were held in 2007 with a broad range of users, including key federal government departments, provinces and territories, local authorities, libraries, academia, the private sector, special interest groups and the general public. The Internet has become the primary vehicle for consultation material. Written submissions, dedicated meetings, conferences and working groups also continue to be highly effective ways to engage with users.

Statistics Canada with the 2011 Census continues to address both internal and external pressures to change its collection, processing and dissemination strategies. Using a number of pilot tests and phasing changes over the 2001, 2006 and 2011 Census cycles, the Agency moves from a decentralized, manually intensive collection and data entry operation to a more centralized and automated approach. This in particular addresses key concerns regarding confidentiality and security of personal census data. At the same time, a more proactive census communications and dissemination strategy has been adopted which has led to substantial increases to the amount of media coverage, to interest by a large number of censuses.

In view of these realities, Statistics Canada has a long tradition of providing guidance in its survey designs by consolidating its experience and conclusions about what constitute "best practices" into a set of Quality Guidelines. The first edition of Quality Guidelines appeared in 1985. Revised editions were released in 1987, 1998 and 2003. In keeping with the need to keep the guidelines evergreen, the present document, i.e. Statistics Canada (2009), has been significantly updated from the previous edition to reflect further advances in survey methodology over the past six years.

Data quality assessment provides an evaluation of the overall quality of census data. The results are used to inform users of the reliability of the data, to make improvements for the next census and, in the case of two coverage studies, to adjust the official population estimates. Quality assessment activities take place throughout the census process, beginning prior to data collection and ending after dissemination.

3.3. British One Number Census project[9]

The author would like to present a synthesis of the British One Number Census project as an example of correctly designed and implemented undertaking. Census 2001 results were the first to represent the entire population. This was achieved through a new strategy known as the *'One Number Census' (*ONC). One of the key elements of the ONC was an independent follow-up survey. The Census Coverage Survey (CCS), as it is known, involved face to face interviews with a sample of 320,000 households from every local authority in the UK. In the past, the total population given by the census was the raw count, reflecting a response rate of 98 per cent. But by combining the results of the census and the

[9] (ONS, 2005); One Number Census: Evaluation Report, Census 2001 Review and Evaluation, More available at: http://www.ons.gov.uk/ons/guide-method/census/census-2001/index.html.

CCS, it was possible in 2001 to estimate the total resident population - the 'one number' - to a high level of precision, plus or minus 0.2 per cent.

The Census Coverage Survey (CCS) was specifically designed to enable census population counts to be adjusted for under-enumeration at the national, local and small area level. It consisted of a completely independent and intensive face-to-face survey of a sample of over 16,000 postcodes containing 320,000 households drawn from all local authorities in England and Wales. The sample design took into account the uneven distribution of under-enumeration across the country by stratifying by a *'Hard to Count'* index based upon characteristics likely to be associated with under-enumeration, such as the number of multi-occupied addresses.

The CCS was operationally independent from the census enumeration exercise. The CCS sample postcodes were kept confidential, CCS interviewers did not have any sight of the address lists produced in carrying out the census, nor the census forms returned in the area in which they were interviewing. The interviewers focused on making as many calls as necessary to achieve an interview, and the timing of these calls was varied to maximise the probability of making contact.

The CCS in England and Wales achieved a response from 91 per cent of the households identified by interviewers. This is a high response rate for such a large-scale voluntary survey when compared to other the U.K. national surveys. The survey succeeded in meeting its objective of identifying households and persons that had been missed by the 2001 census.

For the ONC process to produce unbiased estimates of the population it is necessary for the census and Census Coverage Survey to be as independent of each other as possible. Practical arrangements were put in place to achieve this with census and CCS operations being kept entirely separate on the ground. If the two attempts at enumerating the same population are independent, it is possible to not only estimate those missed by either the census or CCS but to also estimate those missed by both - the dual system approach.

Through this approach, independence of the process was achieved. However, there is an additional component of dependence which needs to be taken into account. This is dependence caused by the fact that those people who are difficult to count in a census are also difficult to count in a post-enumeration survey such as the CCS. This was expected and a methodology was developed to identify those areas where dependency was marked and to adjust for that dependence. This added an additional 230,000 to the ONC population estimates for England and Wales as a whole.

3.4. The Nordic Countries Register-based statistics and Census of Population

The Nordic countries have a long tradition in using administrative registers in the production of official statistics. One of their common experiences is that the use of administrative records in censuses is the last step in a process that begins with producing statistics on different subject areas, depending on the type of registers available. By producing statistics on population or employment based on administrative data, they have learned about the influence the actual registers have on the quality of the statistics. After a period of testing and improvement, they realized that the quality of the administrative data was compatible with the quality recommended for censuses, and decided to also use the registers for census purposes. The time it takes from the establishment of an administrative register to possessing the quality data needed for census purposes may differ from one subject area to another. Nevertheless, the process is similar from subject area to subject area and from country to country. This also seems to be the case for establishing a statistical system based on register information.

In recent years, an increasing number of countries in the UNECE region have been considering the possibility of producing statistics based on administrative registers. Therefore, the National Statistical Institutes (NSIs) of the Nordic countries (Denmark, Finland, Iceland, Norway and Sweden) decided to share their experience and knowledge with the international statistical community, by producing comprehensive documentation of their best practices. The UN Economic Commission for Europe[10] published this document in 2007: The objective of this volume is to give strategic and planning officers in the NSIs and understanding of what register-based statistics are, covering also the necessary technical and administrative capacity, and the possible applications of these methods to produce official statistics. The emphasis of the publication is on the use of administrative registers to produce demographic and social statistics. In publishing the present volume, the United Nations Economic Commission for Europe (UNECE) would like to support the Nordic countries in sharing their experience in this field with the international statistical community at large. The volume represents a valuable tool for all NSIs (both within the UNECE region and outside it) that are planning to produce official statistics based on administrative registers. It also supported the implementation of the 2010 round of population and housing censuses, and preparation of the 2020 round of population and housing censuses.

It is necessary to stress that developing a register-based data system has been a step-by-step process in all Nordic countries over a rather long period. Statistical registers have been established in several areas and by 2011 all Nordic countries, at least according to national plans, will have a totally register-based population and housing census system. And even if we call it a census system, the same data sources are used in the corresponding subject matter statistics.

[10] Register-based statistics in the Nordic countries - Review of best practices with focus on population and social statistic: available at: http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf.

A register-based statistical system should never be regarded as completed once and for all. As new user needs arise and new administrative registers are established, new information should be integrated in the system.

What is a register-based census?

At one end of the line we have traditional censuses collecting data by the use of enumerators and questionnaires, using no register information at all. At the other end we have the totally register-based census. Some countries use mixed mode data collection with a combination of data from registers and questionnaires (as a total count or a sample survey). However, even countries conducting mainly traditional censuses may use register information to some extent, for instance an address list for persons or households for mailing out census forms.

For some countries, such as in the Nordic region, the goal has been to develop fully register-based censuses. For other countries, the goal may be to use registers to a certain extent, even if it is not possible or desirable to conduct a fully register-based census. And for some countries, a register-based census may not be an option at all.

What is required in order to call a census "register-based"? The main requirement must be the existence of a population register and a dwelling register. After all, resident persons and housing units are the basic units in a population and housing census. The link between persons and their dwellings is equally important, giving the household unit. These requirements constitute the cornerstones for developing a register-based census.

Register-based population and housing censuses system

The population and housing census provides the best example of the use of administrative records in statistical production. The Finnish system is used here as an example. The base registers cover the statistical units relevant to a census: *persons* resident in the country, the *buildings and dwellings* in the country as well as all *enterprises and their establishments* (business register). All statistical units can be linked to one another by means of the identification systems: *persons* can be linked to *families* and *households*, to the *dwelling* and *building* in which they live, and to the *employer* for whom they are working. Similarly, all units can be located on maps using geographical coordinates, because all buildings have been provided by coordinates.

3.5. Some Polish experience in census quality assessment

In Poland, demographic analyses of censuses evaluation have been widely used. There are a number of publications devoted to evaluation of census quality results using demographic methods (Gołata, 2009, 2010, 2012, 2014, 2016; Nowak, 2002, 2008; Paradysz, 2010, Strzelecki et al., 2002). The author will refer to some of them later, and begin with post enumeration survey (PES).

In Poland, first PES on sampling basis was applied in 1978 for the 1978 Population Census (Zasepa, 1993), in 1988 for the 1988 Population Census

(Nowak, 1998), the Micro-census 1995 and the 2002 General Population and Housing Census (GUS, 1996; Szabłowski et al., 1996).

In the 2002 Population and Housing Census the PES was conducted three weeks after the main census. A primary sampling unit was Census Enumeration Area (CEA). Out of all 177,591 CEAs for PES 903 CEAs were selected using stratified sampling designs by region with proportional allocation. Altogether 60,029 dwellings were selected and 27 census items were checked.

In the 2011 Population and Housing Census Sampling frame for the PES were sampling units selected for the sampling census, i.e. 2,744 thousand. For PES covered dwellings which had done self-enumeration, dwellings done by census enumerators or CATI, and not covered from different reasons. The PES enumeration was carried out from 1 April to 30th June. Additional condition was added that a dwelling should be contacted by telephone (50, 5%) . . . As by now no PES results have been officially published.

The author would like to refer here shortly to three papers: Gołata (2012, 2016), and Paradysz (2010), where some interesting aspects of census quality are discussed.

Gołata (2012) discusses the quality of population census data; in particular, the quality assessment of census data refers to the Polish experience. The author aims to propose an integrated approach to censuses, based on administrative records and sample surveys. She has stressed that despite a long tradition, as well as a well-developed research methodology, censuses do not provide ‘perfect’ results. First of all, the census as a comprehensive investigation may be burdened with different kinds of non-random errors. Taking into account all the different methods for conducting censuses, including the traditional method, using data from administrative records with sample surveys and the mixed method, different sources of error are analyzed.

The paper also attempts to define the census population in the light of international standards concerning usual residents, permanent residents and actual residents, with the resulting consequences. Sources of errors, methods of their identification and elimination are discussed. Particular attention is paid to coverage errors, which are illustrated by examples from the Polish censuses. The results of previous studies assessing the quality of administrative records, which were conducted in preparation for 2011 Population Census, are taken into account.

Evaluation of population censuses is traditionally carried out based on the results of post-census surveys. Polish experience in this field is rather limited. The discussion also includes harmonization problems, particularly the danger of divergent results and estimates, in terms of several data sources. In this regard, the British One Number Census project is mentioned (ONS, 2005). The idea of the project was to reduce the population underestimate and provide outputs which would be consistent in cross-section of territorial units at the regional and national level.

Gołata (2016): This is the best updated article on quality assessment of the last two population censuses in Poland conducted in 2002 and 2011. The article refers to the shift in methods to conduct a population census: from a conventional enumeration through a sample survey and a mixed approach to administrative data, as a new standard in statistics. The paper compares two Polish censuses of 2002 and 2011. It is aimed at quality assessment in the case of both: the traditional method (2002 census) and the combined approach (2011 census).

The quality of census data is discussed with essential aims and objectives to provide reliable information on the population age and sex structure in detailed territorial division. Therefore, quality assessment is provided for the whole country and at regional level. First of all, coverage errors are considered. She uses multiple sources of data and non-matching methods, in particular: demographic analysis based on previous censuses, vital statistics and a comparison with other existing sources. Different cross-sections according to sex, age and place of residence are considered. In each separate domain adequacy and divergence assessments are accompanied by an attempt to provide substantive explanations.

Paradysz (2010) has stressed that Poland prepares the new population census 2011 on the basis of a Virtual Census, using the administrative registers and combining them with surveys. This approach to modern census needs, however, modern indirect estimators. The paper also deals with nonresponse problems and other biases due to new approach. In the Polish 2011 census calibration and imputation methods are used. It also means indirect estimation on a different level of statistical aggregation of population. In conclusions he stressed that those systematic errors in population censuses are much higher that previously assumed. It concerns not only mobile persons (20 - 29 age group), but also much older persons. He concludes that post enumeration surveys had no important meaning since none of them did not show the most important, evident systematic errors, but no references are mentioned. He suggests that for next census administrative registers, and particularly a population register, should be used as well as indirect estimation. The author agrees with Paradysz’s conclusions, and would like to add that previously conducted post enumeration surveys in Poland were not correctly prepared and implemented if we take into account recommendations for the post enumeration surveys (UN Statistics Division, 2010). Some parts of these recommendations are presented later.

4. Overview of Post Enumeration Surveys

As it has been stressed earlier the primary objective of a census evaluation programme is to determine the sources and magnitude of coverage error and content error (for some selected variables). For many developing countries the post enumeration survey (PES) has become a reasonable independent evaluation programme. This is partly because other independent sources of data with relevant, comprehensive and reliable information are not available (Hogan, 1992; Hogan and Wolter, 1988, 1999; UN Statistics Division, 2010).

In conclusion, it should be pointed out that for many developing countries basic data, to facilitate census evaluation, are lacking or insufficient. For example, to undertake a demographic analysis there is a need for very reliable data to make it possible to calculate the demographic components of the population, namely fertility, mortality and migration. In some developing countries these data are not available. In addition, many developing countries do not have comprehensive vital registration systems; therefore, sophisticated demographic analysis to evaluate the census may not be feasible.

It should, however, be emphasized that the PES can only generate reliable and accurate results if the sample is well and efficiently designed, its implementation is of high standard, the matching exercise is carefully done, and the analysis of results and estimation are correctly executed (UN Statistics Division, 2010). Basic principles of planning and implementing of a PES, using UN Statistics Division extensively, are presented below[11].

4.1. Planning of a Post Enumeration Survey

"The planning of a PES should be preceded by a clear and unambiguous statement of objectives of the evaluation. Planning for a PES should, to the extent possible, be synchronized with the planning for the census. It should start early and adequate resources should be devoted to it as part of the overall census programme. The success of the PES depends mainly on the availability of qualified human and other adequate resources and it has to be independent from the census operations. In some cases, insufficient resources are at the disposal of PES planners to support its thorough conduct. Without adequate resources, the quality of the PES results would seriously be compromised. It is therefore necessary for the organization responsible for the conduct of the PES to develop a plausible survey plan with adequate budgetary and manpower requirements clearly spelled out.

Cost

The cost is a determining factor as to whether a PES should be undertaken or not. There is need to have adequate financial allotment to ensure availability of qualified enumerators and supervisors, competent matching clerks, qualified data processors, adequate training for all involved and effective operational and quality assurance in the whole PES process. The sample size in turn will depend on whether only national estimates are required. In this case the sample size can be relatively smaller compared to an overall sample size aimed at getting reliable results for many different domains. The latter will require independent estimates,

[11] Adapted from: UN Statistics Division (2010), Post Enumeration Surveys, Operational guidelines, Technical Report, New York, April 2010. Available at: http://unstats.un.org/unsd/demographic/standmeth/handbooks.

which can only be reliable if the sample size is reasonably large, implying that adequate sample sizes are obtained for each domain with its specified reliability levels. Common domains include rural/urban, regional provincial or other sub-regional domains. Large samples, for example, will demand the recruitment of a large pool of enumerators, supervisors, data entry clerks, etc.

Publicity

In order to encourage active participation in the PES by as many respondents as possible, it is advisable to plan for and mount publicity campaigns. Advance publicity is necessary because it prepares potential respondents for the PES by soliciting their cooperation. In this way response rates may be increased. Different approaches to publicity may be adopted depending on prevailing circumstances in different parts of a country.

4.2. Implementation of a Post Enumeration Survey

A number of factors contribute to errors in executing a PES, among them: use of faulty maps defining enumeration areas and unclear addresses especially in rural areas, poor publicity, shortage of transport facilities and limited communication during the data collection exercise, poor planning for data collection and data processing activities coupled with resource constraints.

Pilot test. A comprehensive test of all PES procedures is advisable. This can be a dress rehearsal of the actual PES just as the pilot census is a dress rehearsal of the census. The pilot test can cover some selected administrative divisions. The aim is to test the adequacy of the entire PES plan and the PES organization. The PES pilot test should preferably be undertaken in conditions similar to the actual enumeration. This implies that the Pilot PES should immediately follow the census pilot test. The purpose of the pilot PES is to prepare for the main PES; however, while it is not a source of usable substantive data, it provides lessons pertaining to the operational aspects of data collection that can be implemented in the current census. Ideally, the pilot should be taken a year before the actual PES just as the pilot census is taken one year before the planned census

Matching of records between a PES and the census is one the main features of the evaluation exercise. It should, however, be stated that it is one of the complex and challenging undertaking in a PES programme. It has to be done well for the PES results to be useful. The results of a pilot contribute to the establishment of matching rules, reconciliation procedures, and logistical flow of documents between the PES and the census. It may be possible to make broad estimates of precision and accuracy on the pilot PES results, such as sampling errors and certain bias components of the total mean square error.

Data collection. The method commonly used in data collection with respect to the PES is the personal interview method. The method entails enumerators going to households, in selected EAs, and interviewing the respondents, thereby collecting information by asking questions from the PES questionnaire. The main advantage of this approach is that the enumerator has the potential to ask probing questions. This is, in most cases, necessary in a PES. Additionally, enumerators are in an interactive mode with respondents such that they can explain to respondents the objectives of the PES when asked.

