STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

CONTENTS

Volume 20, Number 2, June 2019

EDITORIAL OFFICE **ISSN 1234-7655**

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

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

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. I−IV

FROM THE EDITOR

Let me start this issue with a good news about our journal. On behalf of the Editorial Board and the Editorial Office, I am pleased to share with our partners and supporters – authors, reviewers and readers – information about the continuous growth of the *Statistics in Transition new series* in terms of the scores and recognition by the prestigious indexation bases. For instance, according to the new CiteScore metrics released by Scopus for the year 2018, our journal has obtained 0, 27 (compared to 0, 21 in 2017). Even more, the journal has recently been included into the New Scimago Journal & Country Rank – a portal developed from the information contained in the Scopus database. We are thankful to everyone who contributed to these achievements in any way – they all count!

A set of eleven papers published in this issue of the *Statistics in Transition new series* presents a heterogeneous body of works – from statistical and econometric research articles through the conference papers from the 2nd Congress of Polish Statistics (Warsaw, July 2018) and a paper from the 27th Conference on Classification and Data Analysis (Ciechocinek, September 2018) to a research communicate.

The main part which contains research papers starts with an article ***Statistical inference of exponential record data under Kullback-Leiber divergence measure*** by **Raed R. Abu Awwad, Ghassan K. Abufoudeh** and **Omar M. Bdair.** Using one parameter exponential record data, the authors conduct statistical inferences (maximum likelihood estimator and Bayesian estimator) for the suggested model parameter. They also aim to predict the future (unobserved) records and to construct their corresponding prediction intervals based on observed set of records. In the estimation and prediction processes, the square error loss and the Kullback-Leibler loss functions are employed. Numerical simulations were conducted to evaluate the Bayesian point predictor for the future records. Subsequently, data analyses involving the times (in minutes) to breakdown of an insulating fluid between electrodes at voltage 34 kV have been performed to show the performance of the methods developed on estimation and prediction.

Anoop Chaturvedi's and **Sandeep Mishra's** paper ***Generalized Bayes Estimation of Spatial Autoregressive Models*** deals with the spatial autoregressive (SAR) models, which are widely used in spatial econometrics for analysing spatial. The authors derive a Generalized Bayes estimator for estimating the parameters of a SAR model for data involving spatial autocorrelation structure. The admissibility and minimaxity properties of the estimator have been discussed. For investigating the finite sample behaviour of the estimator, the results of a simulation study have been presented, and the approach was applied to demographic data on total fertility rate for selected

Indian states - an improvement over the usual least squares estimator was demonstrated for a wide range of the parametric settings.

Jana Cibulková, Zdenek Šulc, Sergej Sirota, and Hana Rezankova discuss ***The effect of binary data transformation in categorical data clustering*** in the paper focused on hierarchical clustering of categorical data. They compare two approaches, the first of which embraces performing a binary transformation of the categorical variables into sets of dummy variables and then use the similarity measures suited for binary data, while the second uses similarity measures developed for the categorical data. The comparison of these two approaches is performed on generated datasets with categorical variables and the evaluation is done using both the internal and the external evaluation criteria. In conclusion, the authors demonstrate that the binary transformation is not necessary in the process of clustering categorical data since the second approach yields comparable results under the less demanding conditions.

Grażyna Dehnel's and **Marek Walesiak's** paper ***A comparative analysis of economic efficiency of medium-sized manufacturing enterprises in districts of Wielkopolska province using the hybrid approach with metric and interval-valued data*** describes a hybrid approach to evaluating economic efficiency of medium-sized manufacturing enterprises (employing from 50 to 249 people) in districts of Wielkopolska province, using metric and interval-valued data. The authors employ an approach that combines multidimensional scaling with linear ordering. The analysis was based on data prepared in a two-stage process. First, a data set of observations was obtained for metric variables describing economic efficiency of medium-sized manufacturing enterprises. Next, the unit-level data were aggregated at district level and turned into two types of data: metric and interval-valued data. The analysis of interval-valued data was carried out using symbolic-to-classic and symbolic-to-symbolic approaches, and the results of the two approaches were compared. [The calculations were made with scripts prepared in the R environment.]

Adedayo A. Adepoju's and **Tayo P. Ogundunmade's** paper ***Economic growth and its determinants: a cross-country evidence*** presents evidence from a panel of 126 countries, over the time period of 2010 to 2014, that economic growth is dependent on various factors. The authors found that government expenditure control, reduced inflation and increased trade openness are the factors that boost the economic growth of a country. Significant evidence is seen for government consumption, fiscal policy and trade openness. No significant relationship has been observed between exchange rate and economic growth, whereas unemployment influences output for African countries. The cross regional analysis of Asian, European, African, Caribbean and American countries gives specific determinants for these regions. Fiscal balance has shown a consistent positive relationship with economic growth throughout the analyses. Fiscal balance and unemployment rate played their role in the growth of African countries. Inflation rates and increased openness were significant for some regions. Exchange rate did not return significant coefficients for any of the sub-regions. Government consumption, trade openness, policy interest rate and industrial production rate showed significant effect for different regions of the world.

In the article ***Application of the strategy combining monetary unit sampling and a Horvitz–Thompson estimator of error amount in auditing – results of a simulation study,*** **Bartłomiej Janusz** discusses a possible alternative for testing audit populations with high error rates under a pragmatic assumption that auditors need information on the performance of different statistical methods when applied to audit populations. A strategy combining systematic Monetary Unit Sampling and confidence intervals for the total error based on the Horvitz-Thompson estimator with normality assumption was checked – including its reliability and efficiency – using real and simulated data sets. It was shown that, for the majority of populations, the interval coverage rate was lower than the assumed confidence level. In most cases confidence intervals were too wide to be of practical use to auditors. Confidence intervals tended to become wider as the observed error rate increased. Tests disclosed that the distribution of the Horvitz-Thompson estimator was not normal. A detailed analysis of the distributions of the error amount in the examined real audit populations is also given.

The next paper ***Estimation of Energy Intensity in Indian Iron and Steel Sector: A Panel Data Analysis*** by **Anukriti Sharma, Hiranmoy Roy** and **Narendra Nath Dalei** presents results of the empirical estimation of the energy intensity of Indian Iron and Steel sector accounting for the impact of ECA (Energy Conservation Act, 2001) and PAT (Perform, Achieve and Trade mechanism), Phase-I in dummy variable form. The results indicate that the decline in energy consumption in this sector until 2011 can also be attributed to Energy Conservation Act implemented in the year 2001 along with other factors. The authors conclude that ECA has a significant impact on reduction of energy intensity of the steel firms. PAT does not seem to have a considerable impact on energy intensity alone but in the years where both PAT and ECA are prevalent, i.e. from 2012 to 2015, there seems to be a significant impact of around 0.050 reduction in energy intensity, as accounted for by different models in this paper. In addition, the empirical results suggest that profit margin intensity was found to be negatively related to energy intensity implying more profitable firms invest more in energy efficiency.

The next articles are based on the conference presentations.

In the paper ***Variable selection in multivariate functional data classification*** by **Tomasz Górecki, Mirosław Krzyśko,** and **Waldemar Wołyński,** a new variable selection method is considered in constructing a classification using multivariate functional data approach. The variable selection is a dimension reduction method, which leads to the replacement of the high-dimensional vector process by a low-dimensional vector with a comparable classification error. Various classifiers appropriate for functional data are used. The proposed variable selection method is based on functional distance covariance (dCov) and the Hilbert-Schmidt Independent Criterion (HSIC), which are discussed in the literature. The method employed is a modified version of the procedure described by Kong et al. (2015), and the proposed methodology is illustrated with a real data. The authors consider this approach as an alternative to other variable selection methods.

Roman Zmyślony's and **Arkadiusz Kozioł's** paper ***Testing hypotheses about structure of parameters in models with block compound symmetric covariance structure*** deals with testing the hypotheses of the so-called structured mean vector and the structure of a covariance matrix. To this aim, Jordan algebra properties are used and tests based on best quadratic unbiased estimators (BQUE) are constructed. For convenience coordinate-free approach is employed as a tool for characterization of best unbiased estimators and testing hypotheses. To obtain the test for mean vector, linear function of mean vector with the standard inner product in null hypothesis is changed into equivalent hypothesis about some quadratic function of mean parameters (it is shown that both hypotheses are equivalent and testable). In both tests the idea of the positive and negative part of quadratic estimators is applied to get the test statistics which have F distribution under the null hypothesis. Finally, power functions of the obtained tests are compared with other known tests like LRT or Roy test. For some set of parameters in the model the presented tests have greater power than the above mentioned tests.

The paper by **Barbara Pawełek, *Extreme gradient boosting method in the prediction of company bankruptcy*** discusses the use of machine learning methods to predict company bankruptcy. Comparative studies carried out on selected methods to determine their suitability for predicting company bankruptcy have demonstrated high levels of prediction accuracy for the extreme gradient boosting method in this area. This method is resistant to outliers and relieves the researcher from the burden of having to provide missing data. The special aim of this study is to assess how the elimination of outliers from data sets affects the accuracy of the extreme gradient boosting method in predicting company bankruptcy, with intention to show the advantages of application of the extreme gradient boosting method in bankruptcy prediction based on data free from the outliers. The research was conducted using 64 financial ratios for the companies operating in the industrial processing sector in Poland. The research results indicate that it is possible to increase the detection rate for bankrupt companies by eliminating the outliers reported for companies which continue to operate as a going concern from data sets.