Questionnaire design. Questionnaires for the PES should be designed based on the final census questionnaire in order to facilitate an objective evaluation of the census. The PES questionnaire plays a central role in the survey process in which information is transferred from the respondents to the survey analysts. It is the vehicle through which the information needs of users are expressed in operational terms as well as the main basis of input into the data-processing system. It may be worth noting that if the questionnaire is to be used for recording responses by enumerators in the field, it should be sturdy enough to survive handling. It is also advisable that the questionnaire should be designed to facilitate the collection of accurate information.

A good questionnaire should have the following qualities:

(a) Enable the collection of accurate data to meet the needs of potential data users in a timely manner;

(b) Facilitate the work of data collection, data processing and tabulations;

(c) Ensure economy in data collection avoiding the collection of non-essential information;

(d) Permit comprehensive and meaningful analysis and purposeful utilization of collected data.

Selection and training of enumerators. As earlier stated, enumerators are at the interface with the respondents. Their work is critical to the success of the PES field work. In general, an enumerator should be able to effectively communicate with respondents and should have qualities needed to collect accurate information from respondents in a timely manner. An enumerator for the PES must have an adequate level of education and should be able to record information honestly. It is important that the selected enumerators should follow instructions and use definitions and concepts as provided in the enumerators’ manual.

The selected enumerators should be thoroughly trained before being assigned to do field work. It should be noted that the main objective of the training programme is to enhance uniformity and minimize measurement error, in interviewing procedures of the PES. Qualified instructors, who are well versed in the objectives of a PES, should be responsible for training. It is advisable that the trainers should be part of the PES planning and implementing team. In addition to the following lectures, trainees should take turns in explaining to others the various items in the questionnaire. In addition, practical sessions should be arranged both in the classroom environment and actual field situation. The training programme should result in decision by the PES director of which trainees may require additional training and whether any of them are entirely unsuited for the assignment.

The role of supervisors. It is recognized that training is a prerequisite of effective and successful PES field work. Notwithstanding the foregoing, training without proper supervision may not yield accurate results. In reality the success of PES fieldwork demands dedicated and effective supervision by supervisors that are supposed to be more experienced and better qualified than the enumerators. Like the enumerators, supervisors should undergo extensive training in all aspects of the PES. It should be underscored that a supervisor is an important link between the PES planners or management and the enumerator. The supervisor is supposed to organize work for enumerators, by determining field assignments. They review completed work and maintain a high level of commitment by enumerators to the PES programme. In order to achieve the above it is suggested that about five enumerators, at the maximum, should be assigned to one supervisor. A supervisor can play a central role in making follow-up of non-respondents. Non-response is a common phenomenon during a PES field work, just like in any other surveys. Because supervisors are supposed to have better qualifications and more experience than enumerators, there are therefore in the best position to contact the non-respondents and try to collect the requisite information.

Field data collection. The following are some of the socio-demographic variables included in a census questionnaire and repeated in the PES questionnaire for matching content error. The listed variables below are relatively easy to measure and are considered important demographic and social variables worth measuring response error if any. They include: a) age, b) sex, c) relationship to head of household or reference person, d) marital status, e) education level, f) type of housing unit.

4.3. Independence

The Dual System Estimation methodology is based on the assumption that the PES is an independent collection from the Census. The method is modelled on the technique of capture-recapture commonly used to estimate the population of wildlife. The methodology assumes a closed population. The assumption being that the population remains unchanged during the period of the study. Independence, therefore, requires that the PES must not be influenced by what took place in the census. For example, in theory, the frame should be independent from that of the census, the planning of a PES must be done by and independent group of people; in addition the PES implementation should be independent from that of the census. This implies that different enumerators and supervisors are supposed to be selected and deployed for the PES from the census. However, what is maintained is operational independence from the census at every stage of the PES such as enumeration, data processing and administering the survey. In order to maximize the independence between the two exercises, in some countries, the enumeration in the census is through the self-completed questionnaire, while the PES is conducted face-to-face and the frame for PES is independent from the census enumeration areas. In addition, the estimation procedure is based on the case-by-case matching of two different and independent sources describing the same event.

4.4. Control of non-sampling error

In all surveys there are bound to be non-sampling errors, the focus should be to minimize them. Both sampling and non-sampling errors should be controlled and reduced to a level at which their presence does not compromise the usefulness of the PES results. Non-sampling errors are particularly harmful when they are non-random, because they introduce bias in PES estimates. Such bias is complicated or difficult to measure. The best way to control non-sampling error is to follow the right procedures in all PES activities including planning, sample design, data collection, processing and analysing results. Emphasis is laid on careful and intensive training of field staff.

The following are some of the factors which contribute to nonsampling errors in PES**:**

(i) Vague objectives resulting in inadequate and/or inconsistent specifications with respect to objectives;

(ii) Duplication or omission of units in the PES due to imprecise definition of the boundaries of the EAs;

(iii) Inappropriate methods of interviewing, observation using ambiguous questionnaires, definitions or enumerator or supervisors instructions;

(iv) Lack of trained and experienced field interviewers including lack of good-quality field supervisors;

(v) Incomplete identification particulars of sampling units or faulty methods of interviewing;

(vi) Errors occurring in data processing;

(vii) PES respondent does not know census occupants or remembers incorrectly;

(viii) Adequate information for movers is inadequate or not available;

(ix) Other identifying information is poorly remembered for out-movers".

4.5. Challenges of Post Enumeration Surveys

There are a number of problems which constrain some countries from carrying out post enumeration surveys. UN Statistics Division (2010) briefly presents below some of the key problems and suggested solutions:

(i) "In some developing countries there is lack of technical personnel with requisite skills and experience in survey methodology and in designing and implementing the whole PES process. These challenges are in areas such as sample design, implementation, matching and estimation. However, it can be argued that if a country can design and conduct good and efficient household surveys, the chances of conducting a good post-enumeration survey are high. There is therefore need for countries to develop and maintain capacity in sample survey methodology and implementation;

(ii) Lack of financial resources after investing so much in a census. It is against this background that it is advisable that, with respect to financial resources for the PES, the planning should be an integral part of the overall census Programme;

(iii) It is not possible to maintain the theoretical independence required between the census and the post enumeration survey. The commonly adopted strategy is to devise practical or operational approaches of maintaining independence. For example, planning and management of a post enumeration survey has to be undertaken by personnel that is separate from census personnel. In practice, independence is maintained by putting into place seemingly independent field procedures that are implemented to try and improve the enumeration in the post enumeration survey compared to the census count. These include:

   (a) the use of enumerators and supervisors who are better qualified than those used in the census or the best staff used in the census;

   (b) assigning the post enumeration survey staff to areas in which they did not work during the census;

   (c) using a questionnaire which asks more detailed and probing questions on selected characteristics;

   (d) census results for designated areas should not be known by staff who are assigned to those areas;

   (e) all census materials from the selected PES areas should be collected before PES enumerators go into the field;

(iv) The design of the survey, matching and estimation procedures may be perceived to be complex. These problems can be solved or mitigated by having good sample survey methodologists and analysts including employing qualified and well-trained enumerators and matching clerks and supervisors;

(v) The post enumeration interview can be demanding. It usually incorporates questions to determine if a respondent should really be counted at the residence in question. In addition, the post enumeration interview takes place after the census interview, at which point the respondent may feel overburdened and not be as forthcoming with accurate information.

(vi) In some countries, census planners feel that it is enough to put in place good quality assurance procedures at various stages of a census. The truth is that a census is a massive operation such that despite the assurance procedures put in place, error is bound to creep in. It is against this background that an evaluation of census results is still necessary despite the quality assurance procedures having been implemented in a census;

(vii) Some countries are ambivalent about conducting a post enumeration survey after the fatigue resulting from the census operation. A census being a difficult and taxing operation, which saps the energy of those associated with it, discourages some national statistical/census offices to conduct a PES. Others feel that the exposure of discrepancies between the census and post enumeration survey results to users would be detrimental to the reputation of the census or statistical organization. With respect to fatigue this is the reason why (UNSD, 2010) advocates that the PES should be as independent, to the extent possible, from the census activities. So that a different team would be responsible for the PES.

(viii) In the case of the ambivalence of having evaluation figures for the census, it should be recognised that it is common in all credible statistical studies to have quantitative measures of error. PES evaluation results, therefore, would enhance confidence among informed ***users*** of census data, contrary to the negative view stated above. The evaluation would not diminish the importance of the census as long as users understand the limitations of the data and errors do not affect the major uses of the data."

It should further be emphasized that an evaluation of a census is necessary for a number of reasons, among them is the fact that the population census is the most extensive and expensive data collection exercise for many countries. In addition, censuses have in recent years become complex. With vast amounts of resources spent, there is usually considerable pressure on census takers to ensure that census results are accurate to facilitate informed decision making at national and other domain levels. In addition, because of the massive nature of the census operation, it is inevitable that some inaccuracies such as errors of coverage and content/responses are unavoidable.

5. Strengths and weaknesses of evaluation methods

Methods based on a single source of data provide less insight into the magnitude and types of errors in the census data than the methods based on comparison of two or more sources of data. Examples are age and sex distribution analyses, which provide a general impression of the quality of the census results, but provide little insight on relative contributions to coverage and content error. The advantage is that such methods do not require additional data to be collected for evaluation purposes and, in general, there is no need for sophisticated matching operations. Such methods can, however, complement other methods of evaluation such as post enumeration surveys.

Demographic analysis has the advantage that no additional data is needed to be collected to perform this analysis. Information is already available, therefore, it is less costly and where the national statistical/census office has demographers there may be no need for additional staff to carry out the technical analysis. The limitation of single data source methods is that they provide less insight into the type and magnitude of errors present in the sources of data.

Results of well-designed and implemented interpenetrating subsamples can give good insights into different contributions of component errors to total error. This type of evaluation helps in the identification of operational stages that contribute to census error. However, such studies are relatively costly involving many field staff, intensive training and close supervision.

The PES is an independent evaluation method of a census. It demands adequate financial, human and other resources. A successful PES calls for a good sample design and survey implementation. Mention should be made that the matching exercise can be somewhat complex. The major advantage of matching over non-matching studies (an analysis that does not require matching censuses records with another source) centres on their ability to provide separate estimates of coverage and content error. On the other hand, the non-matching studies, because they review census results at the aggregate rather than unit level, for example, housing units, households or persons, provide only estimates of net census error. The characteristics that can be evaluated from matching studies are much more than those for non-matching studies, which are usually limited to age and sex distributions. Matching evaluation methods, however, require high level technical skills, managerial and financial resources.

In summary, since the demographic analyses are, in general, undertaken irrespective of a PES being conducted, the critical decision as to whether or not to conduct a PES lies in the quality and variety of demographic data available. Accurate data on fertility, mortality and migration levels and trends are needed.

Demographic analysis often depends on previous census data which may also be flawed. In such situations, therefore, the PES approach, though relatively complex, may be the only reliable way of evaluating census error. This, however, does not preclude complementing the PES with demographic analysis approaches in situations where required data are available.

6. Concluding remarks

The author has outlined different methods of quality assessment of population census data, using mainly UN Statistics Division (2010) and National Research Council (2009) recommendations. He has added also UN (2015) Recommendations for the 2020 Censuses of Population and Housing, accepted by the Conference of European Statisticians. He had in mind the Polish experience in this field and preparation of the Polish 2021 Census of Population and Housing. Some of the methods presented here were used in the last Polish population censuses, such as *demographic analysis* and *post enumeration survey, small area estimation methods* and *imputation* (Gołata, 2012, 2016; Paradysz, 2010), but final solutions are not yet known. Nevertheless, he would like to draw attention to the following aspects connected with census quality assessment:

1. ***Different modes of data collection*** were used in the 2011 Polish Census of Population, such as: traditional census using electronic questionnaire and a terminal of hand-held type, CAPI - Computer Assisted Personal Interview, Self-registration by Internet, CAII - Computer Assisted Internet Interview and CATI - Computer Assisted Telephone Interview. Each of these modes is the source of some kind of errors. They should be verified by the 2021 Polish Pilot Census.

2. ***Basic data are to be collected on sampling basis***. On average, 20 percent of dwelling units have been selected in each poviat (county), but sampling fractions were different in each county: the smaller numbers of dwelling units in the county, the higher sampling fraction were selected. Sampling allocation in different strata should be verified for the next census.

3. ***Demographic analysis*** has the advantage of no additional data needed to be collected to perform this analysis. Information is already available, therefore, it is less costly and where the national statistical office has demographers, as it is in Poland, there is no need for additional staff to carry out the technical analysis. The limitation of single data source methods is that they provide less insight into the type and magnitude of errors present in the different sources of data. Demographic analysis often depends on previous census data which may also be defective. In such situations, therefore, the PES approach, though relatively complex, may be the only reliable way of evaluating census error. In summary, since the demographic analyses are, in general, undertaken irrespective of a PES being conducted, the critical decision as to whether or not to conduct a PES lies in the quality and variety of demographic data available. The GUS should give priority to research on improving demographic analysis in the four areas: (i) improving the measurement of undocumented and documented immigrants, (ii) development of sub-national geographic estimates, (iii) assessment of the uncertainty of estimates from demographic analysis, and (iv) refining methods for combining estimates from demographic analysis and post enumeration survey data.

4. ***A post-enumeration survey*** is worth conducting if it is carefully planned and functioning within operational and statistical constraints. Cooperation of the different kind of experts involved in preparation, implementation, processing and publication of a population census is very important for the quality of census results. The dual system estimation methodology, which is key to the PES philosophy, assumes *independence* between *the census* and *the PES*. However, it should be noted that conducting a PES is demanding in terms of planning, sampling design, data collection and supervision, matching of PES and census results. The prerequisites for a successful PES are having adequate resources; qualified enumerators and supervisors; good survey statisticians and analysts; and efficient and careful implementation of all the activities related to the survey. ***If these conditions are not fulfilled, it is better to stop conducting PES*.**

5. ***Application of small area estimation methods.*** To increase the precision of estimates obtained on sampling basis, small area estimation methods may be applied if good additional data are available. Different kinds of registers are to be used in the census if the qualities of these registers are verified. How the registers are to be used is still not clear. It seems that in Poland we have some experience in this field but regarding the last population census, it is necessary to show how effective the methods of small area estimation were for obtaining data from sample surveys and available statistical sources.

6. ***Permanent data quality improvement.*** The author thinks that a PES proposed in the census is not efficient and will produce the same kind of data as post enumeration surveys in years 1978, 1988 and 2002 (Kordos, 2007). There were problems with independent matching and double system of estimation. He suggests conducting post enumeration survey after the main census using methods presented in (UN Statistics Division, 2010) adjusted to the Polish conditions or stop conducing PES.

7. ***Preparation for the Polish 2021 Census of Population and Housing:*** Preparation for the 2021 Polish Census and Housing has been started, but first of all, the assessment of the quality of the last census should be completed. All obtained results of analysis should be published, giving opportunity of other experts for contributions in this field.

REFERENCES

BARTOSIŃSKA, D., (2006). Attempts at Applying Small Area Estimation Methods in Agricultural Sample Surveys in Poland, Statistics in Transition, Vol. 7, No. 6, December 2006, pp. 1203–1218.

DAUPHIN, M., CANAMUCIO, A., (1993). Design and implementation of post-enumeration survey: developing country example. Washington: International Statistical Programs Center, US Bureau of the Census.

EUROSTAT, (2007). Handbook on Data Quality Assessment: Methods and Tools.

EUROSTAT, (2009). Handbook on Quality Reports.

GOŁATA, E., (2009). Economic activity in population census 2011 and administration resources. In: Gołata, E. (Ed.), Methods and Sources of Obtaining Information in Public Statistics, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. Poznań.

GOŁATA, E., (2010). Indirect Estimation of Economic Activity for the Register-based Census. In: Gołata, E. (Ed.), Measurement and Information in Economy, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań, pp. 85–104 (in Polish).

GOŁATA, E., (2012). Population Census and Truth, Studia Demograficzne, No.1, 161, pp. 23–55 (in Polish).

GOŁATA, E., (2014). Census quality of the new generation, Wiadomości Statystyczne, No. 5, pp. 26–53. (in Polish).

GOŁATA, E., (2016). Shift in Methodology and Population Census Quality, Statistics in Transition-new series, December 2016, Vol. 17, No. 4, pp. 1–14.

HOGAN, H., (1992). The 1990 Post-Enumeration Survey: An Overview, The American Statistician, Vol. 46, No. 4, pp. 261–269.

HOGAN, H., (2003). The Accuracy and Coverage Evaluation: Theory and Design. Survey Methodology, 29 (2): 129–138.

HOGAN, H., WOLTER, K. (1988). Measuring Accuracy in a Post-Enumeration Survey, Survey Methodology, Vol. 14, No. 1, pp. 99–116.

KORDOS, J., (1987). Data Accuracy in Social Surveys (in Polish). BWS, GUS, t. 35, Warszawa.

KORDOS, J., (1988). Quality of Statistical Data (in Polish), PWE, Warsaw , pp. 244.

KORDOS, J., (2007). Some Aspects of Post-Enumeration Surveys in Poland, Statistics in Transition-new series, December 2007, Vol. 8, No. 3, pp. 563–576.

NATIONAL RESEARCH COUNCIL, (2009). Coverage Measurement in the 2010 Census. Panel on Correlation Bias and Coverage Measurement in the 2010 Decennial Census, Robert M. Bell and Michael L. Cohen (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press: http://www.nap.edu/catalog/12524/coverage-measurement-in-the-2010-census.