The issue concludes with a research communicate*, **Efficient two-parameter estimator in linear regression,*** by **Ashok V. Dorugade,** who discusses two-parameter estimators in linear model with multicollinearity. The author proposes an alternative efficient two-parameter estimator along with examination of its properties and the results of its comparison with the OLS estimator, as well as the ordinary ridge regression (ORR) estimators. Also, using the mean squared error criterion the proposed estimator performs more efficiently than OLS estimator, ORR estimator and other reviewed two-parameter estimators. A numerical example and simulation study are finally conducted to illustrate the superiority of the proposed estimator.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. V

SUBMISSION INFORMATION FOR AUTHORS

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

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

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

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

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

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl,
GUS/Statistics Poland,
Al. Niepodległości 208, R. 296, 00-925 Warsaw, Poland

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

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

EDITORIAL POLICY

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

***

ABSTRACTING AND INDEXING DATABASES

Statistics in Transition new series is currently covered in:

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

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 1–14, DOI 10.21307/stattrans-2019-011

STATISTICAL INFERENCE OF EXPONENTIAL RECORD DATA UNDER KULLBACK-LEIBLER DIVERGENCE MEASURE

Raed R. Abu Awwad[1], **Ghassan K. Abufoudeh**[2], **Omar M. Bdair**[3]

ABSTRACT

Based on one parameter exponential record data, we conduct statistical inferences (maximum likelihood estimator and Bayesian estimator) for the suggested model parameter. Our second aim is to predict the future (unobserved) records and to construct their corresponding prediction intervals based on observed set of records. In the estimation and prediction processes, we consider the square error loss and the Kullback-Leibler loss functions. Numerical simulations were conducted to evaluate the Bayesian point predictor for the future records. Finally, data analyses involving the times (in minutes) to breakdown of an insulating fluid between electrodes at voltage 34 kv have been performed to show the performance of the methods thus developed on estimation and prediction.

Key words: Bayes estimation, Bayes prediction, record values, Kullback-Leibler divergence measure, exponential distribution.

1. Introduction

Let $X_1, X_2, ...$ be a sequence of independent and identically distributed (iid) random variables from exponential distribution with probability density function (pdf)

$$f(x;\theta) = \begin{cases} \theta e^{-\theta x} & if\ x > 0, \theta > 0 \\ 0 & if\ x \leq 0\ , \end{cases} \tag{1}$$

and cumulative distribution function (cdf)

$$F(x;\theta) = 1 - e^{-\theta x}, x > 0, \theta > 0. \tag{2}$$

Based on the distribution function of exponential distribution, the distribution can be used effectively in analyzing any lifetime data, especially when censoring is used or if the data are grouped. The exponential distribution received considerable attention in the literature during the last three decades and was commonly used in many situations of lifetime data analysis. The exponential distribution was typically used

[1]Department of Mathematics, Faculty of Arts and Sciences, University of Petra, Amman, Jordan. E-mail: raed_abuawwad@yahoo.com

[2]Department of Mathematics, Faculty of Arts and Sciences, University of Petra, Amman, Jordan. E-mail: ghassan_math@yahoo.com

[3]Faculty of Engineering Technology, Al-Balqa Applied University, Amman 11134, Jordan. E-mail: bdairmb@bau.edu.jo. ORCID ID: https://orcid.org/0000-0002-5346-4381.

to model time intervals between random events, such as the length of time between arrivals at a service station. In queuing theory, the service times of agents in a system are often exponentially distributed. It is worth mentioning here that when times between "random events" follow the exponential distribution with rate θ, then the total number of events in a time period of length t follows the Poisson distribution with parameter θt. Reliability theory and reliability engineering also use the exponential distribution extensively. Many authors have developed inference procedures for exponential distribution. Abufoudeh *et al*. (2017) have obtained the Bayes estimate of the parameter of the exponential distribution under Kullback-Leibler divergence measure. Nasiri *et al*. (2012) have used the upper record range statistic to draw inferences from the parameter of the exponential distribution. Based on record data, Janeen (2004) have discussed the empirical Bayes estimators for the parameter of the exponential distribution. An interested reader may refer to Balakrishnan *et al*. (2005). Balakrishnan *et al*. (1995) have established some recurrence relations for single and product moments from exponential distribution based on record values. Ahsanullah and Kirmani (1991) have obtained some characterizations of the exponential distribution based on lower record values.

In many real life situations, we may be interested in the largest value of data such as stock exchange, weather and sports, because in some cases the decisions may depend on the largest values. Chandler (1952) has introduced the study of record values and has reported many of the basic properties for records. Bdair and Raqab (2009) have studied the mean residual lifetime of records and Bdair and Raqab (2012) have studied the upper bounds of the mean residual lifetime of records. Properties of record values have been extensively studied in the literature by Ahsanullah (1988, 1995), Arnold and Balakrishnan (1989), Arnold *et al*. (1998), Nevzorov (2001), Kamps (1995) and Jaheen (2004).

Let $X_1, X_2, ...$ be a sequence (X-sequence) of iid random variables from the exponential distribution given in Eq. (1). The random variable X_j is called an upper record if $X_j > X_i$ for all $i = 1, 2, ..., j-1$. To formalize this concept, let $X_1, X_2, ..., X_n$ be a sample of size n from X-sequence. By convention X_1 is the first record where $U(1) = 1$ is the first record time. For $n \geq 2$, X_j is an upper record if its value exceeds all of the previous observations. To obtain record data, the n^{th} record time $U(n)$ is defined using the recursive formula $U(n) = \min\left\{j : X_j > X_{U(n-1)}\right\}$, then the n^{th} record is $X_{U(n)}$.

In this research work, based on record values we estimate the parameter θ of the exponential distribution using both classical and Bayesian methods of estimation, as well as we predict the future record values depending on a sequence of past records. In the Bayesian estimation method and in prediction of future values, we use two types of loss functions; the first is the square error loss (SEL) function, which is defined as

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2. \tag{3}$$

The second is the Kullback-Leibler divergence measure (KL) as an alternative loss function. The Kullback-Leibler divergence measure (also called relative entropy

measure) has been introduced by Kullback and Leibler (1951). Unlike the square error loss function, KL does not measure the discrepancy between an unknown parameter and its estimate, but between the actual distribution $f(x|\theta)$ of the record sample $\underset{\sim}{x}$ of size n from X-sequence and the approximate distribution $\hat{f}(x|\hat{\theta})$. As a consequence, it is invariant with one-to-one reparametrization of the parameters and, hence, becomes a serious competitor to square error loss function. An interesting property of the KL divergence is $KL(f,\hat{f}) \geqslant 0$ with equality if and only if $f(x|\theta) = \hat{f}(x|\hat{\theta})\ \forall x \in \underset{\sim}{x}$. For more details, one may refer to Abufoudeh *et al*. (2018) and Singh *et al*. (2014). The Kullback-Leibler divergence measure of the true distribution $f(x|\theta)$ from the approximate distribution $\hat{f}(x|\hat{\theta})$ is defined as

$$\begin{aligned} KL(f,\hat{f}) &= E_f\left[\log \frac{f(x|\theta)}{\hat{f}(x|\hat{\theta})}\right] \\ &= E_f\left[\log \frac{\theta e^{-\theta x}}{\hat{\theta} e^{-\hat{\theta} x}}\right] \\ &= \log \frac{\theta}{\hat{\theta}} - (\theta - \hat{\theta})\, E_f(X) \\ &= \frac{\hat{\theta}}{\theta} - \log \frac{\hat{\theta}}{\theta} - 1. \end{aligned} \tag{4}$$

This measure is called Kullback-Leibler error loss (KEL) function and it is denoted by $KL(\theta,\hat{\theta})$.

The rest of the article is organized as follows. In Section 2, based on record data from the exponential distribution of parameter θ, we find the maximum likelihood estimate and the Bayes estimate of θ, under both SEL and KEL functions. In Section 3, we find the point and credible interval of the future records based on previously known records generated from the exponential distribution. A simulation study based on different sizes of record samples from the exponential distribution and real life example is presented in Section 4. Simulation studies that compare all classical and Bayes estimates along with a real life example are presented and discussed in Section 5. Finally, we conclude the results thus obtained in Section 6.

2. Classical method

The most common classical technique in estimating the unknown parameters of a distribution is the maximum likelihood (ML) estimation method. The ML estimation method chooses, as an estimate of θ, the value $\hat{\theta}$, which maximizes the likelihood function. Suppose we observe n upper record values $\underset{\sim}{x} = (x_{U(1)}, x_{U(2)}, ..., x_{U(n)})$ from X-sequence of iid random variables following the exponential distribution with pdf and cdf given in (1) and (2), respectively. According to Arnold *et al*. (1998), the

likelihood function of records based on exponential distribution is given as

$$\begin{aligned} L(\theta|\underset{\sim}{x}) &= \prod_{i=1}^{n-1} \frac{f(x_{U(i)}|\theta)}{1-F(x_{U(i)}|\theta)} f(x_{U(n)}|\theta) \\ &= \theta^n \, e^{-\theta x_{U(n)}}. \end{aligned} \tag{5}$$

Applying $\ln(.)$ for both sides, we obtain the log-likelihood function

$$\ln L(\theta|\underset{\sim}{x}) = n \ln \theta - \theta x_{U(n)}.$$

Differentiating the above equation with respect to θ and equating the resulting term to zero, we obtain the ML estimator of θ as follows

$$\hat{\theta} = \frac{n}{x_{U(n)}}.$$

3. Bayesian method

In this section, we introduce the Bayesian point estimation and credible interval for the unknown parameter of the exponential distribution based on upper record values. The KEL and SEL functions are used to approximate the point estimation of the unknown parameter θ.

3.1. Bayesian estimation

The inference problem concerning the unknown parameter θ can easily be dealt with using the Bayesian method, since the posterior distribution supposedly contains all available information about θ (both sample and prior information). The posterior distribution of θ given $\underset{\sim}{x}$ is defined as

$$\pi(\theta|\underset{\sim}{x}) = \frac{L(\theta|\underset{\sim}{x})\,\pi_1(\theta)}{\int_0^\infty L(\theta|\underset{\sim}{x})\,\pi_1(\theta)\,d\theta}. \tag{6}$$

Assume that θ has the conjugate gamma prior with pdf

$$\pi_1(\theta|a,b) = \begin{cases} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} & \text{if } \theta > 0, \\ 0 & \text{if } \theta \leq 0, \end{cases} \tag{7}$$

where $a > 0$, $b > 0$ are the hyper-parameters. By substituting Eq. (5) and Eq. (7), we immediately obtain

$$\pi(\theta|\underset{\sim}{x}) = \frac{(b+x_{U(n)})^{a+n}}{\Gamma(a+n)} \, \theta^{a+n-1} \, e^{-\theta(b+x_{U(n)})}. \tag{8}$$

That is, the posterior distribution of θ given $\underset{\sim}{x}$, is $Gamma(a+n, b+x_{U(n)})$.
The Bayesian estimator of θ under the SEL function is then given by

$$\hat{\theta}_{B_1} = E_{posterior}(\theta|\underset{\sim}{x}) = \frac{a+n}{b+x_{U(n)}}.$$

The Bayesian estimator of θ under the KEL function $\hat{\theta}_{B_2}$ is obtained by minimizing the risk function

$$E_{posterior}(KL(\theta,\hat{\theta})) = \int_0^\infty \left(\frac{\hat{\theta}}{\theta} - \log\frac{\hat{\theta}}{\theta} - 1\right)\pi(\theta|\underset{\sim}{x})\, d\theta.$$

By differentiating $E_{posterior}(KL(\theta,\hat{\theta}))$ with respect to $\hat{\theta}$ and setting its derivative to zero, we get the equation

$$\int_0^\infty \left(\frac{1}{\theta} - \frac{1}{\hat{\theta}}\right)\pi(\theta|\underset{\sim}{x})\, d\theta = 0.$$

Solving for $\hat{\theta}$ we conclude

$$\hat{\theta}_{B_2} = \frac{1}{E_{posterior}(\frac{1}{\theta}|\underset{\sim}{x})} \tag{9}$$

Using Eq. (8) and Eq. (9), the Bayes estimator under the KEL function is then

$$\hat{\theta}_{B_2} = \frac{\Gamma(a+n)}{(b+x_{U(n)})\Gamma(a+n-1)}.$$

3.2. Credible interval

Since the posterior distribution of θ follows gamma distribution, a credible interval of θ can be obtained as follows:
The $(1-\beta)100\%$ credible interval of θ, (C_L, C_U), satisfies the following two conditions

$$P(C_L < \theta < \infty) = 1 - \frac{\beta}{2}, \tag{10}$$

$$P(C_U < \theta < \infty) = \frac{\beta}{2}. \tag{11}$$

Now, from Eq. (10), we have

$$\int_{C_L}^{\infty} \frac{\left(b+x_{U(n)}\right)^{a+n}}{\Gamma(a+n)}\,\theta^{a+n-1}\,e^{-\theta\left(b+x_{U(n)}\right)}\,d\theta = 1 - \frac{\beta}{2}.$$

Using the transformation $u = \theta(b + x_{U(n)})$, we immediately obtain

$$\int_{(b+x_{U(n)})C_L}^{\infty} u^{a+n-1}\, e^{-u}\, du = \left(1 - \frac{\beta}{2}\right)\Gamma(a+n).$$

Based on the incomplete gamma function, which is defined as

$$\Gamma(c,d) = \int_{d}^{\infty} x^{c-1} e^{-x}\, dx,\ c > 0,\ d > 0, \tag{12}$$

we immediately obtain

$$\Gamma\left(a+n,\ (b+x_{U(n)})C_L\right) = \left(1 - \frac{\beta}{2}\right)\Gamma(a+n). \tag{13}$$

Similarly from Eq. (11), we obtain

$$\Gamma\left(a+n,\ (b+x_{U(n)})C_U\right) = \frac{\beta}{2}\Gamma(a+n). \tag{14}$$

Consequently, we conclude the lower and upper credible interval C_L and C_U by solving Eqs. (13) and (14) using a suitable numerical method, with respect to C_L and C_U, respectively.
In particular, if a is a positive integer, then the chi-square table values can be used to construct the credible interval for θ as follows:
Since θ has $Gamma(a+n,\ b+x_{U(n)})$, then a pivotal statistic $Q = 2\theta(b+x_{U(n)})$ has $\chi^2_{2(a+n)}$. Hence, the $(1-\beta)100\%$ credible interval for θ is given by

$$\frac{\chi^2_{(1-\frac{\beta}{2},\,2(a+n))}}{2(b+x_{U(n)})} < \theta < \frac{\chi^2_{(\frac{\beta}{2},\,2(a+n))}}{2(b+x_{U(n)})},$$

where $\chi^2_{(\beta,r)}$ is the $100\beta^{th}$ upper percentile of chi-square with r degrees of freedom.

4. Bayesian prediction

In this section, we consider the problem of one sample prediction. The idea of this problem is to find the Bayes predictors and bounds of future record values based on observed records which have been taken from X-sequence. We consider the two loss functions SEL and KEL to find the predictors and bounds. One sample prediction problem has been studied by many authors, see Ahsanullah (1980), Dunsmore (1983), Berred (1998) and Bdair and Raqab (2016).

4.1. Predictors of future records

Let $\underset{\sim}{x} = (x_{U(1)}, x_{U(2)}, ..., x_{U(m)})$ be the m observed upper records. To find the Bayes predictor of the n^{th} future upper record $X_{U(n)}$, $1 \leq m < n$, we need to derive the posterior predictive density at any point $y > x_{U(m)}$, as follows:
The conditional probability density function of $y = x_{U(n)}$ given that the observed upper record data $\underset{\sim}{x}$ is indicated by $f_{X_{U(n)}|\underset{\sim}{x}}(y|\theta)$. Since the upper record values satisfy the Markovian property, then $f_{X_{U(n)}|\underset{\sim}{x}}(y|\theta) = f_{X_{U(n)}|x_{U(m)}}(y|\theta)$. The conditional probability density function of $y = x_{U(n)}$ given that $x_{U(m)}$ is given (see Ahsanullah (1995)) by

$$\begin{aligned} f_{X_{U(n)}|x_{U(m)}}(y|\theta) &= \frac{[H(y) - H(x_{U(m)})]^{n-m-1}}{(n-m-1)!} \frac{f(y|\theta)}{1 - F(x_{U(m)|\theta})} \\ &= \frac{\theta^{n-m} e^{\theta x_{U(m)}}}{(n-m-1)!} (y - x_{U(m)})^{n-m-1} e^{-\theta y}, \end{aligned}$$

where $H(.) = -\ln(1 - F(.))$. Using the well-known Binomial expansion, we immediately obtain

$$f_{X_{U(n)}|x_{U(m)}}(y|\theta) = \frac{\theta^{n-m} e^{\theta x_{U(m)}}}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i (x_{U(m)})^i y^{n-m-1-i} e^{-\theta y}.$$

The posterior predictive density at any point $y > x_{U(m)}$, is then

$$\begin{aligned} f^P_{X_{U(n)}|\underset{\sim}{x}}(y|\theta) &= E_{posterior}\left[f_{X_{U(n)}|x_{U(m)}}(y|\theta) \right] \\ &= \int_0^\infty \frac{\theta^{n-m} e^{\theta x_{U(m)}}}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i (x_{U(m)})^i y^{n-m-1-i} e^{-\theta y} \pi(\theta|\underset{\sim}{x}) d\theta. \end{aligned}$$

The Bayes estimator of future records under the SEL function, is given by

$$\begin{aligned} X^{BP1}_{U(n)} &= E_{f^P}(Y|\underset{\sim}{x}) \\ &= \int_{x_{U(m)}}^\infty \left[\int_0^\infty \frac{\theta^{n-m} e^{\theta x_{U(m)}}}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i (x_{U(m)})^i y^{n-m-1-i} e^{-\theta y} \pi(\theta|\underset{\sim}{x}) d\theta \right] dy \\ &= \int_0^\infty \frac{\theta^{n-m} e^{\theta x_{U(m)}}}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i (x_{U(m)})^i \left(\int_{x_{U(m)}}^\infty y^{n-m-1-i} e^{-\theta y} dy \right) \pi(\theta|\underset{\sim}{x}) d\theta. \end{aligned}$$

Using the transformation $u = \theta y$ and Eq. (12), we obtain

$$X_{U(n)}^{BP1} = \int_0^{\infty} \frac{e^{\theta x_{U(m)}}}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i (x_{U(m)})^i \frac{\Gamma(n-m-i+1, \theta x_{U(m)})}{\theta^{1-i}} \pi(\theta|\underset{\sim}{x})\, d\theta.$$

Based on the MCMC samples $\{\theta_j;\ j = 1,2,...,M\}$ generated from Eq. (8), the Bayes predictor of future records becomes

$$\hat{X}_{U(n)}^{BP1} = \frac{1}{M} \sum_{j=1}^{M} \frac{e^{\theta_j x_{U(m)}}}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i (x_{U(m)})^i \frac{\Gamma(n-m-i+1, \theta_j x_{U(m)})}{\theta_j^{1-i}}.$$