NATIONAL RESEARCH COUNCIL, (2010). Envisioning the 2020 Census, 2010. http://www.nap.edu/catalog/12865/envisioning-the-2020-census.

NOWAK, L., (1998). Quality of Census Data, In: Tendencies of Changes in Structure of Population, Households and Families in 1998-1995 GUS, Warsaw, pp. 22–31 (in Polish).

NOWAK, L., (2008). The 2011 Census of Population and Housing – Methodology and Topics, the Committee of Demographic Research of the Polish Academy of Sciences (in Polish). http://www.knd.pan.pl/images/stories/pliki/pdf/Nowak_28_luty_2008.pdf

ONS, (2005). One Number Census: Evaluation Report, Census 2001 Review and Evaluation, http://www.ons.gov.uk/ons/guide-method/census/census-2001/index.html.

PARADYSZ, J., (1989). On non-sampling errors in women's fertility survey in the 1970 National Population Census. In: Problems of Statistical Surveys by Sampling. GUS, Warszawa, BWS, Vol. 36, pp. 154–159 (in Polish).

PARADYSZ, J., (2010). Necessity of Indirect Estimation in National Census. In: Gołata, E. (Ed.); Measurement and Information in Economy, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. Poznań, pp. 45 – 66 (in Polish).

STATISTICS CANADA, (2009), Quality Guidelines, Fifth Edition – October 2009; http://unstats.un.org/unsd/dnss/docs-nqaf/Canada-12-539-x2009001-eng.pdf.

STATISTICS FINLAND, (2004). Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland, Tilastokeskus, Statistikcentralen, Statistics Finland, Helsinki.

SZABLOWSKI, J., WESOLOWSKI, J., WIECZORKOWSKI, R., (1996). Index of Fitting as a Measure of Data Quality — on basis of the Post-enumeration Survey of Micro-census 1995, Wiadomości Statystyczne, No. 4. pp. 43–49 (in Polish).

UN, (2007). UN Principles and Recommendations for Population and Housing Censuses: http://unstats.un.org/unsd/demographic/sources/census/docs/P&R2_February%2012%202007.pdf

UN, (2010). Post Enumeration Surveys Operational Guidelines. Technical Report. United Nations Statistics Division New York. Available at: http://unstats.un.org/unsd/demographic/standmeth/handbooks/.

UN, Economic Commission for Europe, (2011). Using Administrative and Secondary Sources for Official Statistic, A Handbook of Principles and Practices, New York and Geneva.

http://www.bing.com/search?ei=UTF&pc=AV01&q=UN+Economic+Commission+for+Europe+%282011%29+%2CUsing+Administrative+and+Secondary+Sources+for+Official+Statistic%2C+A+Handbook+of+Principles+and+Practices%2C+New+York+and+Geneva.+&FROM=AVASDF.

UN, Economic Commission for Europe, (2015). Recommendations for the 2020 Censuses of Population and Housing, New York, Geneva. http://unstats.un.org/unsd/demographic/sources/census/census3.htm.

WHITFORD, D. C., BANDA, J., (2001). Post-enumeration Surveys: are they worth it? Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects, UN Statistics Division, New York, 7-10 August 2001.

WOLTMAN, H., ALBERTTI, N., MORIARITY, C., (1988). Sample Design for the 1990 Census Post-Enumeration Survey, Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 529–533.

ZASEPA, R., (1993). Use of Sampling Methods in Population Censuses in Poland, Statistics in Transition, Vol. 1, No. 1, pp. 69–78.

ZHANG, L.-C, (2011). A Unit-Error Theory for Register-Based Household Statistics, „Journal of Official Statistics", 27 (3), pp. 415–432.

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 139–150

THE ACHIEVEMENTS OF STUDENTS AT THE STAGES OF EDUCATION FROM THE SECOND TO FOURTH USING FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

Mirosława Sztemberg-Lewandowska[1]

ABSTRACT

Significant demographic phenomena can be observed in Poland – the number of school age population is decreasing. It affects higher education since the immediate effect of demographic changes is the drop in the number of students. The analysis of the level of future students' knowledge also remains an important aspect of the problem.

The purpose of the article is to compare the level of knowledge presented by students at the subsequent stages of education in the period 2009-2015. The research covers the average exam results received on graduation from the second, third and fourth stage of education. Functional principal component analysis, which is based on functional data, will be applied in the study. This method allows an analysis of dynamic data.

Key words: level of knowledge, functional data, functional principal component analysis, longitudinal data.

1. Introduction

Demographic conditions exert a direct impact on higher education. Since 1990 Polish higher education has been going through a period of continuous and dynamic growth resulting from the population boom, lasting for almost 15 years. Starting from 2006, the first symptoms of this trend collapse became noticeable as the number of students was gradually decreasing. A drop in the number of 10-year-olds was recorded in each consecutive year and thus the reversal of the trend favourable for Polish higher education. This process has been continuing until today causing the ongoing decrease in the number of the traditional college-age population.

Figure 1 presents the number of students graduating from the subsequent stages of education. The students who graduated from primary school (PS) in

[1] Wrocław University of Economics. E-mail: miroslawa.sztemberg-lewandowska@ue.wroc.pl.

2003 are marked with the letter A on x-axis, in 2006 - middle school and 2009 - secondary school, i.e. theoretically those born in the same year. It can be observed that slight local extremes of the PS students correspond to larger fluctuations in the number of secondary school students. The number of PS graduates reached the minimum (M) level in 2006 and of secondary school graduates in 2012, whereas a year later – the local maximum was achieved. The number of PS graduates has been stabilizing since 2014, i.e. the students who will graduate from the middle school (MS) in 2017 and the secondary school in 2020.

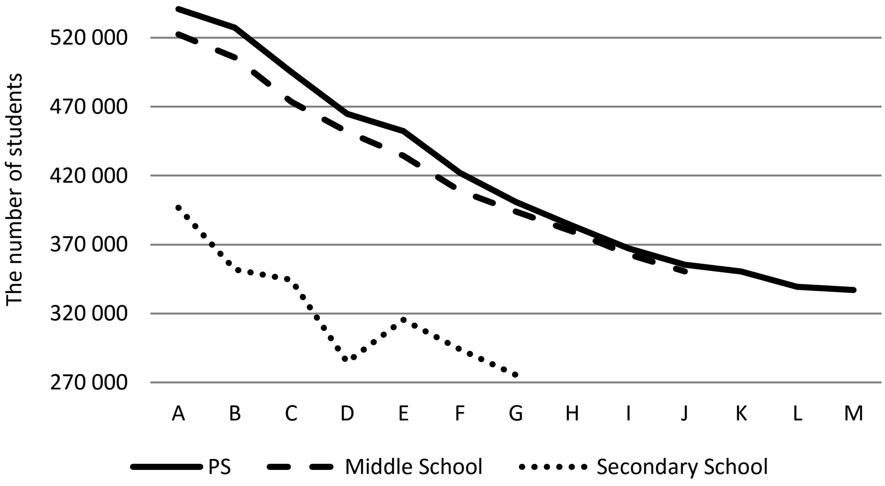


Figure 1. The number of students graduating from the subsequent stages of education

Source: author's compilation based on CKE [Central Examination Commission] data.

A	PS 2003	M 2006	S 2009
B	PS 2004	M 2007	S 2010
C	PS 2005	M 2008	S 2011
D	PS 2006	M 2009	S 2012
E	PS 2007	M 2010	S 2013
F	PS 2008	M 2011	S 2014
G	PS 2009	M 2012	S 2015
H	PS 2010	M 2013	
I	PS 2011	M 2014	
J	PS 2012	M 2015	
K	PS 2013		
L	PS 2014		
M	PS 2015		

The level of knowledge presented by future students is another important aspect affecting higher education. For this reason, the level of knowledge presented by students at the consecutive stages of education in the period 2009-2015 was compared. The research covers average exam results received on graduating from the second, third and fourth stages of education. Functional principal component analysis was applied in the study.

2. Methodology

Principal component analysis (PCA) is based on the transformation of original variables into the set of new and mutually orthogonal variables referred to as principal components [Harman 1975]. Functional principal component analysis (FPCA) is characterized by the advantages of a classical principal component analysis and, moreover, allows for the analysis of dynamic data. The type of data is the basic difference between these two methods: PCA is based on multivariate data, whereas FPCA is based on functional data. Functional data take the form of curves and trajectories, i.e. the sequence of individual observations rather than just a single observation [Hall and Hosseini-Nasab 2006, Krzyśko et al. 2012].

In the case of the functional principal component analysis (FPCA) each principal component is presented as the principal component weight function, also referred to as the time dependent eigenfunction $\xi_j(t)$ [Daniele 2006, Ramsay and Silverman 2005]. Eigenfunction maximizes the principal component function variance:

$$v(t,s) \overset{def}{=} \frac{1}{n-1}\sum_{i=1}^{n}[x_i(t)-\bar{x}(t)][x_i(s)-\bar{x}(s)] \tag{1}$$

Similarly to the classical PCA, in the case of the functional one the problem is the function variance distribution:

$$v(t,s)=\sum_j \lambda_j \xi_j(t)\xi_j(s) \tag{2}$$

where $\lambda_j, \xi_j(t)$ satisfy eigenequation:

$$\langle v(u,),\xi_j\rangle = \lambda_j \xi_j(s). \tag{3}$$

and eigenvalues are positive and non-decreasing:

$$\lambda_j \overset{def}{=} \int_T \xi_j(t)v(t,s)\xi_j(s)dtds. \tag{4}$$

Eigenfunctions satisfy the condition:

$$\int_T \xi_j^2(t)dt=1 \quad \text{and} \quad \int_T \xi_j(t)\xi_i(t)dt=0 \quad (i<j). \tag{5}$$

Eigenfunctions define the principal components of variation between the sampling functions x_i [Ingrassia and Costanzo 2005, Hall et al. 2006].

3. The achievements of students

3.1. The second stage of education

Since April 2002 the exam taken by students graduating from the sixth grade of primary school has been held. The standards of examination requirements constitute the basis for carrying out the final test. Until 2014 the standards were grouped into five trans-subject categories:

- reading,
- writing,
- reasoning,
- use of information,
- practical application of knowledge.

In 2015 the test consisted of two parts:

- the first part – covers tasks in Polish and maths,
- the second part – covers tasks in a modern foreign language.

A sixth grader takes the test in one of the following foreign languages: English, French, Spanish, German, Russian and Italian. A student can choose only the foreign language learned at school as an obligatory subject.

Figure 2 illustrates the average percentage results in particular subjects received by the sixth graders from their graduation test in the period 2002-2014.

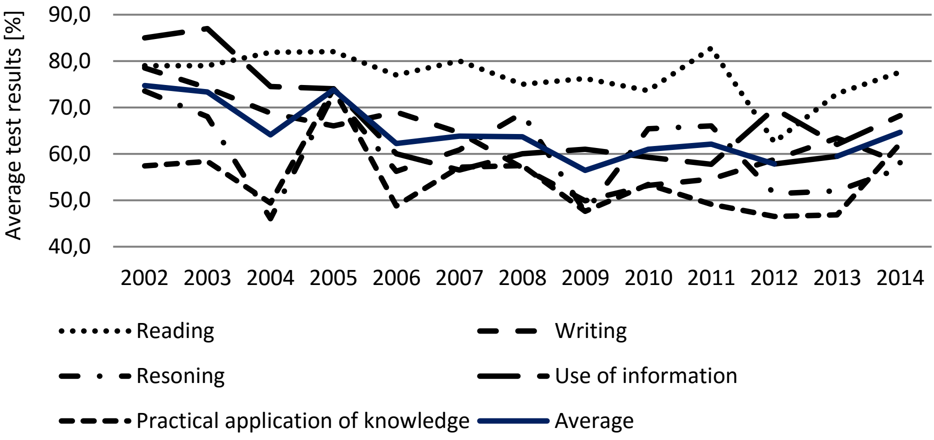
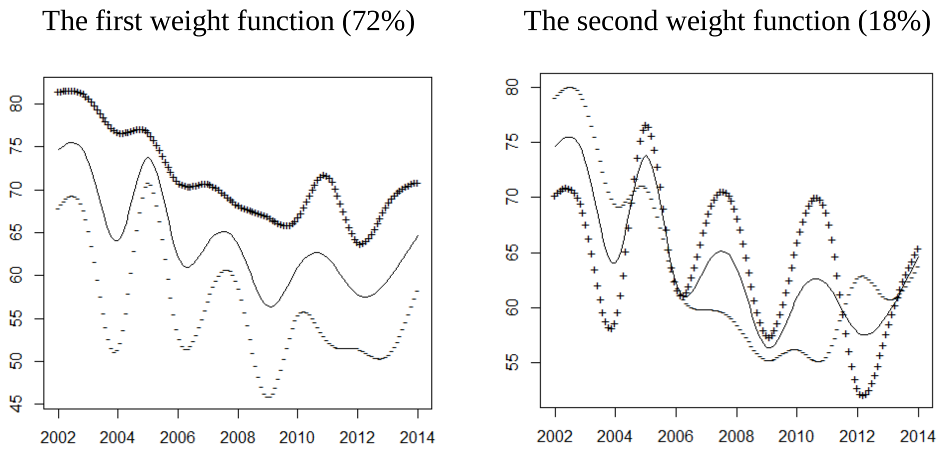


Figure 2. Average test results achieved on graduating from the sixth grade of PS in particular subjects [%]

Source: author's compilation based on CKE [Central Examination Commission] data.

The average result in all subjects is denoted by a solid line and presents the decreasing tendency in the period 2002-2014.

The functional principal component analysis allowed for distinguishing two component functions. The practical explanation of the functional principal components is supported by the graphs showing each component deviation from the average in all subjects (Fig. 3).


The first weight function (72%)

The second weight function (18%)

Figure 3. Weight functions

Source: author's compilation using R program.

The first functional principal component explains 72% of the joint variation, whereas the second one 18%. The first component reflects the overall tendency. The plus sign on this component means that the curve describing the result in a particular subject remains above the average. The second component shows the tendency in both outlier and mid years against the average. The plus sign on the second component means that the test result in a given subject before 2005 and after 2011 was below average, whereas in the years 2005-2011 the result was above average.

Based on the results of the functional factor analysis, data visualization and a comparison of analysed objects can be performed. Figure 4 presents data projection on the plane defined by two functional principal components.

Students received the best results in reading – above average. Writing and using information remained on an average level, the situation was better before 2005 and after 2011. Both practical application of knowledge and reasoning were below average.

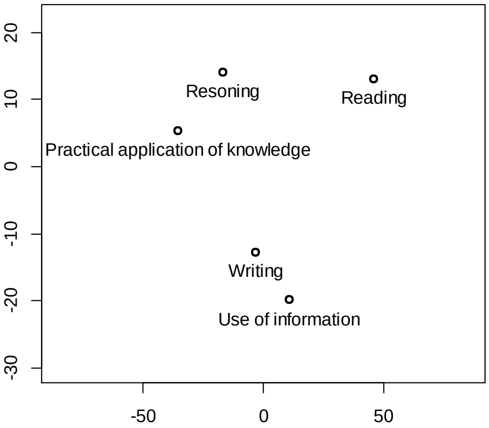


Figure 4. Objects in the components' space

Source: author's compilation using R program.

3.2. The third stage of education

The exam for the third graders of middle school covers the acquired knowledge and skills specified in the core curriculum of general education for the selected subjects taught at the third and earlier stages of education.

Before 2009 middle school students took the exam in humanities as well as maths and natural science only. In the period 2009-2011 middle school graduation exam consisted of three parts:

- humanities,
- maths and natural science,
- modern foreign language.

In the years 2012-2015 the following scopes were identified within the framework of each exam part:

- humanities covering history and social studies and also Polish,
- maths and natural science covering subjects teaching natural sciences and maths,
- modern foreign language at either basic or extended level.

A middle school student takes an exam in one of the following foreign languages: English, French, Spanish, German, Russian, Ukrainian and Italian. A student can choose only the foreign language learned at school as an obligatory subject.

Every middle school student is obliged to take an exam in a modern foreign language at the basic level. An extended level exam is obligatory only for those students who choose to take an exam in the language they used to learn in a primary school. The other middle school students can also take it if they wish to check the level of their language skills.

The exam has a written form. Taking it is the condition to graduate from the middle school, however, the minimum result to be achieved by a student is not defined and, therefore, it is not possible to fail the exam.

Figure 5 presents the average percentage test results in particular subjects taken after the third grade of middle school in the years 2006-2015. The average result in all subjects is denoted by a solid line and shows the decreasing tendency in the period 2006-2015.