The Bayes predictor of future records under the KEL function, is given by

$$X_{U(n)}^{BP2} = \frac{1}{E_{f^P}(\frac{1}{Y}|\underset{\sim}{x})}$$

Using a similar argument, the Bayes predictor of future records will be

$$\hat{X}_{U(n)}^{BP2} = \frac{M}{\sum_{j=1}^{M} \frac{e^{\theta_j x_{U(m)}}}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i (x_{U(m)})^i \frac{\Gamma(n-m-i-1, \theta_j x_{U(m)})}{\theta_j^{-1-i}}}.$$

4.2. Bounds of future records

Under SEL function, we present the Bayesian predicted bounds of the $(1-\beta)100\%$ interval of the future record value, $Y = X_{U(n)}$, (Y_L, Y_U).
The lower bound Y_L can be obtained by solving the following equation for Y_L

$$\int_{Y_L}^{\infty} f_{X_{U(n)}|\underset{\sim}{x}}^{P}(y|\theta)\, dy = 1 - \frac{\beta}{2},$$

or equivalently

$$\int_{Y_L}^{\infty} f_{X_{U(n)}|X_{U(m)}}^{P}(y|\theta)\, dy = 1 - \frac{\beta}{2}.$$

This is equivalent to solve the equation

$$\int_{Y_L}^{\infty} \left[\int_0^{\infty} \frac{\theta^{n-m} e^{\theta x_{U(m)}}}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i (x_{U(m)})^i y^{n-m-1-i} e^{-\theta y} \pi(\theta|\underset{\sim}{x})\, d\theta \right] dy = 1 - \frac{\beta}{2},$$

which yields to

$$\int_0^{\infty}\left[\frac{e^{\theta x_{U(m)}}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i(x_{U(m)})^i\frac{\Gamma(n-m-i,\theta Y_L)}{\theta^{-i}}\right]\pi(\theta|\underset{\sim}{x})\,d\theta=1-\frac{\beta}{2}.$$

Based on the MCMC samples $\{\theta_j;\ j=1,2,...,M\}$, the lower bound Y_L can be found by solving the equation

$$\frac{1}{M}\sum_{j=1}^{M}\left[\frac{e^{\theta_j x_{U(m)}}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i(x_{U(m)})^i\frac{\Gamma(n-m-i,\theta_j Y_L)}{\theta_j^{-i}}\right]=1-\frac{\beta}{2}.$$

Following the same approach, the upper bound Y_U can be found by solving the equation

$$\int_{Y_U}^{\infty} f^P_{X_{U(n)}|X_{U(m)}}(y|\theta)\,dy=\frac{\beta}{2}.$$

Under the KEL function, the lower and upper bounds can be obtained by solving the following two equations for Y_L and Y_U, respectively.

$$\left[\frac{1}{M}\sum_{j=1}^{M}\left[\frac{e^{\theta_j x_{U(m)}}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i(x_{U(m)})^i\frac{\Gamma(n-m-i,\theta_j Y_L)}{\theta_j^{-i}}\right]^{-1}\right]^{-1}=1-\frac{\beta}{2},$$

and

$$\left[\frac{1}{M}\sum_{j=1}^{M}\left[\frac{e^{\theta_j x_{U(m)}}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i(x_{U(m)})^i\frac{\Gamma(n-m-i,\theta_j Y_U)}{\theta_j^{-i}}\right]^{-1}\right]^{-1}=\frac{\beta}{2}.$$

5. Simulation study and illustrative example

Here, we perform a simulation study based on the exponential distribution with $\theta=2$ $(E(2))$. From $(E(2))$ we generate different sample size cases of records $n=5,7,10$. A sample of size n upper record can be generated using the transformation

$$X_{U(k)}=\frac{\sum_{i=1}^{k}e(i)}{\theta},k=1,2,...,n,$$

where $\{e(i),i\geq 0\}$ is a sequence of iid $E(1)$ [see Arnold *et al.* (1998), p.20].
Using the mean square error (MSE), we investigate the performance of the maximum likelihood estimator (MLE) and the Bayesian estimator of the parameter θ, based on 1000 replications. In Bayesian method, we use an informative prior $\pi_1(\theta)$ of gamma distribution with hyper parameters $a=2$ and $b=1$, based on the two

suggested types of error loss functions SEL and KEL, to estimate the unknown parameter. The MCMC samples, which are used in computing the Bayes estimates, are generated from gamma distribution with hyper parameters $a = 2$ and $b = 1$. We compare the performance of the Bayes estimates of θ under two different priors for θ; the noninformative prior $(a = b = 0)$ (Prior 0), where the prior density becomes improper and not specifically related to the gamma density, and the informative prior $(a = 2, b = 1)$ (Prior 1). In the prediction process, we consider only the one sample prediction problem to find the predictors of the future records as well as the predicted intervals of the predictors based on the informative prior under the suggested loss functions SEL and KEL for all cases of records $n = 5, 7, 10$.

In Tables 1 and 2, we present the MLEs and the Bayes estimators of θ under SEL and KEL functions when Prior 0 and 1 are used.

Table 1. MLEs and Bayes estimators when $\theta = 1$, MSEs are reported in parentheses.

		Prior 0		Prior 1	
Cases	MLE	SEL	KEL	SEL	KEL
$n = 5$	1.2268	1.2268	0.9814	1.3337	1.1432
	(0.6233)	(0.6233)	(0.3664)	(0.3801)	(0.2180)
$n = 7$	1.1530	1.1530	0.9883	1.2439	1.1057
	(0.2605)	(0.2605)	(0.1743)	(0.2397)	(0.1535)
$n = 10$	1.1067	1.1067	0.9960	1.1731	1.0754
	(0.1795)	(0.1795)	(0.1362)	(0.1692)	(0.1227)

Table 2. MLEs and Bayes estimators when $\theta = 2$, MSEs are reported in parentheses.

		Prior 0		Prior 1	
Cases	MLE	SEL	KEL	SEL	KEL
$n = 5$	2.3990	2.3990	1.9192	2.1909	1.8779
	(1.7493)	(1.7493)	(1.0242)	(0.5327)	(0.3795)
$n = 7$	2.3899	2.3899	2.0484	2.2062	1.9611
	(1.2336)	(1.2336)	(0.7970)	(0.4922)	(0.3568)
$n = 10$	2.2073	2.2073	1.9866	2.1596	1.9796
	(0.5432)	(0.5432)	(0.4054)	(0.4094)	(0.3230)

It can be observed from Tables 1 and 2 that the Bayes estimators show superior behaviour over the MLEs of θ as these estimates provide smaller MSEs. Furthermore, it is evident that the Bayes estimators obtained under Prior 1 compete quite well with those obtained under Prior 0 in terms of the MSE criterion. It can be also noted, as expected, that the MSEs tend to be smaller as the number of observed records increases.

In Table 3, we present the average credible interval length (AL) and coverage probability (CP) for the 95% confidence interval when $\theta = 2$ under SEL and KEL for $n = 5, 7, 10$. Table 4 contains different percentiles of the generated value of θ, which are basically generated when $\theta = 2$ as an initial value.

Table 3. AL and CP for the 95% confidence intervals when $\theta = 2$.

Cases		Prior 0		Prior 1	
		AL	CP	AL	CP
$n = 5$	θ	6.1430	0.93	4.1050	0.96
$n = 7$	θ	4.7036	0.95	3.6353	0.96
$n = 10$	θ	3.8140	0.94	3.1574	0.95

Table 4. Percentiles for the generated values of θ.

	0.005	0.025	0.05	0.5	0.95	0.975	0.995
Percentiles	0.3685	0.6226	0.7772	1.9565	3.7570	4.1501	5.046

It can be noticed from Table 3 that the ALs of the 95% confidence intervals are better when using Prior 1 than that when using Prior 0, and that the ALs decrease as the number of observed records increases.

Table 5 contains the predicted values and the corresponding 95% predicted intervals for the future record $X_{U(n)}$, $1 \leq m < n$ based on observed records under SEL and KEL functions when Prior 1 is used.

Table 5. Predicted values and the corresponding 95% predicted intervals for $X_{U(n)}$, $1 \leq m < n$ under SEL and KEL functions when Prior 1 is used.

Cases		SEL		KEL	
	$X_{U(n)}$	Predicted value	95% predicted interval	Predicted value	95% predicted interval
$n = 5$	$X_{U(6)}$	0.5595	$(0.3363, 1.2840)$	0.4883	$(0.3362, 0.8862)$
	$X_{U(7)}$	0.7878	$(0.3761, 1.8599)$	0.6587	$(0.3757, 1.1092)$
	$X_{U(8)}$	1.0160	$(0.4413, 2.3957)$	0.8371	$(0.4402, 1.2898)$
$n = 7$	$X_{U(8)}$	2.7196	$(2.3172, 3.9666)$	2.6615	$(2.3171, 3.4318)$
	$X_{U(9)}$	3.1313	$(2.3937, 4.9242)$	3.0183	$(2.3933, 3.9143)$
	$X_{U(10)}$	3.5431	$(2.5208, 5.8033)$	3.3774	$(2.5191, 4.3131)$
$n = 10$	$X_{U(11)}$	2.6875	$(2.3913, 3.5917)$	2.6546	$(2.3912, 3.2473)$
	$X_{U(12)}$	2.9908	$(2.4490, 4.2793)$	2.9258	$(2.4487, 3.6192)$
	$X_{U(13)}$	3.2940	$(2.5454, 4.9082)$	3.1978	$(2.5444, 3.9231)$

We can observe from Table 5 that all predicted values are located within the predicted intervals and it is worth to note that the predicted intervals get to be wider as the values n increases, i.e. when we try to predict future values that are much wider than the observed data.

Example (real data):

To illustrate the results of this work thus obtained, we analyse the real data of times (in minutes) to breakdown of an insulating fluid between electrodes at voltage 34 kv. These data are originally reported in Lawless (1982, Table 1.1, p.3). The complete data set consists of 19 times to breakdown: 0.96, 4.15, 0.19, 0.78, 8.01, 31.75, 7.35, 6.50, 8.27, 33.91, 32.52, 3.16, 4.85, 2.78, 4.67, 1.31, 12.06, 36.71, 72.89 and this involves a substantial extrapolation from the exponential data. From these data, we extract $n = 7$ upper record values which are: 0.96, 4.15, 8.01, 31.75,

33.91, 36.71, 72.89. Table 6 contains the ML estimator, the Bayes estimator of θ under the SEL and KEL functions when Prior 0 and 1 are used, and the corresponding 95% credible interval of the unknown parameter θ.

Table 6. MLE and Bayes estimates of θ using SEL and KEL functions under Prior 0 and 1.

		Prior 0		Prior 1		
Cases	MLE	SEL	KEL	SEL	KEL	95% credible interval
$n=7$	0.9604	0.9604	0.8232	1.0858	0.9651	$(0.4274, 2.0254)$

Table 7 contains the 8th, 9th and 10th future records, and also their 95% predicted intervals based on the $n=7$ observed upper records.

Table 7. 8th, 9th, 10th future records and their 95% predicted intervals using SEL and KEL functions.

		SEL		KEL	
Cases	$X_{U(n)}$	Predicted value	95% predicted interval	Predicted value	95% predicted interval
$n=7$	$X_{U(8)}$	85.136	$(73.185, 119.869)$	83.616	$(73.185, 111.765)$
	$X_{U(9)}$	97.382	$(75.653, 145.063)$	94.522	$(75.645, 130.039)$
	$X_{U(10)}$	109.628	$(79.830, 167.541)$	105.563	$(79.803, 145.677)$

Based on the previously known records, we find that the MLE of θ to be 0.9604 while the Bayes estimators under Prior 1 are 1.0858, 0.9651 using SEL and KEL, respectively. The 8th, 9th and 10th future records are computed to be 85.136, 97.382, 109.628 under SEL function and 83.616, 94.522, 105.563 under KEL function. The predicted intervals are also computed for all cases and it is found that they include the predicted values.

6. Conclusion

In this work, we have considered the problem of estimating the parameter of exponential distribution and predicting the future (unobserved) records based on an observed set of exponential record data. We have computed the maximum likelihood estimator of the parameter of the exponential distribution and also the Bayes estimator under both SEL and KEL functions. We have computed the MSEs to make a comparison between the MLEs and Bayes estimators. The MCMC samples are used to compute the predictors and the predicted intervals of the future records. Simulation and data analyses are performed to study the behaviour of the proposed methods on estimation and prediction, as well as real data example is presented for illustrative purposes. Based on our study, we recommend the use of KEL function over the well-known SEL function in both estimation and prediction problems depending on the values of the MSEs, which are reported in the previous tables, used in making our comparisons.

Acknowledgements

The authors would like to thank the referees and the editor for their helpful comments and suggestions that contributed to the improvement of this version of the paper.

REFERENCES

ABUFOUDEH, G., BDAIR, O. M., ABU AWWAD, R., (2019). Bayesian Estimation Under Kullback-Leibler Divergence Measure Based on Exponential Data, Investigacion Operacional, 40 (1), pp. 61–72.

AHSANULLAH, M. (1980). Linear Prediction of Record Values for the Two Parameter Exponential Distribution, Annals of the Institute of Statistical Mathematics, 32, pp. 363–368.

AHSANULLAH, M., (1988). Introduction to Record Statistics, Ginn Press, Needham Heights, MA, U.S.A.

AHSANULLAH, M., (1995). Record Statistics, Nova Science Publishers, Commack, NY.

AHSANULLAH, M., KIRMANI, S., (1991). Characterizations of the Exponential Distribution Through a Lower Records, Communications in Statistics - Theory and Methods, 20, pp. 1293–1299.

ARNOLD, B. C., BALAKRISHNAN, N., (1989). Relations, Bounds and Approximations for Order Statistics, Lecture Notes in Statistics, 53, Springer-Verlag, New York.

ARNOLD, B. C., BALAKRISHNAN, N., NAGARAJA, H. N., (1998). Records, Wiley, New York.

BALAKRISHNAN, N., AHSANULLAH, M., CHAN, P.S., (1995). On the Logistic Record Values and Associated Inference, Journal of Applied Statistical Science, 2, pp. 233–248.

BALAKRISHNAN, N., LIN, C. T., CHAN, P.S., (2005). A Comparison of Two Simple Prediction Intervals for Exponential Distribution, IEEE T. Reliab., 54 (1), pp. 27–33.

BDAIR, O. M., RAQAB, M. Z., (2009). On the Mean Residual Waiting Time of Records, Statistics and Decisions, 27 (3), pp. 249–260.

BDAIR, O. M., RAQAB, M. Z., (2012). Sharp Upper Bounds for the Mean Residual Waiting Time of Records, Statistics, 46 (1), pp. 69–84.

BDAIR, O. M., RAQAB, M. Z., (2016). One-Sequence and Two-Sequence Prediction for Future Weibull Records, J. Stat. Theory Appl., 15 (4), pp. 345–366.

BERRED, A. M., (1998). Prediction of Record Values, Communications in Statistics - Theory and Methods, 27, 2221-2240.

CHANDLER, K.N., (1952). The Distribution and Frequency of Record Values, Journal of the Royal Statistical Society - Series B, 14 (2), pp. 220–228.

DUNSMORE, I. R., (1983). The Future Occurrence of Records, Ann. Inst. Statist. Math., 35, pp. 267–277.

JANEEN, Z.F., (2004). Empirical Bayes Analysis of Record Statistics Based on LINEX and Quadratic Loss Functions, Comput. Math. Appl., 47, pp. 947–954.

KAMPS, U., (1995). A Concept of Generalized Order Statistics, J. Statist. Plan. Inference, 48, 1–23.

KULLBACK, S., LEIBLER, R.A. (1951). On Information and Sufficiency, Annals of Mathematical Statistics, 22(1), pp. 79–86.

LAWLESS, J. F., (1982). Statistical Models and Methods for Lifetime Data, 2nd Edition, Wiley, New York.

NASIRI, P., HOSSEINI, S., YARMOHAMMADI, M., (2012). A New Approach to Statistical Inference for Exponential Distribution Based on Record Values, Canadian Journal of Pure and Applied Sciences, 6, pp. 2033–2038.

NEVZOROV, V. B., (2001). Records: Mathematical Theory. Translations of Mathematical Monographs, American Mathematical Society, 194: Providence, RI.

SINGH, S., SINGH, U., SHARMA, V., (2014). Bayesian Estimation and Prediction for the Generalized Lindley Distribution Under Asymmetric Loss Function, Hacettepe Journal of Mathematics and Statistics, 43 (4), pp. 661–678.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 15–32, DOI 10.21307/stattrans-2019-012

GENERALIZED BAYES ESTIMATION OF SPATIAL AUTOREGRESSIVE MODELS

Anoop Chaturvedi[1], **Sandeep Mishra**[2]

ABSTRACT

The spatial autoregressive (SAR) models are widely used in spatial econometrics for analyzing spatial data involving spatial autocorrelation structure. The present paper derives a Generalized Bayes estimator for estimating the parameters of a SAR model. The admissibility and minimaxity properties of the estimator have been discussed. For investigating the finite sample behaviour of the estimator, the results of a simulation study have been presented. The results of the paper are applied to demographic data on total fertility rate for selected Indian states.

Key words: spatial autoregressive model, prior and posterior distributions, generalized Bayes estimator, admissibility and minimaxity; total fertility rate (TFR).

1. Introduction

Spatial data analysis has attracted considerable attention in econometrics literature for modelling data involving spatial dependence. The Spatial Autoregressive (SAR) models assume that the level of response variable depends on the levels of response variable in the neighbouring regions and thus models such spatial spillover effect. Anselin (1988) provided the theoretical aspects of spatial econometrics. Lesage and Pace (2009) discussed various spatial econometric models including SAR model, spatial Durbin model (SDM), and spatial error model (SEM), along with the classical and Bayesian inference procedures for these models and their various applications.

The Bayesian approach involves combining the data distribution embodied in the likelihood function with prior distributions for the parameters assigned by the practitioner, to produce posterior distributions. However, a major drawback of Bayes procedures is lack of robustness with respect to underlying prior assumptions. As mentioned in Berger (1980), the Bayes estimator derived under normal prior has infinite Bayes risk when true prior is Cauchy distribution. One may consider pre-test estimators, but a serious problem with pre-test estimators is that these estimators provide improvement in specific region of parameter space but perform much worse than the usual maximum likelihood estimator

[1] Department of Statistics, University of Allahabad, Allahabad, 211002, India.
E-mail: anoopchaturv@gmail.com. ORCID ID: https://orcid.org/0000-0002-7322-3331.
[2] Department of Statistics, University of Allahabad, Allahabad, 211002, India.
E-mail: sandeepstat24@gmail.com. ORCID ID: https://orcid.org/0000-0001-5631-3354.

(MLE) or least squares estimator outside this region. Rubin (1977) and Berger (1980) demonstrated that the generalized Bayes estimators are a viable alternative to incorporate prior belief and are more robust with respect to underlying prior assumptions. These estimators provide uniform improvement over MLE/least squares estimator and satisfy minimaxity and admissibility properties.