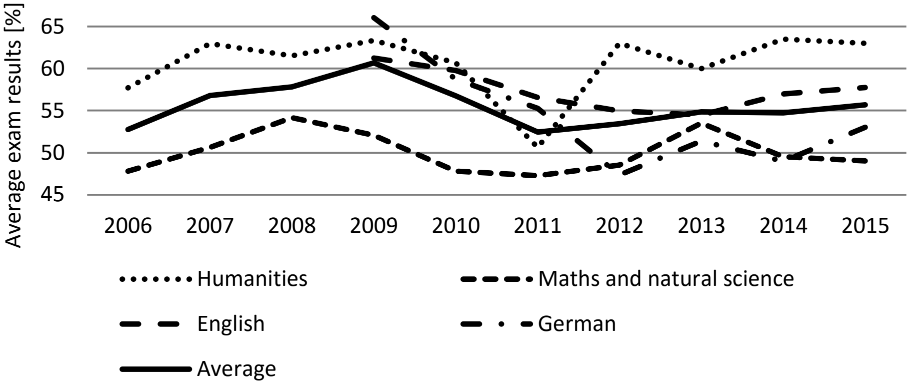


Figure 5. Average exam results achieved on graduating from the third class of middle school in particular subjects [%]

Source: *author's compilation based on CKE [Central Examination Commission] data.*

Two component functions were distinguished by means of the functional principal component analysis. Fig. 6 presents the graphs of each component deviation from the average in all subjects.

The first weight function (74%) The second weight function (24%)

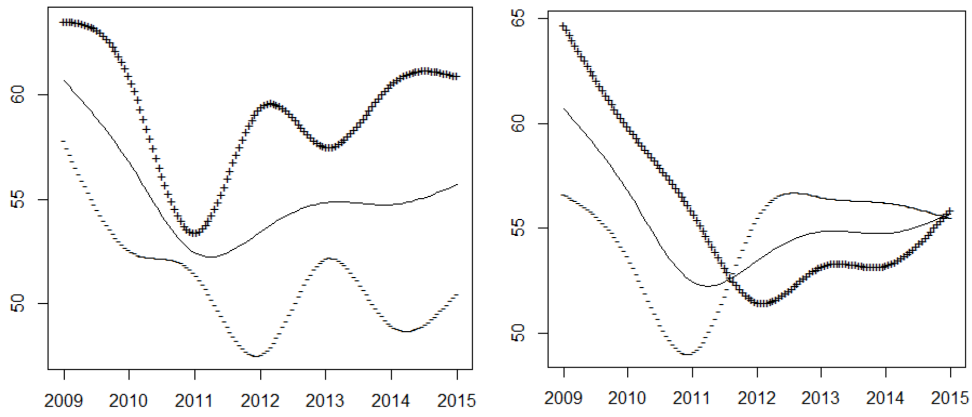


Figure 6. FPCA weight functions

Source: author's compilation using R program.

The first functional principal component explains 74% of the joint variation, whereas the second one 24%. The first component is responsible for the overall tendency. The plus sign of this component means that the curve describing the result in a particular subject remains above the average. The second component shows the tendency in the initial and final years against the average ("the beginning vs. the end") and compares the period until 2011 and after 2012 against the average result. The plus sign on the second component means that the test result in a given subject at the beginning of the analysed period was higher than the average, whereas in the years 2012-2014 the result was below the average.

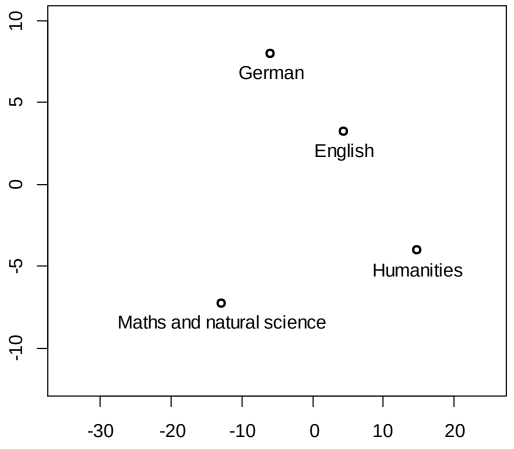


Figure 7. Objects in the components' space

Source: author's compilation using R program.

The visualisation of objects in the space of the component function (fig. 7) allows for drawing the following conclusions:

- humanities: the test result below the average, the situation was worse at the beginning of the analyzed period than in the years 2012-2014,
- English and German: average test result, higher than average at the beginning of the studied period, after 2012 – lower than average,
- maths and natural science: test result lower than average, at the beginning of the analysed period the situation was worse than after 2012.

3.3.The fourth stage of education

A graduate taking "the old type of graduation exam" (before 2015) is obligated to take: two exams in the oral part and three exams in the written part. The obligatory exams in the oral part are as follows:

- an exam in Polish (without defining the level),
- an exam in a modern foreign language (without defining the level).

The obligatory exams in the written part are as follows:

- an exam in Polish (basic level),
- an exam in maths (basic level),
- an exam in a modern foreign language (basic level).

A graduate taking “the new type of graduation exam” (from 2015) in the written part is obliged to take an exam in the chosen additional subject (extended level).

In order to receive the graduation diploma a student has to get at least 30% points at the exam in each obligatory subject in the oral part and receive at least 30% points at the exam in each obligatory subject in the written part.

Figure 8 presents the number of students who passed / failed the secondary school graduation exam in the period 2009-2015. It is noticeable that the number of students taking the graduation exam is decreasing each consecutive year and, moreover, the number of those who failed this exam remains disturbingly high in the recent two years.

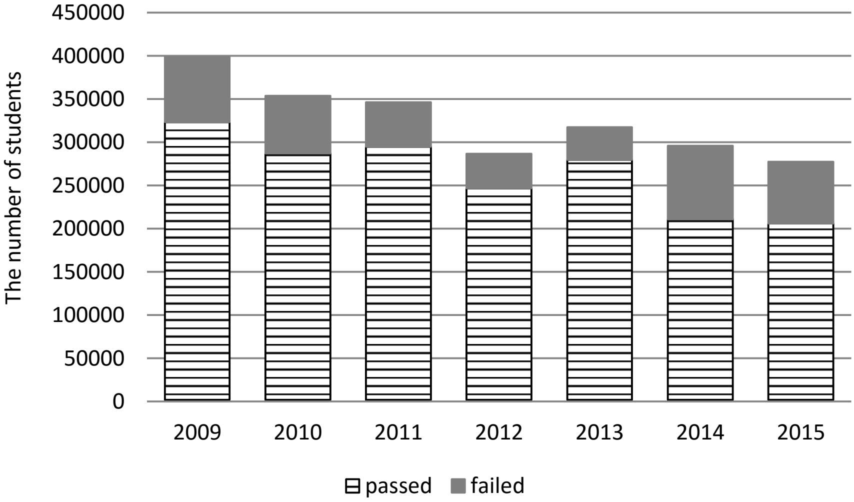


Figure 8. The number of students taking graduation exam

Source: author’s compilation based on CKE [Central Examination Commission] data.

Figure 9 presents the average percentage results obtained at the secondary school graduation exam in particular subjects in the period 2008-2015. The bold line shows an average result in all subjects – a decreasing tendency is noticeable.

The functional principal component analysis was applied to distinguish two component functions. The graphs of each component deviation from the average in all subjects are presented in Figure 10.

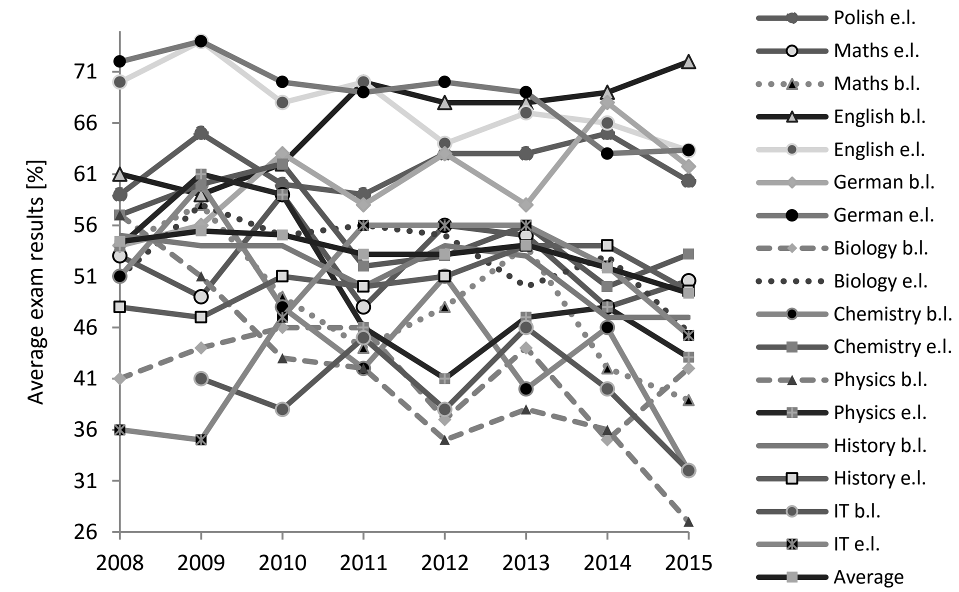


Figure 9. Average results of the secondary school graduation exam

Source: author's compilation based on CKE [Central Examination Commission] data.

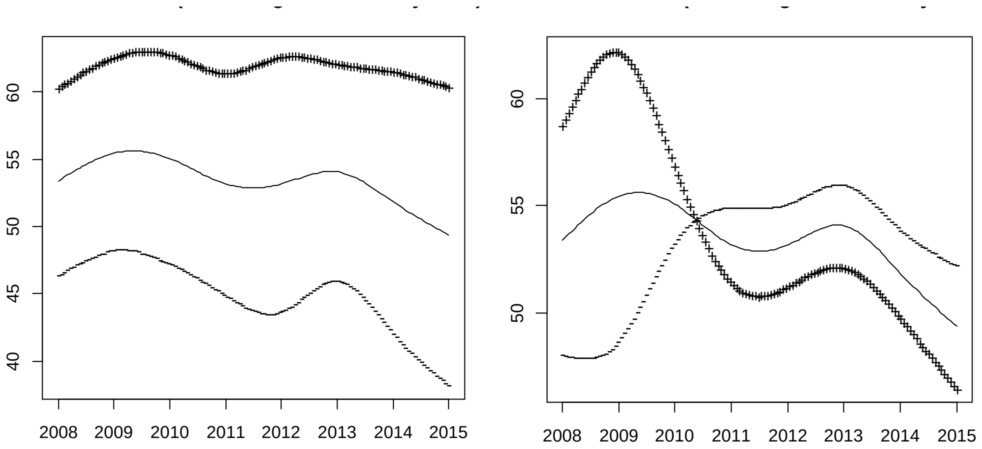

The first component function (80%) The second component function (15%)

Figure 10. FPCA components

Source: author's compilation using R program.

The first functional principal component explains 80% of the joint variation, whereas the second one 15%. The first component is responsible for the overall tendency. The plus sign of this component means that the curve describing the result in a particular subject remains above the average. The second component shows the tendency in the initial and final years against the average ("the beginning vs. the end") and compares the period until 2010 and after 2010 against the average result. The plus sign on the second component means that the secondary school graduation exam result in a given subject was higher than the average at the beginning of the analysed period, whereas at the end the result was worse than the average.

Next the data were projected on the plane determined by two functional principal components (Fig. 11).

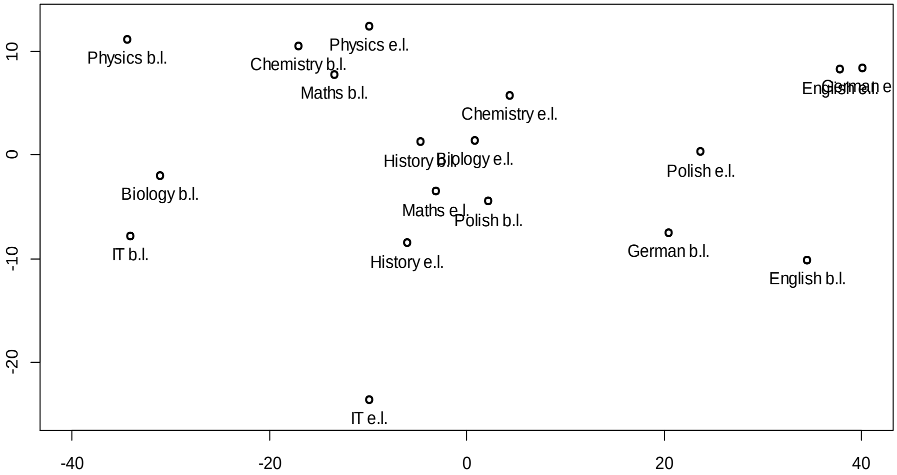


Figure 11. Objects in the component's space

Source: author's compilation using R program.

The best exam results were achieved by the secondary school graduates in German and English at the basic level (plus sign on the first and minus sign on the second component). The worst results were achieved in basic physics, basic chemistry and extended maths (minus sign on the first and plus sign on the second component).

4. Final remarks

The number of sixth graders in primary schools has stabilized since 2014, the number of final year students in middle schools and secondary schools has been continuously decreasing.

The decreasing tendency in average exam results is noticeable at all levels of education. Starting from primary school, the students receive better results in humanities, whereas science causes more problems. Primary school leavers achieved an above average result in reading, but their practical application of knowledge was below average in all subjects. An above average result was obtained in humanities part at the middle school graduation exam, whereas maths and natural science part showed below average results in all subjects. Secondary school graduates have biggest problems in science, especially in physics (basic and extended level), in basic chemistry and extended maths. The best results were recorded in German and English at the basic level.

Due to the demographic situation and the decrease in the population representing basic educational age groups the number of students has been decreasing since 2006. This situation will persist until 2020, when the students who graduated from primary schools in 2014 will start their university education.

Similarly to the classic analysis of principal components, FPCA makes this visualisation possible. It makes the analysis easier and shows nontrivial correlations, which are difficult to find in a different way. A vital property of both methods is data reduction with maximum information being kept. Thanks to the functional analysis of principal components, which enriches possibilities of the classic analysis of principal components, it is possible to analyse dynamic data, showing both the tendency as well as the pace of changes in time.

REFERENCES

DANIELE, M., (2006). Functional principal components analysis to study environmental data, the article available on the website

http://www.old.sis-statistica.org/files/pdf/atti/Spontanee%202006_677-680.pdf

HALL, P., HOSSEINI-NASAB, M., (2006). On properties of functional principal components analysis, Journal of the Royal Statistical Society, Series B (Statistical Methodology), Vol. 68, No.1, pp. 109–126.

HALL, P., MÜLLER, H. G., WANG, J. L., (2006). Properties of Principal Component Methods for Functional and Longitudinal Data Analysis, The Annals of Statistics, Vol. 34, No. 3., pp. 1493–1517.

HARMAN, H., (1975). Modern Factor Analysis, The University of Chicago Press.

INGRASSIA, S., COSTANZO, G. D., (2005). Functional principal component analysis of financial time series, in: Vichi M., Monari P., Mignani S., Montanari A. (ed.), New Developments in Classification and Data Analysis, Springer-Verlag, Berlin, pp. 351–358.

KRZYŚKO, M., GÓRECKI, T., DERĘGOWSKI, K., (2012). Jądrowa i Funkcjonalna Analiza Składowych Głównych [Nuclear and Functional Analysis of Principal Components], the meeting of Polish Statistical Association Division in Poznań, presentation available on http://stat.gov.pl/cps/rde/xbcr/pts/Krzysko_wyklad_7_11_12.pdf (access date 1st March 2015).

RAMSAY, J. O., SILVERMAN, B. W., (2005). Functional Data Analysis, Springer.

RAMSAY, J. O., HOOKER, G., GRAVES, S., (2009). Functional Data Analysis with R and MATLAB, Springer.

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 151–165

THE APPLICATION OF BŰHLMANN-STRAUB MODEL TO THE ESTIMATION OF NET PREMIUM RATES DEPENDING ON THE AGE OF THE INSURED IN THE MOTOR THIRD LIABILITY INSURANCE

Anna Szymańska[1]

ABSTRACT

One of the basic variables used in the process of tariff calculation of premiums in motor liability insurance is the age of the insured. In this type of insurance offered by insurers operating on the Polish market, this variable is taken into account in the ratemaking by discounts and increases in assigned premium, known as the net premiums rates. The aim of this work is to propose a method of rate estimation of net premiums in the groups of the motor third liability insurance portfolio of individuals created by the age of the insured. For the premium estimation, one of the maximum likelihood models, called the Bűhlmann-Straub model, was used.

Key words: a posteriori ratemaking, credibility theory, premium for a group of insurance contracts, motor third liability insurance

1. Introduction

In motor liability insurance the premium calculation process consists of two stages. The first, called a priori ratemaking, is the determination of net premiums, using actuarial methods (Ostasiewicz (ed.), 2000) based on certain risk factors, known as the basic ratemaking variables. The premium defined in this way, increased by the costs of insurance operations and the security addition among other things, is known as the base premium. The second stage of ratemaking is *a posterior* ratemaking, consisting in the base premium increases and discounts depending on individual risk factors of the insured. The bonus-malus system are one of the components of *the posterior* ratemaking commonly used in Europe. The bonus-malus systems (Lemaire, 1995, p. 3) differentiate the premium, with respect to the number of claims reported by the insured in the previous insurance period, which is based on the damage history of the insured. In addition to the

[1] Department of Statistical Methods, University of Lodz. E-mail:szymanska@uni.lodz.pl

bonus-malus system insurance companies may use other discounts and increases in the premium dependent on additional ratemaking variables, such as the age of the insured, the period of driving license holding, possession or not of children under the age of 12, profession of the insured, the age of the car, using car for business purposes, having or not any other insurance with the same company, continuation of insurance, etc. Competition on the insurance market should force the insurers to use different ratemaking variables in the ratemaking process to match premiums to the risk, represented by the insured, which should lead to lower premiums especially for those insured who do not cause damage. European countries use a few (usually from one to four) basic ratemaking variables. Most countries, including Poland, use the registration area of the vehicle and engine capacity as a major ratemaking factor pricing. In Europe, additional variables most often used in the ratemaking are the age of the insured, the use of the vehicle for commercial purposes and the age of the car. In Poland, an increase in the insured under the age of 25 is as high as 300% of the base premium. In some countries, such as France or Norway, the age is an element of *the prior* ratemaking. The aim of this paper is to present the method for determining the increases and discounts in the premium depending on the age of insured on the basis of estimation of premiums by the credibility estimation method and evaluation of premium rates related to the age applied in the audited insurance company. The author's proposition is to use the Bűhlmann-Straub model for this purpose. An example of the application of the new method is presented based on the data obtained from one of the insurance companies operating on the Polish market, which has reserved the right to stay anonymous.

2. Bűlmann-Straub model

Let X_{ij} denote the total amount of claims paid (or the number of claims) for the *i*-th insured (the *i*-th sub-group) in the *j*-th year of insurance. Suppose that the insurer has observations x_{ij}, i=1,...,N, j=1,...,t, which are the realizations of random variables X_{ij}. The amounts of payments $x_{i,t+1}$ in year t+1 are not known.

Let us assume that for each i the distribution of the random variable X_{ij} depends on parameter θ_i and that random variables X_{ij} by given $\Theta_i = \theta_i$ are independent and have the same distribution. Random vector $\mathbf{X}_i = (X_{i1},..., X_{it})$ denotes the individual history of insurance for the policy i (i-th sub-group) in a portfolio consisting of N policies (subgroups). The aim of the insurer is to determine the net premium in the year t+1 for the contract and the (i-th sub-group), given the vector $\mathbf{x}_i = (x_{i1},..., x_{it})$.

Assuming the equivalence of claims and premiums – net premium $m(\theta_i)$ for contract i (i-th sub-group) is defined by the formula:

$$m(\theta_i) = E\left(X_{i,t+1} \middle| \Theta_i = \theta_i\right) \tag{1}$$

Since we do not know θi parameter value, the value of net contributions is unknown.

The premium calculated as a weighted average from the premium for the entire portfolio, i.e. collective premiums $\mu = EX_{ij} = \frac{1}{Nt}\sum_{i=1}^{N}\sum_{j=1}^{t} x_{ij}$ and individual premium $\bar{x}_i = \frac{1}{t}\sum_{j=1}^{t} x_{ij}$ in the form:

$$m(\theta_i) = Z_i \bar{x}_i + (1 - Z_i)\mu \tag{2}$$

is called a credibility premium for the i-th contract (the i-th sub-group), where $Z_i \in [0,1]$ is a credibility factor (Kowalczyk, Poprawska and Ronka-Chmielowiec, 2006).

The estimator of variable $X_{i,t+1}$ is called a predictor of this variable, while the predictor's value is called a forecast for $X_{i,t+1}$ based on observations $x_{i1},..., x_{it}$. The basis of the credibility theory is the Bayesian statistical analysis with quadratic loss function (Krzyśko,1996).

One of the tasks of the credibility theory is to determine the values of the credibility Z_i factor. A small value of the coefficient means that the collective premium is more credible for the insurer than the individual premium. The factor Z_i is approximately equal to one when the history of damage to the policy or a group policy is long and has small variation with respect to time, or when contracts (group of policies) are very different from one another in terms of the history of damage.

Historically, the first model of the theory of credibility was the Bűhlmann model (Bűhlmann,1967), in which it is assumed that the portfolio policies can be divided into N sub-groups, each of which contains the same number of policies for which the data on t damage periods is available.

The Bűlmann-Straub model is a modified Bűlmann model, in which the number of policies included in the portfolio of individual subgroups does not have to be equal and which takes into account the importance of contracts in the portfolio. Also, the number of policies may vary periodically (Denuit, Marechal, Pitrebois and Walhin, 2007).

The model finds its application especially when a single policy or a small subset of policies differs significantly, in terms of risk profile, from the others. It is a one-way classification model. The model takes into account the weights (i.e. the volume of risk) w_{ij} of random variables X_{ij}. If the random variable X_{ij} denotes the arithmetic average of w_{ij} independent random variables with the same distributions, then the numbers w_{ij} are natural weights. The actuary, however, may establish its own weights, which do not have to be integers. In this model, insurance histories may have different lengths t_i for different contracts i. The structure of the data in the model is presented in Table 1 (Jasiulewicz, 2005).

Table 1. Structure of data in the Bűlmann-Straub model

Groups of policies	Years of insurance			
	1	2	...	t
1	x_{11} w_{11}	x_{12} w_{12}	...	x_{1t} w_{1t}
2	x_{21} w_{21}	x_{22} w_{22}	...	x_{2t} w_{2t}
...	...	...	...	...
N	x_{N1} w_{N1}	x_{N2} w_{N2}	...	x_{Nt} w_{Nt}

As previously assigned, let $\mathbf{X}_i = (X_{i1},..., X_{it})$ be a vector of observation of the number of damages for i-th policy (i-th subgroup of policies) during last t years, and let random variable Θ_i, represent the structure of risk in the portfolio.

The assumptions of the Bűlmann-Straub model (Bűlmann and Straub,1970):

1. For given i and $\Theta_i=\theta_i$, random variables $X_{i1},...,X_{it}$ are conditionally independent and

$$\mathrm{E}(X_{ij}|\theta_i) = m(\theta_i), \tag{3}$$

$$Var(X_{ij}\ |\theta_i) = \frac{s^2(\theta_i)}{w_{ij}} \tag{4}$$

for i=1,...,N, j=1,...,t, wherein variables w_{ij} are known.

2. Pairs $(\Theta_1, \boldsymbol{X}_1),...,(\Theta_N, \boldsymbol{X}_N)$ are mutually independent and random variables $\Theta_1,..., \Theta_N$ are independent and have the same distributions.

Let there be given:

– the average amount of damages for the i-th sub-group of policies:

$$\overline{X}_{iw} = \sum_{j=1}^{t} \frac{w_{ij}}{w_i} X_{ij}, \quad w_i = \sum_{j=1}^{t} w_{ij} \ , \tag{5}$$

– the average amount of damage for the portfolio:

$$\overline{X}_{ww} = \sum_{i=1}^{N} \frac{w_i}{w} \overline{X}_{iw}, \quad w = \sum_{i=1}^{N} w_i \ , \tag{6}$$

– structural parameters of risk in the portfolio:

$$\mu = Em(\Theta_i) = EX_{ij}, \quad \varphi = Es^2(\Theta_i), \quad \psi = Var(m(\Theta_i)), \tag{7}$$

where:

μ – collective net premium, which is a weighted average of the individual net premiums $m(\theta_i)$; the overall mean, it is the expected value of the claim amount for an arbitrary policyholder in the portfolio

φ – describes the average volatility of claims in a group (variation within the group)

ψ – describes the variation of claims between groups.

It can be proved that if the assumptions of the Bűlmann-Straub model are met, then (Kass, Goovaerts, Dhaene and Denuit, 2001):

1. best inhomogeneous linear predictor $\tilde{m}_i = E(X_{i\,n+1} | \mathbf{X}_i)$ of the credibility premium $m(\Theta_i)$ in the sense of least mean square error is of the form:

$$\tilde{m}_i = Z_i \overline{X}_{iw} + (1 - Z_i)\mu, \tag{8}$$

where the trust factor is $Z_i = \dfrac{w_i \psi}{w_i \psi + \varphi}$

2. best homogeneous linear predictor $\tilde{m}_i^*$ of the credibility premium $m(\Theta_i)$ in the sense of least mean square error is of the form:

$$\tilde{m}_i^* = Z_i \overline{X}_{iw} + (1 - Z_i)\overline{X}_{zw}, \tag{9}$$

where the trust factor is $Z_i = \dfrac{w_i \psi}{w_i \psi + \varphi}$ and $\overline{X}_{zw} = \sum_{i=1}^{N} \dfrac{Z_i}{Z} \overline{X}_{iw}, \quad Z = Z_1 + \ldots + Z_N$.

It can be proved that if the assumptions of the Bűlmann-Straub model are met, then unbiased estimators of structural parameters in the portfolio are of the form (Kass, Goovaerts, Dhaene, and Denuit, 2001):

$$\hat{\mu} = \overline{X}_{zw}, \ \hat{\varphi}_N = MSW, \ \hat{\psi} = \frac{w(N-1)(MSB - MSW)}{w^2 - \sum_{i=1}^{N} w_i^2}, \tag{10}$$

where:

$SSW=\sum_{i=1}^{N}\ \sum_{j=1}^{t}w_{ij}(X_{ij}-\overline{X}_{i})^{2}$ - the weighted sum of squares of deviations within groups (*sum-of-squares-within*);

$MSW=\frac{SSW}{(t-1)N}$ - the average weighted sum of squares of deviations within groups (*mean-square-within*);

$SSB=\sum_{i=1}^{N}w_{i}(\overline{X}_{iw}-\overline{X}_{ww})^{2}$ - the weighted sum of squares of deviations between groups (*sum-of-squares-between*);

$MSB=\frac{SSB}{N-1}$ - the average weighted sum of squares of deviations between groups (*mean-square-between*).

If the assumptions of the Bűlmann-Straub model are met, the average square error of inhomogenous and homogeneous predictor of credibility premium $m(\Theta_i)$ are respectively (Daykin, Pentikäinen and Pesonen,1994):

$$MSE_i=E(m(\Theta_i)-\tilde{m}_i)^2=(1-Z_i)\psi\,, \tag{11}$$

$$MSE_i^*=E(m(\Theta_i)-\tilde{m}_i^*)^2=(1-Z_i)\psi\left(1+\frac{1-Z_i}{Z}\right) \tag{12}$$

for $i=1,...,N$.

3. Example of application of the model to evaluate the rates of premium in groups separated by the age of insured

An empirical research was carried out based on the data from the portfolio of the third party liability insurance of motor vehicle owners individuals from the period of four years. For the sake of the study more than 100,000 policies were drawn for each year analyzed (the exact sample size is not specified due to the anonymity of the data). In what follows, this sample will be called portfolio. The data, in an aggregated form, on the number and value of claims paid with respect to the age groups of the insured are presented in Tables 2 and 3. The division of the insured into age groups is consistent with the classification of the insurer. The specified number of claims and the division into classes according to the value of claims paid is consistent with the tariffs of the insurer.

Table 2. The structure of the insured by the age and the number of claims paid in the motor third liability insurance portfolio in the years analyzed [%]

year	age	number of claims				portfolio
		0	1	2	3	
1	18-25	1.2560	0.0889	0.0049	0.0005	1.3503
	25-28	2.5665	0.1186	0.0049	0.0000	2.6899
	28-43	41.6109	1.6970	0.0674	0.0038	43.3790
	43-53	27.0127	1.1649	0.0361	0.0032	28.2170
	53-90	23.3994	0.9299	0.0323	0.0022	24.3638
	∑	95.8455	3.9993	0.1455	0.0097	100.0000
2	18-25	1.0846	0.0869	0.0040	0.0004	1.1758
	25-28	2.4324	0.1145	0.0047	0.0007	2.5524
	28-43	42.3952	1.7107	0.0625	0.0040	44.1724
	43-53	25.5445	1.0824	0.0385	0.0018	26.6672
	53-90	24.4352	0.9617	0.0345	0.0007	25.4321
	∑	95.8918	3.9562	0.1443	0.0076	100.0000
3	18-25	1.3342	0.1013	0.0052	0.0005	1.4411
	25-28	2.8953	0.1418	0.0079	0.0007	3.0457
	28-43	42.5567	1.7898	0.0611	0.0025	44.4100
	43-53	23.8434	1.0471	0.0486	0.0017	24.9408
	53-90	25.0821	1.0385	0.0397	0.0020	26.1624
	∑	95.7117	4.1185	0.1624	0.0074	100.0000
4	18-25	0.9935	0.0675	0.0063	0.0000	1.0672
	25-28	2.0259	0.0983	0.0048	0.0003	2.1292
	28-43	42.5934	1.5122	0.0549	0.0020	44.1626
	43-53	23.3443	0.8661	0.0344	0.0005	24.2452
	53-90	27.4171	0.9433	0.0336	0.0018	28.3958
	∑	96.3742	3.4874	0.1339	0.0045	100.0000

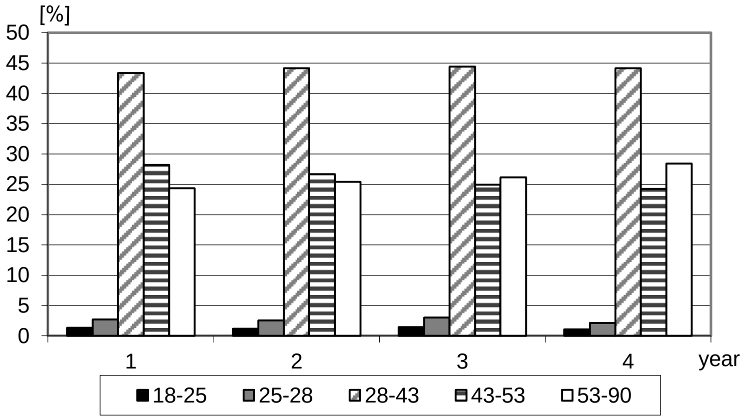


Figure 1. The structure of the insured by the age in the motor third liability insurance portfolio in the years analyzed [%]

In the years analyzed, people aged under 25 accounted for just over 1% of the insured in the analyzed portfolio. The largest group are the insured at the age from 28 to 43 (approx. 44%). Figure 2 presents the mean number of claims paid in each age group in the years analyzed. The highest loss ratio of above 0.07 in the studied period was observed in the group of under 25. Among the insured aged from 25 to 28 there has been an annual average of 0.048 to 0.052 damages. It should be noted that in other age groups, the average number of claims submitted per year was smaller, it varied from 0.036 to 0.046 and was close to the portfolio mean, which ranged from 0.038 to 0.45.

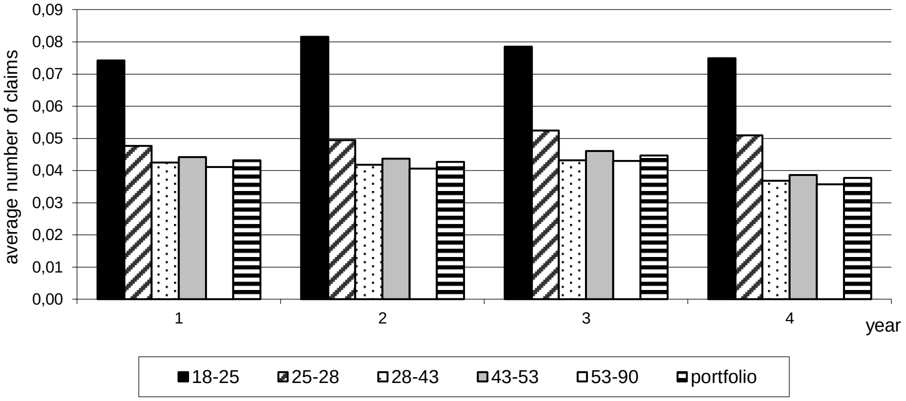


Figure 2. The average number of claims paid in the motor third liability insurance portfolio in the years analyzed according to the insured age groups

Table 3. The structure of the insured by the age and the value of claims paid in the motor third liability insurance portfolio in the years analyzed

year	age [years]	value of claim [thousands of zlotys]											portfolio
		(0,1]	(1,3]	(3,5]	(5,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,100]	(100,200]	> 200	
1	18- 25	0.31	0.65	0.39	0.52	0.23	0.09	0.03	0.00	0.04	0.01	0.00	2.27
	25-28	0.58	1.25	0.45	0.34	0.16	2.81	0.01	0.01	0.08	0.00	0.00	2.98
	28-43	8.99	17.10	6.63	5.64	2.21	16.97	0.44	0.19	0.48	0.12	0.01	42.55
	43-53	5.73	12.07	4.69	3.50	1.66	0.84	0.26	0.09	0.40	0.05	0.01	29.01
	53-90	4.42	9.68	3.65	3.11	1.33	1.13	0.14	0.06	0.30	0.04	0.01	23.20
	∑	20.04	40.75	15.82	13.10	5.59	5.33	0.88	0.36	1.30	0.15	0.04	100.00
2	up to 25	0.27	0.78	0.33	0.40	0.26	0.08	0.02	0.03	0.03	0.02	0.00	2.21
	25-28	0.49	1.04	0.48	0.47	0.25	0.12	0.04	0.02	0.03	0.00	0.00	2.92
	28-43	7.74	18.29	6.67	5.78	2.60	1.05	0.35	0.19	0.44	0.12	0.04	43.27
	43-53	5.49	10.95	4.47	3.32	1.73	0.74	0.20	0.13	0.22	0.06	0.02	27.34
	53-90	4.89	9.73	3.98	3.06	1.51	0.50	0.21	0.14	0.22	0.02	0.01	24.26
	∑	18.88	40.78	15.93	13.02	6.35	2.49	0.82	0.50	0.94	0.21	0.07	100.00

Table 3. The structure of the insured by the age and the value of claims paid in the motor third liability insurance portfolio in the years analyzed (cont.)

year	age [years]	value of claim [thousands of zlotys]											portfolio
		(0,1]	(1,3]	(3,5]	(5,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,100]	(100,200]	> 200	
3	up to 25	0.40	0.84	0.45	0.42	0.23	0.06	0.05	0.02	0.02	0.01	0.01	2.50
	25-28	0.49	1.48	0.59	0.48	0.25	0.07	0.04	0.05	0.05	0.01	0.01	3.51
	28-43	7.34	17.90	7.16	5.89	2.82	1.05	0.42	0.19	0.31	0.11	0.02	43.23
	43-53	4.29	10.97	4.00	3.38	1.74	0.55	0.23	0.12	0.23	0.05	0.02	25.60
	53-90	4.59	10.46	4.22	3.39	1.39	0.53	0.26	0.09	0.19	0.03	0.01	25.17
	∑	17.11	41.66	16.42	13.55	6.43	2.27	1.01	0.48	0.81	0.21	0.06	100.00
4	do 25	0.30	0.75	0.34	0.37	0.17	0.05	0.02	0.01	0.01	0.01	0.01	2.03
	25-28	0.42	1.00	0.64	0.44	0.18	0.10	0.03	0.01	0.02	0.00	0.00	2.84
	28-43	8.53	17.74	7.56	5.08	2.90	0.66	0.37	0.12	0.26	0.03	0.02	43.28
	43-53	4.79	10.13	4.30	3.13	1.73	0.42	0.11	0.03	0.14	0.03	0.03	24.84
	53-90	5.61	10.91	4.71	3.27	1.67	0.38	0.12	0.10	0.19	0.02	0.02	27.00
	∑	19.66	40.54	17.56	12.28	6.64	1.61	0.65	0.27	0.62	0.10	0.08	100.00

Figure 3 presents the average value of claims paid in separate age groups of the insured.