Stein (1973) considered the generalized Bayes estimator for the multivariate normal mean vector under a scale mixture of prior distributions and suggested that the estimator may dominate the James-Stein and positive part James-Stein estimators. Efron and Morris (1976) presented minimax family of estimators for matrix of multivariate normal means in MANOVA model. Berger (1980) provided robust generalized Bayes estimator for multivariate normal mean and obtained confidence region based on generalized Bayes estimator for the mean vector. Brown (1971) derived a powerful condition for the admissibility of generalized Bayes estimators. Berger (1976), and Maruyama (1998) developed classes of admissible minimax generalized Bayes estimators using Brown's (1971) condition. Kubokawa (1991, 1994) showed that the generalized Bayes estimator dominates the usual James-Stein estimator and derived the sufficient dominance condition. Maruyama (1999) considered the extended Stein's prior distribution, which is the scale mixture of multivariate normal distribution, and demonstrated its admissibility and minimaxity. He also showed that the estimator dominates positive part Stein rule estimator. Pal *et al.* (2016) proposed a family of shrinkage estimators for the coefficients vector of a SAR model and investigated its asymptotic properties.

The present paper considers SAR model involving one period lag spatial dependent variable and derives a generalized Bayes estimator for the regression coefficients vector. The admissibility and minimaxity properties of the estimator are investigated. A simulation study has been carried out to assess the finite sample behaviour of the estimator. For illustration purpose, the results of the paper are applied to demographic data on total fertility rate for selected Indian states.

2. The SAR model and estimators

Let us consider the SAR model:

$$y = \rho W y + X\beta + u \ , u \sim N(0, \sigma^2 I_n), \tag{2.1}$$

where y is $(n \times 1)$ vector of the observations on a dependent variable collected at each of *n* locations, X is $(n \times p)$ matrix of observations on exogenous variables, β is $(p \times 1)$ vector of regression parameters, ρ is the spatial autoregressive parameter, W is known $n \times n$ spatial weight matrix which indicate the potential interaction between contiguous positions and has been standardized to have row sum of unity. This model is termed as spatial autoregressive model as it combines the standard regression model with spatially lagged dependent variable.

When ρ is known, the ordinary least squares (OLS) estimator of β is

$$b(\rho) = (X'X)^{-1}X'(y - \rho W y)$$

$$= (X'X)^{-1}X'y(\rho), \tag{2.2}$$

where $y(\rho)=y-\rho Wy$.

The OLS estimator can alternatively be written as

$b(\rho) = b - \rho b_w$.

Here $b = (X'X)^{-1}X'y$ and $b_w = (X'X)^{-1}X'Wy$. Further, the maximum likelihood estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{v}{n}$$

where

$v = \big(y(\rho) - Xb(\rho)\big)'\big(y(\rho) - Xb(\rho)\big) = [y - \rho Wy]'M[y - \rho Wy]$.

Here $M = I_n - X(X'X)^{-1}X'$. When ρ is unknown, we replace it by its estimator

$$\hat{\rho} = \frac{y'W'My}{y'W'MWy} \tag{2.3}$$

in (2.2) to obtain feasible least squares estimator of β as

$$b(\hat{\rho}) = b - \hat{\rho}b_w.$$

Then the estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\hat{v}}{n}, \text{with } \hat{v} = [y - \hat{\rho}Wy]'M[y - \hat{\rho}Wy]. \tag{2.4}$$

3. Generalized Bayes estimator

For obtaining the Generalized Bayes estimator of regression coefficients vector β, let us write the model (2.1) as

$$y(\rho) = X\beta + u, \tag{3.1}$$

where $y(\rho) = y - \rho Wy$. Then the pdf of $y(\rho)$ is given by

$$p(y(\rho)|\beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} exp\left\{-\frac{1}{2\sigma^2}(y(\rho) - X\beta)'(y(\rho) - X\beta)\right\}. \tag{3.2}$$

We assume that β follows a g-prior $N(0, \sigma^2 gX'X)$, (see Zellner, 1986) with $g = \lambda^{-1}(1-\lambda), 0 < \lambda < 1$. Hence the pdf of prior distribution of β is given by

$$p(\beta|\sigma^2, \lambda) \propto \sigma^{-p}\left(\frac{\lambda}{(1-\lambda)}\right)^{p/2} exp\left\{-\frac{\lambda}{2\sigma^2(1-\lambda)}\beta'X'X\beta\right\}. \tag{3.3}$$

We take the prior distribution for λ as

$$p(\lambda) \propto \lambda^{-a}(1-\lambda)^c I_{(0,1)}(\lambda). \tag{3.4}$$

If c > -1, the prior distribution for λ is proper for $a < 1$ and improper for $a \geq 1$. Let us assume σ^2 to be known. The joint density of $(y(\rho), \beta, \lambda)$ is

$p(y(\rho), \beta, \lambda)$

$= p(y(\rho)|\beta, \sigma^2)p(\beta|\sigma^2, \lambda)p(\lambda)$

$$\propto \sigma^{-(n+p)} exp\left\{-\frac{1}{2\sigma^2}(y(\rho)-X\beta)'(y(\rho)-X\beta)\right\}\lambda^{\frac{p}{2}-a}(1-\lambda)^{-\frac{p}{2}+c}$$
$$\times exp\left\{-\frac{\lambda}{2\sigma^2(1-\lambda)}\beta' X'X\beta\right\}$$

$$\propto \sigma^{-(n+p)}\lambda^{\frac{p}{2}-a}(1-\lambda)^{-\frac{p}{2}+c}e^{-v/2\sigma^2} exp\left\{-\frac{1}{2\sigma^2}b(\rho)'X'Xb(\rho)\right\}$$
$$\times exp\left\{-\frac{1}{2\sigma^2(1-\lambda)}[\beta' X'X\beta - 2(1-\lambda)\beta' X'Xb(\rho)]\right\}$$

$$\propto \sigma^{-(n+p)}\lambda^{\frac{p}{2}-a}(1-\lambda)^{-\frac{p}{2}+c}e^{-v/2\sigma^2} exp\left\{-\frac{\lambda}{2\sigma^2}b(\rho)'X'Xb(\rho)\right\}\times$$
$$exp\left\{-\frac{1}{2\sigma^2(1-\lambda)}\left[\left(\beta-(1-\lambda)b(\rho)\right)'X'X\left(\beta-(1-\lambda)b(\rho)\right)\right]\right\}. \quad (3.5)$$

Integrating (3.5) with respect to β, the joint density of $(y(\rho),\lambda)$ is obtained as $p(y(\rho),\lambda)$

$$\propto \sigma^{-(n+p)}\lambda^{\frac{p}{2}-a}(1-\lambda)^{-\frac{p}{2}+c}e^{-v/2\sigma^2} exp\left\{-\frac{\lambda}{2\sigma^2}b(\rho)'X'Xb(\rho)\right\}\times$$

$$\int_{R^p} exp\left\{-\frac{1}{2\sigma^2(1-\lambda)}\left[\left(\beta-(1-\lambda)b(\rho)\right)'X'X\left(\beta-(1-\lambda)b(\rho)\right)\right]\right\}d\beta$$

$$\propto \sigma^{-n}\lambda^{\frac{p}{2}-a}(1-\lambda)^{c}e^{-v/2\sigma^2} exp\left\{-\frac{\lambda}{2\sigma^2}b(\rho)'X'Xb(\rho)\right\}. \quad (3.6)$$

Then the marginal density of $y(\rho)$ is

$$m\left(y(\rho)\right) \propto \sigma^{-n}e^{-v/2\sigma^2}\int_0^1 \lambda^{\frac{p}{2}-a}(1-\lambda)^{c} exp\left\{-\frac{\lambda}{2\sigma^2}b(\rho)'X'Xb(\rho)\right\}d\lambda. \quad (3.7)$$

Further, the posterior expectation of λ given $y(\rho)$ is obtained as

$$E\left(\lambda|y(\rho)\right) = \frac{\int_0^1 \lambda^{\frac{p}{2}-a+1}(1-\lambda)^{c} exp\left\{-\frac{\lambda}{2\sigma^2}b(\rho)'X'Xb(\rho)\right\}d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-a}(1-\lambda)^{c} exp\left\{-\frac{\lambda}{2\sigma^2}b(\rho)'X'Xb(\rho)\right\}d\lambda}$$
$$= \phi_{a,c}\left(\frac{b(\rho)'X'Xb(\rho)}{\sigma^2}\right), \quad (3.8)$$

where

$$\phi_{a,c}\left(\frac{b(\rho)'X'Xb(\rho)}{\sigma^2}\right)$$
$$= \frac{\int_0^1 \lambda^{\frac{p}{2}-a+1}(1-\lambda)^{c} exp\left\{-\frac{\lambda}{2\sigma^2}b(\rho)'X'Xb(\rho)\right\}d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-a}(1-\lambda)^{c} exp\left\{-\frac{\lambda}{2\sigma^2}b(\rho)'X'Xb(\rho)\right\}d\lambda}$$
$$= \frac{\Gamma\left[2-a+\frac{p}{2}\right]\Gamma\left[2-a+c+\frac{p}{2}\right]\ {}_1F_1\left[2-a+\frac{p}{2},3-a+c+\frac{p}{2},-\frac{b(\rho)'X'Xb(\rho)}{2\sigma^2}\right]}{\Gamma\left[1-a+\frac{p}{2}\right]\Gamma\left[3-a+c+\frac{p}{2}\right]\ {}_1F_1\left[1-a+\frac{p}{2},2-a+c+\frac{p}{2},-\frac{b(\rho)'X'Xb(\rho)}{2\sigma^2}\right]}. \quad (3.9)$$

The Kummer confluent hypergeometric function ${}_1F_1[a;c;z]$ used in expression (3.9) is defined as

$$ {}_1F_1[a;c;z] = \sum_{k=0}^{\infty} \frac{(a)_k}{(c)_k} \frac{z^k}{k!}; $$

where $(a)_0 = 1, (a)_k = a(a+1)\ldots(a+k-1)$, is the rising factorial.