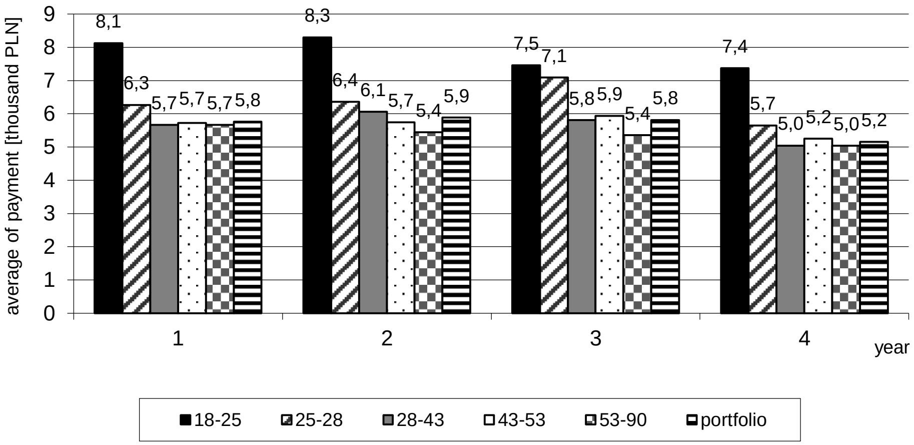


Figure 3. The average value of claims paid in the motor third liability insurance portfolio in the years analyzed, according to the age of the insured

As one can see in Figure 3, the highest average value of payments was observed in the group of insured under the age of 25. Analyzing the data from Table 3, one can observe that the structure of claims paid in the group of insured persons under the age of 25 differs from the structure of payments in other age groups. Persons under the age of 25 cause less damage of low value and more damage of higher value. For example, the payout structure of different age groups surveyed for the first years is shown in Figure 4. Damages up to 5 thousand zlotys constitute 59% of all payments in the group of insured persons under the age of 25, in other age groups such damages constitute approx. 77% of the claims paid. At the same time the group of insured persons under the age of 25 has a greater share of the compensation values from 5 thousand zlotys to 10 thousand zlotys (23% of payments in this group) and from 10 thousand zlotys to 20 thousand zlotys (10% of payments in this group), in other age groups it is approx. 12% and 6% of payments, respectively.

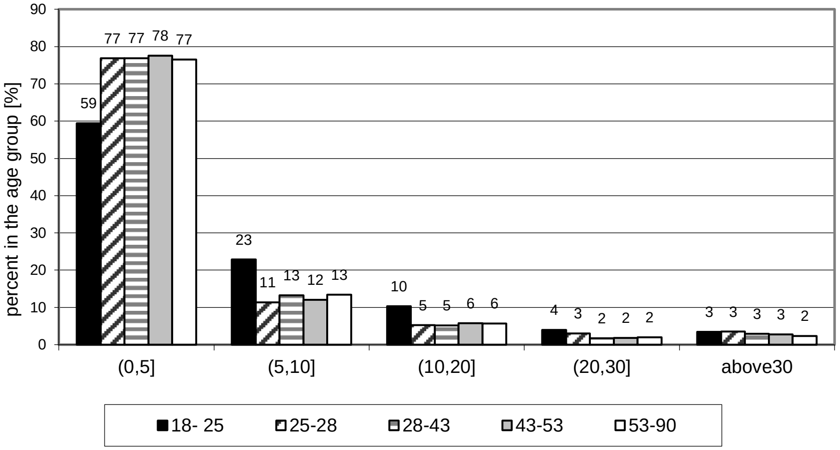


Figure 4. The structure of claims paid in the motor third liability insurance portfolio in the years analyzed, according to the age of the insured

The analysis of the number and the value of claims paid confirms the justification of taking into account the age of the insured in the ratemaking, especially for drivers aged up to 25. In motor insurance, the individual net premium in the period t+1 is determined by the equation (Szymańska, 2014):

$$\Pi(X,K) = (EX)\cdot(EK)\cdot b_{t+1} \tag{13}$$

where $\Pi(X,K)$- individual net premium in period $t+1$, EX – the expected value of a single loss in the portfolio, EK – the expected number of claims for individual insurance portfolio, b_{t+1} – the rate of premium in period $t+1$. In actuarial literature the independence of random variables of the amount and number of damages is assumed.

The aim of this study is to determine the coefficient b_{t+1} constituting the increase or discount of the premium dependent on the age of the insured. Net premiums were estimated using the Bűlmann-Straub model. Models based on the theory of credibility do not require assumptions about the form of the random variable describing the size of individual loss in the portfolio and the values of the parameters of this distribution.

The credibility premium is determined by multiplying the expected value of the estimated payments by the Bűlmann-Straub model in particular age groups of the insured (based on data from Table 4), and the expected number of claims is also estimated based on the Bűlmann-Straub model (see the data in Table 6). Two cases were considered: when the contribution of the credibility is a heterogeneous or homogeneous predictor of the net premium. Tables 5 and 7 show the results of the estimation. Premium rates ($_{i}b_{t+1}$ and $_{i}b_{t+1}^{*}$) in different age groups were calculated as the ratio of the credibility premiums in a given age group and the credibility premium in the portfolio:

$$_{i}b_{t+1} = \frac{\tilde{m}_i \cdot {}^{K}\tilde{m}_i}{\tilde{m}_{portf} \cdot {}^{K}\tilde{m}_{portf}} \cdot 100\% \tag{14}$$

$$_{i}b_{t+1}^{*} = \frac{\tilde{m}_i^{*} \cdot {}^{K}\tilde{m}_i^{*}}{\tilde{m}_{portf}^{*} \cdot {}^{K}\tilde{m}_{portf}^{*}} \cdot 100\% \tag{15}$$

The value of the credibility premiums, premium rates and net premiums is presented in Table 8.

Table 4. The average value of compensation paid [thousand zlotys] in the portfolio according to the age of the insured in the years analyzed

i (age group)	*j* (year)							
	1		2		3		4	
	X_{ij}	w_{ij} [%]	X_{ij}	w_{ij} [%]	X_{ij}	w_{ij} [%]	X_{ij}	w_{ij} [%]
1	8.12571	2.27	8.30120	2.21	7.45528	2.50	7.37585	2.03
2	6.26419	2.98	6.36364	2.92	7.08891	3.51	5.65085	2.84
3	5.66982	42.55	6.06428	43.27	5.81042	43.23	5.04389	43.28
4	5.72357	29.01	5.74814	27.34	5.93312	25.60	5.24798	24.84
5	5.66751	23.20	5.44195	24.26	5.35467	25.17	5.04204	27.00

X_{ij} – the average value of claim paid in *i*-th group in period *j* [thousand zlotys],

w_{ij} – the share of policies in *i*-th group of the portfolio in period *j* [%],

1 – group of the insured at the age of 18-25, 2 – group of insured at the age of 25-28, 3 – group of insured at the age of 28-43, 4 – group of insured at the age of 43-53, 5 – group of insured at the age of 53-90.

Table 5. Coefficients of credibility, credibility premium [thousand zlotys] and estimation errors according to age groups

i	Z_i	$\tilde{m}_i^*$	$\tilde{m}_i$	MSE_i^*	MSE_i
1	0.34	6.51136	6.35357	0.13214	0.11071
2	0.42	6.11444	5.97458	0.11497	0.09814
3	0.91	5.65019	5.62833	0.01575	0.01534
4	0.86	5.70044	5.66652	0.02479	0.02380
5	0.85	5.40268	5.36751	0.02575	0.02468

Table 6. The average number of claims in the portfolio according to the age of the insured in the years analyzed

i (age group)	j (year)							
	1		2		3		4	
	K_{ij}	w_{ij}	K_{ij}	w_{ij}	K_{ij}	w_{ij}	K_{ij}	w_{ij}
1	0.0743	1.3503	0.0816	1.1758	0.0785	1.4411	0.0750	1.0672
2	0.0477	2.6899	0.0494	2.5524	0.0524	3.0457	0.0510	2.1292
3	0.0425	43.3790	0.0418	44.1724	0.0432	44.4100	0.0369	44.1626
4	0.0442	28.2170	0.0437	26.6672	0.0461	24.9408	0.0386	24.2452
5	0.0411	24.3638	0.0406	25.4321	0.0430	26.1624	0.0358	28.3958

K_{ij} – the average number of claims in the i-th group in period j ,

w_{ij} – the share of policies in i-th group of the portfolio in period j [%],

1 – group of insured at the age of 18-25, 2 – group of insured at the age of 25-28, 3 – group of insured at the age of 28-43, 4 – group of insured at the age of 43-53, 5 – group of insured at the age of 53-90.

Table 7. Coefficients of credibility, the number of damages estimated by means of the credibility method and estimation errors for age groups of the insured

i	Z_i	${}^K\tilde{m}_i^*$	${}^K\tilde{m}_i$	MSE_i^*	MSE_i
1	0.38	0.0582	0.0554	0.0000192	0.0000164
2	0.56	0.0488	0.0468	0.0000131	0.0000117
3	0.96	0.0411	0.0409	0.0000012	0.0000012
4	0.93	0.0433	0.0429	0.0000020	0.0000020
5	0.93	0.0403	0.0399	0.0000019	0.0000019

Taking into account equations (13), (8) and (9), the value of the net premium was determined from the formulas:

$$\Pi_i(X,K) = \tilde{m}_i \cdot {}^K\tilde{m}_i \cdot {}_ib_{t+1} \tag{16}$$

$$\Pi_i^*(X,K) = \tilde{m}_i^* \cdot {}^K\tilde{m}_i^* \cdot {}_ib_{t+1}^* \tag{17}$$

The value of the credibility premiums, premium rates and net premiums is presented in Table 8.

Table 8. Net premiums and contributions of net premiums according to the age of the insured.

Age [years]	Credibility premium [thousand zlotys]		Premium rates		Net premium [PLN]	
	$\tilde{m}_i^* \cdot {}^K\tilde{m}_i^*$	$\tilde{m}_i \cdot {}^K\tilde{m}_i$	${}_ib_{t+1}^*$	${}_ib_{t+1}$	$\Pi_i^*(X,K)$	$\Pi_i(X,K)$
18-25	0.37867	0.35186	1.3920	1.4931	527.13	525.37
25-28	0.29818	0.27958	1.0961	1.1864	326.84	331.69
28-43	0.23196	0.22995	0.8527	0.9758	197.79	224.38
43-53	0.24655	0.24321	0.9063	1.0321	223.46	251.00
53-90	0.21749	0.21435	0.7995	0.9096	173.88	194.97
portfolio	0.27203	0.23565	1.0000	1.0000	272.03	235.65

4. Conclusions

The analysis of the number and the value of claims paid in the analyzed portfolio justifies the use of the age of the insured as one of the ratemaking variables. Insured persons aged from 18 to 25 cause on average more damage per year with higher average value, and therefore should pay higher premiums. The estimation results indicate a contribution rate of between 140% and 150% of the basic premium. In the analyzed insurance company, in the years investigated, the insured under the age of 25 pay 300% of the basic premium if they buy insurance for the first time in this company, and 200% of the basic premium on the continuation of insurance. Also, people aged 25 to 28, cause on average more damage with the value slightly exceeding the average value. According to the estimation method, they should pay a contribution rate of 119%. In the insurance company analyzed, for drivers aged 25 to 28 who took out insurance for the first time in the company paying the rate of 170% of the basic premium, the rate on the continuation of insurance was 130% of the premium. Another group that should have raised contributions are the insured at the age from 43 to 53, according to estimates. Insured persons in this age group should pay premiums increased by 4%. In the studied company they were at a 10% rise in the premium unless they signed a declaration about not sharing a vehicle with any person aged up to 25. The study reveals that persons aged 28-43 and over 53 could have a small discount. However, in the analyzed insurance company there were no discounts due to the age of the insured. There was also no discount due to the time of the possession of a driving license.

REFERENCES

BÜHLMANN, H., (1967). Experience Rating and Credibility, ASTIN Bulletin, 4 (3), pp. 199–207.

BÜHLMANN, H., STRAUB, E., (1970). Glaubwürdigkeit für Schadensätze, Mitteilungen der Vereiningung scheizerischer Vesicherungsmathematiker, 1970, pp. 111–133.

DAYKIN, C. D., PENTIÄINEN, T., PESONEN, M., (1994). Practical Risk Theory for Actuaries, Chapman & Hall, London.

DENUIT, M., MARECHAL, X., PITREBOIS, S., WALHIN, J. F., (2007). Actuarial Modelling of Claim Counts. Risk classification, Credibility and Bonus-Malus Systems, J. Wiley & Sons, England.

JASIULEWICZ, H., (2005). Teoria zaufania. Modele aktuarialne, Wydawnictwo AE im. Oskara Langego we Wrocławiu, Wrocław.

KAAS, R., GOOVAERTS, M., DHAENE, J., DENUIT, M., (2001). Modern Actuarial Risk Theory, Kluwer, Boston.

KRZYŚKO, M., (1996). Statystyka Matematyczna, Wydawnictwo Naukowe UAM, Poznań.

KOWALCZYK, P., POPRAWSKA, E., RONKA-CHMIELOWIEC, W., (2006). Metody aktuarialne, PWN, Warszawa.

LEMAIRE, J., (1995). Bonus-Malus Systems in Automobile Insurance, Kluwer, Boston.

OSTASIEWICZ, W., (red.), (2000). Modele Aktuarialne, Wydawnictwo Akademii Ekonomicznej im. O. Langego we Wrocławiu, Wrocław.

SZYMAŃSKA, A., (2014). Statystyczna analiza systemów bonus-malus w ubezpieczeniach komunikacyjnych, Wydawnictwo UŁ, Łódź.

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 167–180

EVALUATION OF THE EU COUNTRIES' INNOVATIVE POTENTIAL – MULTIVARIATE APPROACH

Elżbieta Roszko-Wójtowicz[1], Jacek Białek[2]

ABSTRACT

The aim of the article is to work out a synthetic measure for estimating country's innovation potential (CIP) of EU economies. For the purpose of the research, data from the European Statistical Office (Eurostat) are used and several indicators are organized by four different areas of analysis, i.e. investment expenditure, education, labour market and effects. Applying multi-dimensional statistics allows us to reduce the primary set of diagnostics variables and, simultaneously, identify those which best describe the potential. The final step is linear ordering of EU countries according to their innovative potential on the basis of CIP synthetic measure. The rating is compared with other ratings based on the recognized *Summary Innovation Index* and *Global Innovation Index*. The main conclusion is that the methodology of innovativeness assessment remains an open issue and requires further research. The most important task is the selection of indicators, followed by statistical verification in relation to their importance to innovativeness. The results show that there is a tendency to between the author's ratings and other already published ratings of innovativeness.

Key words: innovativeness, Innovation Union Scoreboard, European Union, cluster analysis, factor analysis.

1. Introduction

Changes in knowledge resources as well as ability to utilize them determine the possession of country in the contemporary world. Capability of using knowledge and information as well as efficient application of modern technology form the basis of building up innovativeness (compare Soete (2000), OECD (2005), Pilat & Woelfl (2003)). Innovativeness represents capability of performing creative acts, inventing new ideas and inventions. Innovativeness manifests itself in an attempt to search for new combinations of production factors, introducing new value added to competitive products as well as

[1] Department of Economic and Social Statistics, University of Lodz, Poland.
E-mail: eroszko33@gmail.com.

[2] Department of Statistical Methods, University of Lodz, Poland. E-mail: jbialek@uni.lodz.pl.

application of knowledge achievements in the production process (Granstrand (1999)). Innovations are a significant factor of the competitiveness of the economy. They are an inherent part of constant and sustainable economic development. Moreover, their importance increases when the country's economy becomes more developed (Cornelius & McArthur (2002)). The question of innovativeness and innovation on micro, meso and macroeconomic level was reflected in the theory of economy and management as well as multiple articles, e.g. by P. Drucker (2004), J. Schumpeter (1960), M. Porter (2001), E. Rogers (2003), and others. Nevertheless, it is Joseph A. Schumpeter (1883 – 1950) who is believed to have coined the term 'innovation'. He described economic process as a creative act which means creating, designing and implementing innovation (Schumpeter (1960)). The authors of the article were encouraged to raise the topic by the lack of unanimity in terms of measuring innovative potential of economies. The purpose of the article is a statistical analysis of factors influencing innovativeness of EU economies. The result of the quantitative analyses is linear ordering of EU countries according to the level of their innovative potential. The rating was compared with the outcome presented in *Innovation Union Scoreboard* (*IUS*) based on *Summary Innovation Index* (*SII*).

2. Measuring innovativeness

Measuring innovation remains a relatively new branch of statistics, although it is gaining a wide interest from both practitioners as well as theorists. One of the best known studies on innovativeness is *Global Innovation Index – GII*[3]. It is an annual report released by experts of Johnson Cornell University, one of the largest management and business schools in the world – INSEAD – the Business School for the World and *The World Intellectual Property Organization – WIPO*). The framework is composed of 79 individual indicators describing innovation, which was divided into 7 categories, i.e. institutions, human capital and research, infrastructure, market sophistication, business sophistication, knowledge and technology output and creative output (Dutta, Lanvin & Wunsch-Vincent (2015)).

European Innovation Scoreboards presents another source of information on innovative activity in particular member states. EIS distinguishes the following products: *Innovation Union Scoreboard*, *Regional Innovation Scoreboard* and a new element with its pilot implementation in 2013 – *European Public Sector Innovation*[4]. Data used for creating *EIS* come from multiple primary resources but also public data obtained from European Patent Office and Office for Harmonization in the Internal Market. Individual indicators collected for *EIS*

[3] The Global Innovation Index, https://www.globalinnovationindex.org/content/page/GII-Home, access 22.10.2015.

[4] For more details see: European Commission portal, http://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards/index_en.htm, European Commission (2013).

allow for working out the *Innovation Union Scoreboard* based on a composite innovation indicator *Summary Innovation Index* (*SII*[5]). Currently, 25 indices divided into five categories are used to estimate SII. The first three sub-groups are input indicators whereas the next two – output ones. *Input* comprises: a) innovation enablers that illustrate conditions for innovation development, which are not directly related to the activity of enterprises, b) *firm's activities* – present innovative activity of a company. *Output* stands for effects that demonstrate results of innovative activity in business (European Commission (2015)). SII index ranges from 0 to 1, however, the closer the index value to 1, the higher the innovativeness level of a given country's economy. Estimated SII value gives basis for classifying EU countries into four groups according to the level of economy innovativeness.

In 2013 the European Commission introduced additional index – *The Innovation Output Indicator* focusing on measuring innovative activity output. It emerged in response to an objective formulated in the Europe 2020 Strategy, concerning increased expenditure on R&D. The new indicator allows the assessment of the progress of member states in achieving established benchmarks. Simultaneously, it supplements Innovation Union Scoreboard (IUS) and Summary Innovation Index (SII). The new indicator suggested by the European Commission is based on four elements significant in terms of EU policy: (1) European technological innovations are measured by the number of patents granted, (2) Employment level in knowledge-intensive activities, expressed by percentage of total employment, (3) Competitiveness of knowledge-intensive products and services, (4) Employment level in fast-growing enterprises in innovative sectors (European Commission (2013)).

3. Description of empirical research

3.1. Research objective

The research aims at working out a synthetic measure estimating country's innovation potential – CIP. The authors' main objective is reducing the primary set of diagnostics variables and simultaneously distinguishing variables which best describe innovation potential of particular member states. The goal shall be achieved by application of various yet complementary methods of multidimensional statistics. The specific objective of the paper is linear ordering of EU countries according to their innovation level based on *CIP* synthetic measure. The following ranking will be compared with ratings based on recognized *Summary Innovation Index and Global Innovation Index*.

[5] SII reaches values from 0 (low innovativeness) to 1 (high innovativeness).

3.2. Description of diagnostics indicators

Data presented in the article come from the European Statistical Office – Eurostat. For the analysis of innovative potential of EU member states, 25 variables were selected in total, and categorized into four different areas of analysis, i.e. investment expenditure, education, labour market, effects. It is a common practice to make clusters of data into specific areas of analysis when building innovativeness ratings (compare *SII* and *GII*). Making the first selection of features for the analysis of the innovative potential of economies, the authors aimed at creating a unique personal attitude, acknowledging the outcomes achieved in the discussed field at the same time. Therefore, two rules were applied when selecting variables: at least two variables representing each distinguished area are also included in *SII* and/or *GII* (1), each area of the analysis is dominated by variables suggested by the authors of the research (2). Moreover, the authors suggest that when creating innovation ratings too little attention is given to society treated as part of the process of creating innovation. Society presents a starting point for creating innovation, its needs and deliberate pursuit of applying innovation are the driving force. This is why the analysis included variables illustrating the employment level, education level, society's interest in information and communication technologies. This particular aspect makes the approach closer to that presented in *GII* rather than in *SII*. Nevertheless, the authors believe that *Global Innovation Index* sees innovativeness from a too broad perspective. As a result, real advantages of particular economies become hard to establish. Furthermore, the aim of multivariate analysis should be to identify only those determinants that are crucial for socio-economic growth through innovation. The subjects of the analysis include EU-28 countries, also referred to as analysis units. For the purpose of estimating EU countries' innovative potential, each presented diagnostic variable is treated as a stimulant, which means that the growth of the value influences the analysed phenomenon in a positive way. In constructing an index of a country's innovation potential, *Global Innovation Index* (Dutta et al. (2015) and *Summary Innovation Index* (European Commission (2015)) methodology were used as a framework for selecting and placing the diagnostic variables into four areas (investment, education, labour market, effects). As a supplement, policy recommendations of the OECD Working Paper (Freudenberg (2003)) and OECD Growth Project (OECD (2001)) were applied. The classification scheme consists of four core areas that combine between five to eight diagnostic variables. To analyse EU countries' innovative potential the initial set comprises 25 indicators, mostly derived from the statistical office of the European Union – Eurostat databases (http://ec.europa.eu/eurostat) – see Table 1. The first core area sees investment expenditure both from public and private perspective. The second core area aggregates variables related to the educational achievements of a country. In the third area labour market is presented. The last area looks at effects of innovative activities including patents, community designs and trademarks, as well as a share of innovative enterprises.

Table 1. Initial and final dataset

Investment expenditure	**X1**	Public R&D expenditure as % of GPD;	**Labour market**	**X2**	Employment rate in population aged 20-64;
	X7	Total public expenditure on education as % of GPD;		**X3**	Employment in technology and knowledge-intensive sectors as a percentage of total employment;
	X13*	Business enterprises R&D expenditure as % of GPD;		**X12**	Percentage of SMEs (10-249 employees) that employed ICT/IT specialists;
	X17	Percentage of households with a broadband Internet connection;		**X14***	Total R&D personnel as a percentage of total employment;
	X22*	Percentage of households with Internet access;		**X5***	Percentage of individuals aged 25-64 using computer at least once a week but not every day;
Education	**X4**	Percentage of population aged 30-34 with tertiary education degree;		**X10**	Percentage population/individuals aged 25-64 using Internet at least once a week including every day;
	X6	Students (ISCED 1_6) aged 15-24 – as % of a corresponding age population	**Effects**	**X8***	Patents granted by United States Patent and Trademark Office per 1 million inhabitants;
	X11	Percentage of individuals aged 25-64 with competences in terms of using computers (at least 5 out of 6 activities listed in the research);		**X9**	Patents filed to European Patent Office per 1 million inhabitants;
	X15	Percentage of individuals aged 15-24 having participated in tertiary education (ISCED 5-8) (formal education, ISCED 5-8);		**X19**	Community trade mark (CTM) registrations per 1 million inhabitants;
	X16	Percentage of individuals aged 25-64 participating in non-formal education and training;		**X20***	Innovative enterprises (including enterprises with abandoned/suspended or ongoing innovation activities) as a percentage of total number of enterprises;
	X18	Graduates (ISCED 5-6) in science, mathematics and technology aged 20-29 per 1000 citizens;		**X21***	Community design (CD) applications per 1 million inhabitants;
	X23*	New doctorate graduates (ISCED 6) per 1 million inhabitants;		**X25***	SMEs introducing product or process innovations as percentage of SMEs;
	X24	Graduates (ISCED 5-6) in science, mathematics and computing, engineering, manufacturing and construction as percentage of all graduates;			

Note: Implementation of the proposed statistical procedure resulted in reduction of 9 variables from the initial dataset – see variables marked with "*".

Source: own elaboration based on data from the statistical office of the European Union – Eurostat – http://ec.europa.eu/eurostat

Building the database, the authors aimed at selecting most up-to-date data available in Eurostat. For this reason, variables used for the research come from different years since Eurostat database is not completed on a regular basis. In the case of 12 variables, the data regard 2014 (X_2, X_4, X_5, X_{10}, X_{11}, X_{12}, X_{15}, X_{16}, X_{17}, X_{19}, X_{21}, X_{22}), 6 variables represent 2013 values (X_1, X_3, X_{13}, X_{14}, X_{23}, X_{24}), another 5 – 2012 (X_6, X_9, X_{18}, X_{20}, X_{25}) and one – 2011 (X_7) and one – 2009 (X_8).

3.3. Methodology

Primary reduction of diagnostic variables was conducted by correlation and cluster analysis. Another reduction of diagnostic variables was based on factor analysis carried out by means of normalized Varimax rotation. For further analysis, variables included in selected factors were used. The presented approach concerns only several statistical methods whereas many others, being potentially helpful in selecting indicators and measuring innovativeness, are discussed widely in papers: Saisana & Tarantola (2002), Freudenberg (2003) or Cherchye et. al. (2005). As an added value of the paper, an analytical strategy, which allows for the application of a combination of complementary statistical methods, is adopted here. The selection of reducing a pre-defined set of variables, based on criteria that are meant to sort out redundant information, is a step forward to the conceptual model of innovativeness. The ultimate EU countries' innovativeness rating was created by applying a non-pattern linear ordering, with weighted and unweighted variant. Methods of reducing the set of diagnostics variables, the form of their weights used in the analysis and the proposition of some measures of the innovativeness can be treated as a new approach in the discussed area due to the fact that the existing methodology (connected with SII or GII) has a poor statistical justification[6]. The analyses were carried out by means of Statistica 8.0 and MS Excel.

4. Innovativeness determinants

4.1. Prselection of data

During the first stage, correlation analysis was applied to reduce the number of variables. From all pairs of variables where Pearson correlation coefficient was at least 0.95, it was the variable with a higher deviation coefficient based on standard deviation that was selected for further analysis. This procedure allowed elimination of co-linearity of explanatory variables, maintaining the most significant variables for the research at the same time. The exception is X_8 and X_9 pair of variables, which are strongly positively correlated, $(\rho_{8,9} = 0.95)$, being comparable in terms of variability. For further analysis, X_9 variable was chosen

[6] For instance, diagnostic variables are highly correlated and they have the same weights in the final index formula.

where data on EU countries are more up to date[7]. At this stage, variable X_{17} remained despite the high level of correlation with X_{10} and X_{22} variables ($\rho_{10,17} = 0.93$ and $\rho_{17,22} = 0.99$ respectively), assuming the access to broadband Internet an indirect determinant of changes in EU countries' innovativeness. Consequently, (compare Tab. 1) variables X_8, X_{20} and X_{22} were initially eliminated.

4.2. Reducing the number of variables by means of cluster analysis

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). A cluster can be described largely by the maximum distance needed to connect parts of the cluster (see Everitt (2011)). The next step towards dimensionality reduction of explanatory variables was clustering of variables[8]. The analysis was supposed to distinguish variables creating clusters, i.e. most similar variables (of the lowest value of Euclidean value). Clusters obtained through the lowest level of aggregation were later compared with correlation matrix identified a priori. It was concluded that X_{14} variable may be omitted without a significant loss of information, which results from the fact that its distance to X_{13} variable is the closest of the observed Euclidean distances[9], the variables represent a high correlation $(\rho_{13,14} = 0.83)$; in addition, variable X_{14} has a much lower volatility.

4.3. Reducing the number of variables by factor analysis

We used factor analysis to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors (see Child (2006), Thomson (2004)). Mathematically speaking, the object of factor analysis is a matrix of data containing n number of m variables $X = [x_{ij}]_{nxm}$, where $i = 1,2,\ldots, n$, $j = 1,2,\ldots, m$. As a result of transforming the value of variables by means of standardization formula we achieve variables of identical expected value (equals 0) and unit standard deviation: $Z = [z_{ij}]_{nxm}$. In this research, the reduced set of 21 variables underwent factor analysis. Principal components method was used to distinguish most relevant factors and corresponding factor loadings[10] (compare Walesiak (1996)). Yet, Varimax

[7] In the case of variable X_8 the latest data come from 2009, while for X_9 variable – 2012.

[8] Generally, clustering is conducted for object class recognition by searching most homogenous clusters (of closest possible distance within the cluster and maximum possible distance to other clusters). It may be referred to as one of methods used for reduction of variables.

[9] Most homogenous clusters are built up by variables X_1, X_3, X_7, X_{13}, X_{14}, with variables X_{13} and X_{14} being closest by Euclidean distances.

[10] The factor loadings, also called component loadings, are the correlation coefficients between the cases (rows) and factors (columns). The squared factor loading is the percent of variance in that indicator variable explained by the factor. As a rule of thumb, in confirmatory factor analysis

normalized rotation[11] was introduced to maximize the variance of primeval factor loadings on variables. The following variables X5, X13, X31, X23 and X25 are removed from further analysis.

5. Results and discussion

The five-factor solution, implicitly identified by factor loadings, corresponds to a priori chosen classification scheme. This is confirmed by the fact that all the core areas (*Investment expenditure*, *Education*, *Labour market*, *Effects*) have their representatives in the final dataset, which comprises 16 variables. The reduction of the number of indicators from the predefined set is presented in Table 1. It is worth noticing that the statistical procedure proposed in the paper allowed for removing two variables from each of the core areas apart from *effects*, where reduction was made by 4 variables. In total, this makes the procedure more input than output oriented. In the case of synthetic measurement of innovative potential this is a very important issue.

The last stage of the analysis is establishing the EU countries' rating by their innovative potential. For this reason, the linear ordering method was applied, weighed and unweighted variant. Variables which serve as stimulators for innovativeness potential were first standardized, then two synthetic measures were created: M_{1k} (unweighted variant) and M_{2k} (weighted variant) for each country $k = 1,2,\ldots,28$, i.e.:

$$M_{1k} = \frac{1}{m}\sum_{i=1}^{m} z_{ik} \, , \tag{1}$$

$$M_{2k} = \frac{1}{\lambda}\sum_{i=1}^{m} \lambda_i z_{ik} \, , \tag{2}$$

where:

z_{ik} – standardized value of each variable (i) established for a specific country (k); $m -$ the number of other *analyzed* variables ($m = 15$), λ_i – weight related to $i -$ the variable set. The $i -$th weight is the quotient where the numerator is an identified variance multiplied by the factor from which the variable is derived, divided by summary percentage of the variance identified by all factors, while denominator is the number of variables creating a particular factor, i.e. $\lambda = \sum \lambda_i$. The aggregation methods described in formulas (1) and (2) are widely

(CFA), loadings should be 0.7 or higher to confirm that independent variables identified a priori are represented by a particular factor.

[11] Varimax rotation is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by the extracted factor. Each factor will tend to have either large or small loadings of any particular variable. A varimax solution yields results which make the identification of each variable with a single factor as easy as possible.

used in linear ordering of objects (see Bąk (2015)). Further, in their report prepared for the European Commission, Saisana and Tarantola (2002) state that this approach is commonly applied and (…) "*The composite indicator is based on the standardized scores for each indicator which equal the difference in the indicator for each country and the EU mean, divided by the standard error.*" In fact, the presented method of aggregation makes the final index more robust when dealing with outliers.

The higher the synthetic factor value, the higher a given country's innovative potential. In the case of the 5 clusters of variables, they were assigned the following weights: 51.4%, 20.6%, 11.8%, 9.3%, 6.9%. Table 2 contains EU countries' rating by the descending values of M_{1k} and M_{2k} measures.

Table 2. EU countries' rating by innovative potential assessed by M_{1k} and M_{2k} compared to SII and GII results.