Then, the generalized Bayes estimator of β is

$$ \hat{\beta}(\rho) = \left[1 - \phi_{a,c}\left(\frac{b(\rho)'X'Xb(\rho)}{\sigma^2}\right)\right] b(\rho). \tag{3.10} $$

If we substitute

$$ \phi_h(w) = w\frac{\int_0^1 \lambda^{\frac{p}{2}+1-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}} d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}} d\lambda} = w\phi_{a,c}\left(\frac{b(\rho)'X'Xb(\rho)}{\sigma^2}\right), $$

then the generalized Bayes estimator $\hat{\beta}(\rho)$ can be represented as

$$ \hat{\beta}(\rho) = \left[1 - \frac{\sigma^2}{b(\rho)'X'Xb(\rho)}\phi_h\left(\frac{b(\rho)'X'Xb(\rho)}{\sigma^2}\right)\right] b(\rho). $$

Theorem 1: *Under the loss function*

$$ L(\hat{\beta},\beta) = \frac{1}{\sigma^2}(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta) \tag{3.11} $$

the GB estimator $\hat{\beta}(\rho)$ has finite risk.

Proof: Let us write

$$ Z = \frac{1}{\sigma}(X'X)^{\frac{1}{2}}b(\rho);\ \theta = \frac{1}{\sigma}(X'X)^{\frac{1}{2}}\beta. $$

Then

$$ R[\hat{\beta}(\rho),\beta] = \frac{1}{\sigma^2}E\left[(\hat{\beta}(\rho)-\beta)'X'X(\hat{\beta}(\rho)-\beta)\right] $$
$$ = E\left[(Z-\theta)'(Z-\theta) + \frac{1}{\|Z\|^2}{\phi_h}^2(\|Z\|^2) - 2\frac{(Z-\theta)'Z\phi_h(\|Z\|^2)}{\|Z\|^2}\right]. $$

Since $Z \sim N(\theta, I_p)$, we have

$$ R[\hat{\beta}(\rho),\beta] = p + E\left[\frac{1}{\|Z\|^2}{\phi_h}^2(\|Z\|^2) - 2\frac{(Z-\theta)'Z\phi_h(\|Z\|^2)}{\|Z\|^2}\right]. $$

We observe that $0 \le \phi_h(w) \le w$, so that

$$ E\left[\frac{1}{\|Z\|^2}{\phi_h}^2(\|Z\|^2)\right] \le E[\|Z\|^2] = p + \theta'\theta < \infty. $$

Further by Schwarz's inequality

$$ E\left[\frac{(Z-\theta)'Z\phi_h(\|Z\|^2)}{\|Z\|^2}\right] \le \left[E(Z-\theta)'(Z-\theta)E\left\{\frac{\phi_h(\|Z\|^2)^2}{\|Z\|^2}\right\}\right]^{1/2} $$

$$\leq \left[pE[\|Z\|^2]\right]^{1/2}$$
$$= [p(p+\theta'\theta)]^2 < \infty.$$

Hence the risk of $\hat{\beta}(\rho)$ is finite■

Theorem 2: The GB estimator is admissible if and only if $a \leq 2$.

Proof: We have

$$f_R(\|z\|^2) = \int_0^1 e^{-\lambda\|z\|^2/2}\lambda^{\frac{p}{2}-a}(1-\lambda)^c d\lambda$$
$$= 2^{\frac{p}{2}-a+1}\int_0^{\frac{1}{2}} e^{-t\|z\|^2}\, t^{\frac{p}{2}-a}(1-2t)^c dt$$
$$= 2^{\frac{p}{2}-a+1}\int_0^{\infty} e^{-t\|z\|^2}\, t^{\frac{p}{2}-a}(1-2t)^c I_{\left(0,\frac{1}{2}\right)}(t)dt.$$

Using Tauberlian theorem (see Maruyama, 2000, p. 37), we observe that as $t \to 0$, $t^{\frac{p}{2}-a}(1-2t)^c I_{\left(0,\frac{1}{2}\right)}(t) \sim t^{\frac{p}{2}-a}$. Hence we have

$$f_h(\|z\|^2) \sim 2^{\frac{p}{2}-a+1}\Gamma\left(\frac{p}{2}-a+1\right)\|z\|^{-2\left(\frac{p}{2}-a+1\right)}. \tag{3.12}$$

Following Maruyama (2000), to show that the GB estimator is admissible it is necessary and sufficient to show that $\int_1^{\infty} f_h^{-1}(t)\, t^{\frac{p}{2}}dt$ diverges. Using equation (3.12), we have

$$\int_1^{\infty} f_h^{-1}(t)\, t^{\frac{p}{2}}dt \sim 2^{\frac{p}{2}-a+1}\Gamma^{-1}\left(\frac{p}{2}-a+1\right)\int_1^{\infty} t^{(-a+1)}\, dt,$$

which diverges as long as $a \leq 2$. This leads to the required result■

Theorem 3: The generalized Bayes estimator $\hat{\beta}(\rho)$ *is minimax whenever* $3-\frac{p}{2} \leq a \leq \frac{p}{2}+1$.

Proof: Under the loss function (3.11) the difference between the risks of GB estimator $\hat{\beta}(\rho)$ and the OLS estimator $b(\rho)$ is given by

$$R\left[\hat{\beta}(\rho),\beta\right] - R[b(\rho),\beta] = E\left[\frac{1}{\|Z\|^2}\Phi^2{}_h(\|Z\|^2) - 2\frac{(Z-\theta)'Z\phi_h(\|Z\|^2)}{\|Z\|^2}\right].$$

Now

$$E\left[(Z-\theta)'Z\frac{\phi_h(\|Z\|^2)}{\|Z\|^2}\right] = E\left[\frac{\partial}{\partial Z'}\left\{Z\frac{\phi_h(\|Z\|^2)}{\|Z\|^2}\right\}\right]$$
$$= E\left[p\frac{\phi_h(\|Z\|^2)}{\|Z\|^2} - 2\frac{\phi_h(\|Z\|^2)}{\|Z\|^2} + 2\phi'{}_h(\|Z\|^2)\right].$$

Hence

$R\left[\hat{\beta}(\rho),\beta\right] - R[b(\rho),\beta]$

$$= E\left[\frac{1}{\|Z\|^2}\phi^2{}_h(\|Z\|^2) - 2(p-2)\frac{\phi_h(\|Z\|^2)}{\|Z\|^2} - 4\phi'{}_h(\|Z\|^2)\right] \tag{3.13}$$

Now we have

$$\frac{\phi_h(w)}{w} = \frac{\int_0^1 \lambda^{\frac{p}{2}+1-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}}d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}}d\lambda},$$

so that

$$\frac{\partial}{\partial w}\left[\frac{\phi_h(w)}{w}\right] = \frac{1}{2}\frac{\left\{\int_0^1 \lambda^{\frac{p}{2}+1-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}}d\lambda\right\}^2 - \left\{\int_0^1 \lambda^{\frac{p}{2}+2-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}}d\lambda\right\}\left\{\int_0^1 \lambda^{\frac{p}{2}-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}}d\lambda\right\}}{\left\{\int_0^1 \lambda^{\frac{p}{2}-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}}d\lambda\right\}^2} \tag{3.14}$$

Let us write

$$f_h(\lambda) = \frac{\lambda^{\frac{p}{2}-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}}}{\int_0^1 \lambda^{\frac{p}{2}-a}(1-\lambda)^c e^{-\frac{\lambda w}{2}}d\lambda}, 0 < \lambda < 1.$$

Then

$$\frac{\partial}{\partial w}\left[\frac{\phi_h(w)}{w}\right] = -\frac{1}{2}\left[E_{f_h(\lambda)}(\lambda^2) - \left\{E_{f_h(\lambda)}(\lambda)\right\}^2\right] \leq 0.$$

Again

$$\begin{aligned}
&\phi'{}_h(w) = \frac{\partial}{\partial w}\left\{w\frac{\phi_h(w)}{w}\right\} \\
&= \frac{\phi_h(w)}{w} + w\frac{\partial}{\partial w}\left\{\frac{\phi_h(w)}{w}\right\} \\
&= E_{f_h(\lambda)}(\lambda) - \frac{1}{2}E_{f_h(\lambda)}(\lambda^2) + \frac{1}{2}\left\{E_{f_h(\lambda)}(\lambda)\right\}^2 \\
&\geq \frac{1}{2}\left\{E_{f_h(\lambda)}(\lambda)\right\}^2 \geq 0.
\end{aligned}$$

Notice that $0 \leq \lambda \leq 1$, so that $E_{f_h(\lambda)}(\lambda) - \frac{1}{2}E_{f_h(\lambda)}(\lambda^2) \geq 0$.