Country	M1k		M2k		SII		GII	
	Rank	Value	Rank	Value	Rank	Value	Rank	Value
Denmark	1	1.15	2	1.07	2	0.74	7	57.70
Finland	2	1.12	1	1.12	3	0.68	4	59.97
Sweden	3	0.92	3	0.96	1	0.74	2	62.40
Germany	4	0.69	4	0.82	4	0.68	8	57.05
Netherlands	5	0.47	5	0.44	5	0.65	3	61.58
United	6	0.39	7	0.39	7	0.64	1	62.42
Austria	7	0.37	6	0.4	11	0.59	9	54.07
Luxembourg	8	0.26	8	0.31	6	0.64	6	59.02
Estonia	9	0.22	9	0.29	13	0.49	11	52.81
France	10	0.21	10	0.27	10	0.59	10	53.59
Ireland	11	0.19	13	0.03	8	0.63	5	59.13
Slovenia	12	0.16	11	0.16	12	0.53	16	48.49
Czech	13	0.09	12	0.11	14	0.45	12	51.32
Belgium	14	0.07	14	-0.01	9	0.62	13	50.91
Malta	15	-0.06	17	-0.21	18	0.40	14	50.48
Spain	16	-0.09	16	-0.1	19	0.39	15	49.07
Portugal	17	-0.14	15	-0.09	17	0.40	17	46.61
Hungary	18	-0.27	20	-0.36	20	0.37	21	43.00
Lithuania	19	-0.3	18	-0.25	25	0.28	23	42.26
Slovakia	20	-0.33	19	-0.31	22	0.36	22	42.99
Italy	22	-0.45	21	-0.41	16	0.44	18	46.40
Latvia	21	-0.45	22	-0.46	26	0.27	19	45.51
Poland	23	-0.5	23	-0.47	24	0.31	27	40.16
Cyprus	24	-0.57	26	-0.66	15	0.44	20	43.51
Croatia	25	-0.58	24	-0.55	23	0.31	25	41.70
Greece	26	-0.63	25	-0.6	21	0.36	26	40.28
Bulgaria	27	-0.87	27	-0.91	27	0.23	24	42.16
Romania	28	-1.05	28	-0.97	28	0.20	28	38.20

Note: calculations carried out by means of Statistica 8.0 and MS Excel

Source: own elaboration based on data from the statistical office of the European Union – Eurostat – http://ec.europa.eu/eurostat

The next step was comparing the EU member states' ratings created by means of linear ordering with *Summary Innovation Index* as well as *Global Innovation Index* (compare Tab. 2). The convergence of all the ratings was assessed with Spearman correlation coefficients. We obtained all correlation coefficients over 0.9 although we observe differences between ratings for rank positions from the middle. High correlations are justified due to the fact that positions of the most innovative countries and these with the lowest innovativeness performance are not threatened regardless of the set of diagnostic variables – primary or reduced.

The applied statistical tools (correlation analysis, cluster analysis, factor analysis) enabled reducing the number of diagnostic variables from 25 to 16 (compare Tab. 1). In this way, the authors reached their primary objective of maximum reduction of the set of features and distinguishing those which best identify the analysed phenomenon. Factor analysis led to identifying five principal factors explaining almost 80% of the total variance of variables. It is worth noticing that the identified factors are of multidimensional nature, which relates to multi variable factors. It means that one factor comprises features covering various areas of analysis, i.e. investment spending, education, labour market and effects (compare Section 3.2). The obtained results are relevant to the ones presented in the literature, where innovativeness is described by means of sets of variables representing different areas. For instance, the first distinguished factor (identifying almost 40% of the total variance) includes both variables of the *investment spending* (e.g. variable X_1) or *labour area* (variable X_2), as well as of other areas: *effects* (X_9) or *education* (X_{16}). Similarly, the second and third factor consist of variables representing different analysis areas, i.e. variables from the second factor include: X_6 and X_{18} (*education*), X_{19} (*effects*), and variables from the third factor include: X_3 (*labour market*), X_7 and X_{12} (*investment expenditure*). It should be emphasized that next two one-element factors relate to *education* area. One may draw a conclusion that human capital is a significant factor in building economy's innovative potential (compare Wheatley (2001), Klingbeil (2008)). According to R. E. Lucas, one of the basic factors stimulating economic innovativeness is human capital, which leads to technological progress when combined with the size and efficiency of R&D investment (Lucas (1988)).

5. Conclusions

The purpose of the research was to check an alternative approach to the approach that prevails in studying innovative potential of selected economies. For this reason, the authors attempted to create a rating based on possibly narrowest yet carefully selected set of diagnostic variables. A comparative analysis of the authors' rating and the rating based on *SII* lead to important conclusions on the ultimate assessment of EU countries' innovative potential. There is a great deal of convergence between authors' and *SII* rating, especially when it comes to top (Denmark, Finland, Sweden) as well as bottom positions (Bulgaria, Romania), which is confirmed by high rank correlation coefficients established for comparative ratings. Central positions in the ratings reveal major differences (Tab. 2). As some contrast, the proposed rankings differ from the GII rating, especially in the case of top 9 countries (for instance, *GII* evaluates the United Kingdom as a winner while this country is ranked 6th or 7th in other rankings). The proposal of the new innovativeness measure and the fact that linear ordering for the EU member countries with *CIP* index is convergent with the rating based on *SII* is to provide additional support for the adopted strategy. Further, the statistical procedure applied in the article may serve as a tool supporting creation of innovativeness conceptual framework and the initial selection of indicators. Nevertheless, the research outcome confirms a commonly shared view that the methodology of innovativeness assessment requires further research.

REFERENCES

BĄK A., (2015). Zagadnienie wyboru optymalnej procedury porządkowania liniowego w pakiecie pllord, [w:] Prace naukowe Uniwersytetu Ekonomicznego we Wrocławiu, No. 384, pp. 33–41, Wrocław.

CATTELL R. B., (1966). The scree test for the number of factors, Multivariate Behavioral Research, 1, pp. 245–276.

CHERCHYE, L., LOVELL, C.A.K., MOESEN, W., VAN PUYENBROECK, T., (2005). One market, one number? A composite indicator assessment of EU internal market dynamics, Discussion Paper Series DPS 05.13. Center for Economic Studies, Leuven University, http://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/KEI-WP5-D5.2.pdf.

CHILD, D., (2006). The Essentials of Factor Analysis (3rd ed.), Continuum International.

CORNELIUS, P., MCARTHUR, J. W., (2002). The Global Competitiveness Report 2001-2002, World Economic Forum, Oxford University Press, New York, Oxford, http://www.nectec.or.th/pld/indicators/documents/WEF-%20Global%20Competitiveness%20Report%202001.pdf.

DRUCKER, P. F., (2004). Natchnienie i fart, czyli innowacja i przedsiębiorczość, Studio Emka, Warszawa 2004.

DUTTA, S., LANVIN, B., WUNSCH-VINCENT, S. (ed.), (2015). The Global Innovation Index 2015. Effective Innovation Policies for Development, Cornell University, INSEAD, the World Intellectual Property Organization (WIPO), Fontainebleau, Ithaca, and Geneva, https://www.globalinnovationindex.org/userfiles/file/reportpdf/gii-full-report-2015-v6.pdf.

EUROPEAN COMMISSION, (2013). Innovation Union. A pocket guide on a Europe 2020 Initiative, Publications Office of the European Union Luxembourg, Belgium.

EUROPEAN COMMISSION, (2013). Powering European Public Sector Innovation: Towards a New Architecture, Report of the Expert Group on Public Sector Innovation, Directorate General for Research and Innovation, Publications Office of the European Union, Luxembourg, https://ec.europa.eu/research/innovation-union/pdf/psi_eg.pdf.

EUROPEAN COMMISSION, (2013). Measuring innovation output in Europe: towards a new indicator, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Brussels.

EUROPEAN COMMISSION, (2015). Innovation Union Scoreboard 2015, Internal Market Industry, Entrepreneurship and SMEs, Belgium.

EVERITT, B., (2011). Cluster analysis, Chichester, West Sussex, U.K: Wiley.

FREUDENBERG, M., (2003). Composite indicators of country performance: a critical assessment, STI Working Paper 2003/16, http://www.oecd-ilibrary.org/docserver/download/405566708255.pdf?expires=1483002854&id=id&accname=guest&checksum=C8780649DE56D4475FDEE57AF84A4F45.

GRANSTRAND, O., (1999). The Economics and Management of Intellectual Property. Towards Intellectual Capitalism, Edward Elgar, Cheltenham 1999

KAISER H. F., (1960). The application of electronic computers to factor analysis, Educational and Psychological Measurement, 20, pp. 141–151.

KLINGBEIL, M., (2008). Measuring Human Capital in the Knowledge Economy, Ministry of Advanced Education, Canada.

LUCAS, R. E., (1988). On the Mechanics of Economic Development, Journal of Monetary Economics, 1988, No. 21.

OECD, (2001). The New Economy: Beyond the Hype. The OECD Growth Project.

OECD, (2005). Good Practice Paper on ICTs for Economic Growth and Poverty Reduction, Paris, http://www.oecd.org/dac/35284979.pdf.

PILAT, D., WOELFL, A., (2003). ICT and Economic Growth – New Evidence From International Comparisons, OECD, Paris, http://www.tiger.edu.pl/konferencje/kwiecien2003/papers/Pilat_Woelfl.en.pdf.

PORTER, M. E., (2001). Porter o konkurencyjności, PWE, Warszawa.

ROGERS, E. M., (2003). Diffusion of Innovations, Free Press, New York.

SAISANA, M., TARANTOLA S., (2002). State-of-the-art Report on Current Methodologies and Practices for Composite Indicator Development, European Commission: Institute for the Protection and the Security of the Citizen Technological and Economic Risk Management Unit, Italy.

SOETE, L., (2000). Governing the Information Society, Communications & Strategies, Vol. 37, pp. 155–167.

THOMPSON, B., (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications, Washington DC: American Psychological Association.

WALESIAK, M., (1996). Metody analizy danych marketingowych, PWN, Warszawa 1996.

WHEATLEY, M., (2001). Innovation means relying on everyone’s creativity. Leader to Leader, (Spring), http://www.margaretwheatley.com/articles/innovationmeans.html.

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 181

The 18th Scientific Conference
Quantitative Methods in Economics 2017

organized by
Department of Econometrics and Statistics,
Warsaw University of Life Sciences – SGGW

Conference date: June 19th – 20th, 2017
Conference Venue: The Center for Nature and Forestry Education in Rogów

Invited Speakers:

- Prof. Ryszard Budziński – University of Szczecin
- Prof. Czesław Domański – University of Lodz,
- Prof. Stanisław Gędek – Rzeszow University of Technology,
- Prof. Karol Kukuła – University of Agriculture in Krakow,
- Prof. Włodzimierz Okrasa – Central Statistical Office of Poland,
- Prof. Tadeusz Trzaskalik – University of Economics in Katowice,
- Prof. Andrzej Wiatrak – University of Warsaw
- Prof. Dorota Witkowska – University of Lodz.

The Scientific Conference *Quantitative Methods in Economics* is a yearly conference organized by Department of Econometrics and Statistics. The main aim of the conference is a presentation of the latest research results as well as the integration of polish and international academic staff dealing with a mathematics and informatics applications in economics.

Conference Scope and Topics:

- statistics, econometrics and biometrics,
- multidimensional data analysis,
- financial engineering and operational research,
- financial mathematics and insurance,
- informatics applications in economics.

Call for papers: www.mibe.sggw.pl

STATISTICS IN TRANSITION new series, March 2017
Vol. 18, No. 1, pp. 183–185

ABOUT THE AUTHORS

Białek Jacek is working as an Professor at the University of Lodz, Department of Statistical Methods. He graduated with master of science degree in Applied Mathematics from the Technical University of Lodz (Department of Physics, Informatics and Applied Mathematics, 2003). In 2007 he defended his PhD thesis in economics at the Faculty of Economy and Sociology of the University of Lodz and he habilitated in 2015. His main scientific interest concentrates on price index theory. His second field of research interest is the innovativeness and efficiency measurement and entrepreneurship development.

Budny Katarzyna is an Assistant Professor in the Department of Mathematics at the Cracow University of Economics. Her main areas of interest include mathematical statistics and probability theory. Her research concentrates on the theory of multivariate distributions, estimation methods and multivariate probabilistic inequalities (especially generalizations of Chebyshev's inequality). She is also interested in applications of statistical methods in finance.

Chandra Shalini is an Associate Professor (Statistics) in the Department of Mathematics & Statistics, Banasthali Vidyapith, Banasthali-304022, Rajasthan, India. She/He has received her/his master’s degree in Statistics from Lucknow University and qualified CSIR-NET JRF in Mathematical Sciences. She/He received her/his PhD in Statistics from Lucknow University and partially from Maharshi Dayanand University, Rohtak, Haryana, India in 2006. She/He visited Indian Statistical Institute, Tezpur center as a visiting scientist between June 2011 and June 2012. Her/His research interests include regression analysis, time series analysis, econometrics and biostatistics. She/He has a number of research articles published in reputed national and international journals. She/He has supervised 5 research scholars to date. She/He has been teaching statistics to undergraduate and postgraduate students for the last 14 years.

Dziechciarz-Duda Marta (PhD) is an Assistant Professor in the Department of Econometrics and Computer Science at Wroclaw University of Economics, Poland. Her major areas of research interest include multivariate statistical analysis and econometric modelling of socio-economic data. Her research concentrates on areas of consumer durable goods market, its structure analysis, segmentation and sales forecasting. She uses information on households’ endowment with durable consumer goods as a proxy for the measurement and assessment of households’ material situation. Her second field of research interest is problems of measurement and modelling of education quality and effectiveness.

Kordos Jan graduated from the Jagiellonian University, Krakow (in mathematics, 1953), and the University of Wroclaw (in mathematical statistics, 1955); PhD in Econometrics from the Academy of Economics, Katowice, Poland (1965), and Professorship (1990). He worked as the Chief of the Methodology Section at the Division of Living Conditions, Central Statistical Office/CSO (1955-1966) and of the Laboratory of Mathematical Methods at the Research Center of Statistics and Economics (CSO, 1966-74). He served as the FAO Adviser in Agricultural Statistics in Ethiopia (1974-80). He acted as Director of the Division of Demographic and Social Surveys (1981-92) and as Vice-President of the CSO Poland (1992-96). He was lecturing and training on agricultural statistics in China in the late 1980s, and also in Kathmandu, Nepal (1989, 1991). During 1994-96 he served as the World Bank Consultant in Household Budget Surveys in Latvia and Lithuania. He was President of the Polish Statistical Association (1985-94). He was the founder and editor-in-chief of *Statistics in Transition* (1993-2007). Now, he is Professor of Statistics at the Warsaw Management University. His publications include four books and over three hundreds articles and other papers. He has been an elected member of the International Statistical Institute since 1974.

Król Anna is a PhD student in the Department of Econometrics at Wrocław University of Economics. She is interested in application of econometrics and multivariate statistical methods, especially in the fields of quality-adjusted price indexes, marketing research and education research.

Mehta Vishal is a Visiting Scientist in the Indian Statistical Institute (ISI), North-East Centre, Tezpur- 784028, Assam, India. He received his PhD in Statistics from the Vikram University, Ujjain, Madhya Pradesh, India in 2016. Dr. Mehta’s research interest include the areas of statistics, sampling theory; statistical inference - use of prior information in estimation procedure.

Roszko-Wójtowicz Elżbieta (PhD) is working as an Assistant Professor at the University of Lodz, Department of Economic and Social Statistics. Her main scientific interests concentrate on measurement and assessment of socio-economic phenomena especially in the field of innovativeness, entrepreneurship development and lifelong learning. Since 2007 she has been actively involved as an expert, external evaluator or coordinator in research projects conducted on regional or national level (European Social Fund or Ministry of Higher Education) and international level (Leonardo da Vinci). Her second field of research interest is the development of family businesses that include the transgenerational succession process. She is a member of Association of Career Development (Lodz, Poland) and Polish Economic Society (Branch in Lodz, Poland).

Singh Housila P. is a Professor in the School of Studies in Statistics, Vikram University, Ujjain-456010, Madhya Pradesh, India. He has published his methodological and applied research extensively in the leading journals of statistics. Prof. Singh’s research interest include survey sampling and statistical inference - use of prior information in estimation procedure.

Sztemberg-Lewandowska Mirosława received her MSc in mathematics from the Faculty of Mathematics and Computer Science of University in Wrocław, Poland, in 1997, where she received also her PhD in economics in 2003. Currently, she is an Assistant Professor at University of Economics in Wroclaw, Department of Econometrics and Computer Science in Jelenia Góra. Her research interests include methods of statistical multidimensional analysis in marketing research. Her scientific interests include methods of statistical multidimensional analysis in marketing research: principal component analysis, factor analysis, structural equation models, multi-group analysis, IRT models, functional principal component analysis and independent principal components analysis. Recent publications concern the analysis of the situation of education at different stages of education in Poland.

Szymańska Anna is an Professor Dr in the Department of Insurance, Institute of Finance, Faculty of Economics and Sociology, University of Lodz. Her research interests focus on insurance statistics and actuarial mathematics. She graduated from the Faculty of Mathematics, University of Lodz.

Tyagi Gargi is an Assistant Professor (Statistics) in Department of Mathematics & Statistics, Banasthali Vidyapith, Banasthali-304022, Rajasthan, India. She/He has received her/his MSc in Mathematical Sciences with specialization in Statistics from Banasthali University, Banasthali, Rajasthan in 2010 and MPhil from the same university in 2011. She/He qualified UGC-NET (Population Studies) in December 2010. She/He is currently pursuing PhD in Statistics from Banasthali University. Her/his research interests include econometrics and regression analysis. She/he has been teaching statistics to undergraduate and postgraduate students for the past 5 years.

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).