We also observe that $\phi_h(w)$ and $\frac{\phi_h(w)}{w}$ are monotone in opposite directions. Therefore we obtain

$$\begin{aligned}
&R\left[\hat{\beta}(\rho), \beta\right] - R[b(\rho), \beta] \\
&\leq E\left[\frac{\phi_h(\|Z\|^2)}{\|Z\|^2}\right]E[\phi_h(\|Z\|^2) - 2(p-2)] - 4E\left[\phi'{}_h(\|Z\|^2)\right] \\
&\leq E\left[\frac{\phi_h(w)}{w}\right]E[\{\phi_h(w) - 2(p-2)\}],
\end{aligned}$$

which is less than or equal to zero whenever

$$0 \leq E[\{\phi_h(w)\}] \leq 2(p-2). \tag{3.15}$$

When w is large, we may approximate $\phi_h(w)$ as

$$\phi_h(w) \approx 2\left(\frac{p}{2} - a + 1\right).$$

Further $\phi_h(w)$ is an increasing function of w. Hence, a sufficient dominance condition is

$$0 \leq 2\left(\frac{p}{2} - a + 1\right) \leq 2(p-2),$$

or

$$3 - \tfrac{p}{2} \leq a \leq \tfrac{p}{2} + 1.$$

This leads to the required result■

When ρ and σ^2 are unknown, we replace them by their estimators $\hat{\rho}$ and $\hat{\sigma}^2$ defined in (2.3) and (2.4) respectively to obtain feasible generalized Bayes estimator of β.

4. Simulation study

In this section we carry out a simulation study using R Software to assess the finite sample behaviour of proposed generalized Bayes estimator. The observations on response variable *y* are generated by using the model (2.1). In simulation study we compare the risks of the usual feasible least squares estimator $b(\hat{\rho}) = (X'X)^{-1}X'(y - \hat{\rho}Wy)$ with the following feasible version of GB estimator:

$$\hat{\beta}(\hat{\rho}) = \left[1 - \phi_{a,c}\left(\frac{b(\hat{\rho})'X'Xb(\hat{\rho})}{\hat{\sigma}^2}\right)\right] b(\hat{\rho}).$$

The matrix *X* has been generated from multivariate normal distribution MVN[(1, 3, 5, 4, 7, 5, 6, 4, 7, 4), diag(0, 1.6, 0.7, 3.2, 1.5, 1, 2.8, 2, 1.4, 2.2)]. In the weight matrix *W,* the weights assigned to nearest neighbour values, say $(w_1, w_2), \ldots, (w_{n-1}, w_n)$, are double the weights assigned to the second nearest neighbour values, say, $(w_1, w_3), \ldots, (w_{n-2}, w_n)$ and other neighbour weights are taken as zero. The property of weight matrix to be row stochastic is also satisfied. Further, to ensure the stationarity, the values of ρ are selected in the range $\left(\frac{1}{W_{max}}, \frac{1}{W_{min}}\right)$, where W_{max} and W_{min} are, respectively, the maximum and minimum eigen values of *W*. We select *c* = 1, *a* = 0.5 and the results are depicted in figures 1-6. Figures 1 and 2 plot the percentage gain in efficiency of GB estimator over feasible least squares estimator when we vary ρ in the range (-0.95, 0.95). For p=5, $\beta'\beta$=1.525 and for p=10 $\beta'\beta$=1.8819. Further figures 3-6 plot percentage gain in efficiency for variation in $\beta'\beta$ and fixed ρ, *n*, and *p*. The selected values of ρ in figures 3-6 are 0.25 and 0.75, selected values of *n* are 20, 50, 100, 200 and those of *p* are 5 and 10. For each setting of parameters, the experiment is replicated 5000 times. We have used maximum likelihood estimator of ρ for evaluating feasible least squares and feasible GB estimators. The percentage

gain in efficiency due to feasible GB estimator $\hat{\beta}(\hat{\rho})$ over feasible least squares estimator $b(\hat{\rho})$ is calculated using formula:

$$\% \text{ gain in efficiency } GE(\hat{\beta}(\hat{\rho})) = \frac{ER(b(\hat{\rho}))-ER(\hat{\beta}(\hat{\rho}))}{ER(b(\hat{\rho}))} \times 100.$$

Empirical risk of the estimator $\hat{\beta}(\hat{\rho})$ based on 5000 replications has been evaluated as

$$ER(\hat{\beta}(\hat{\rho})) = E\left(\left(\hat{\beta}(\hat{\rho}) - \beta\right)'\left(\hat{\beta}(\hat{\rho}) - \beta\right)\right)$$
$$\approx \frac{1}{5000}\sum_{r=1}^{5000}\left(\hat{\beta}(\hat{\rho})_r - \beta\right)'\left(\hat{\beta}(\hat{\rho})_r - \beta\right),$$

where $\hat{\beta}(\hat{\rho})_r$ is the estimated β based on r-th replication.

The main findings of the simulation are as follows:

1. GB estimator performs better than the FLS estimator, in all the selected parametric settings.
2. From Figures 1 and 2 we observe that the percentage gain in efficiency remains almost constant for $\rho < 0$ and then it starts increasing gradually except for *n*= 20, where it increases for $\rho > -0.25$.
3. For *n*=20, *p*=5 and *n*=20, *p*=10, the gain in efficiency is maximum when ρ is close to 0.6 and then again it starts decreasing with increasing ρ.
4. For *n*= 50, *p*= 5, the gain in efficiency increases for $\rho > 0.25$ and for *n*= 50, *p*= 10, it increases for $\rho > 0.5$.
5. For *n*= 100, *p*= 5, the gain in efficiency increases for $\rho > 0.35$ and for *n*= 100, *p*= 10 it increases for $\rho > 0.5$.
6. For *n*= 200, both for *p*= 5, and *p*=10, the gain in efficiency usually keeps on increasing for $\rho > 0.5$.
7. For fixed *n* the gain in efficiency decreases as *p* increases from 5 to 10.
8. Figures 3-6 show that the percentage gain in efficiency increases as the value of ρ increases from 0.25 to 0.75. The gain in efficiency decreases with increasing $\beta'\beta$.
9. For *n*= 20, *p*= 5, 10, $\rho = 0.25$ the percentage gain in efficiency is almost constant for $\beta'\beta > 9$. For $\rho = 0.75$ it gradually decreases with increasing $\beta'\beta$ up to $\beta'\beta = 20$ and, after that, it remains almost constant. For all others combination of parameters the gain in efficiency remains almost constant as long as $\beta'\beta > 3$.

5. Application to TFR Data

In this section we present an application of SAR model for modelling the total fertility rates (TFR) of selected Indian states. We use the causal variables Female literacy rate (FLIT), Headcount poverty ratio (HCPR), and Percentage of urban population (PUP), which control the socio economic conditions influencing TFR, see table 4. For incorporating the influence of spatial structure of states in India, first order spatial autoregressive term is also included. The spatial weight matrix is

formed using spatial contiguity matrix. To form the contiguity matrix, define Vij =1 for two spatial units (states in our example) that own a common border of non-zero length, else equal to zero. Since an element is not contiguous or neighbouring to itself, the main diagonal elements of the matrix are zero. The matrix **V** is then scaled to make it row stochastic. Denoting such a standardized first order contiguity matrix by W, its *(i,j)*-th element, say W_{ij} is given by

$$W_{ij} = \frac{{w_{ij}}'}{\sum_{\substack{j=1 \\ i \neq j}}^{n} {w_{ij}}'} \tag{2.4}$$

where ${W_{ij}}' = 1$ if i is linked to j, and 0 otherwise. Moran's I statistic for W is $I = z'Wz/z'z$, where z is $n \times 1$ vector of variables expressed as deviations from the mean. The global Moran's *I* statistic is used to examine the variables in our data set for global autocorrelation. If the observed value of *I* is greater than its expected value, then corresponding observation tend to be surrounded by neighbours with similar values. On the other hand if *I* is less than its expected value, the observation tend to be surrounded by dissimilar values, see Schabenberger and Gotway (2005) for details.

The regression coefficients of fitted SAR model are estimated using feasible LS and feasible GB estimators. In sample predicted values of TFR for different states are also computed based on both the estimators, see table 5. Table 6 gives the estimated coefficients using both of these estimators. The estimated values of spatial autocorrelation coefficient is 0.5923. Table 7 gives the observed and expected value of the Moran's *I* for each of the variables considered in the analysis. We observe that HCPR shows the highest degree of spatial correlation, followed by the FLIT while the PUP shows the lowest degree of spatial autocorrelation among the independent variables. The results from the empirical investigation indicate that the feasible GB estimator performs better than FLS estimator of regression coefficients in terms of predictive efficiency.

6. Concluding remarks

With the objective of achieving robustness with respect to prior distribution and satisfying admissibility and minimaxity properties, we have developed a family of generalized Bayes estimators for the regression coefficients vector of a SAR model. The simulation study has been carried out to examine the efficiency properties of GB estimator and it was observed that GB estimator provides improvement over the usual least squares estimator for a wide range of the parametric settings.

REFERENCES

ANSELIN, L., (1988). Spatial econometrics, methods, and models, Dordrecht: Kluwer Academic.

ANSELIN, L., REY, S., (2010). Perspectives on spatial data analysis, Berlin: Springer, DOI:10.1007/978-3-642-01976-0.

BERGER, J., (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. Ann. Statist., 4, pp. 223–226.

BERGER, J. O., (1980): Statistical Decision Theory: Foundations, Concepts and Methods, Springer, N.Y.

BROWN, L. D., (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems, Ann. Math. Statist., 42, pp. 855–903.

EFRON, B., MORRIS, C., (1976). Families of minimax estimators of the mean of a multivariate normal distribution. Ann. Statist., 4, pp. 11–21.

FOURDRINIER, D., STRAWDERMAN. W. E., WELLS, T., (1998). On the construction of Bayes minimax estimators, Ann. Statist., 26, pp. 660–671.

KUBOKAWA, T., (1991). An approach to improving the James-Stein estimator, J. Multivariate Anal., 36, pp. 121–126.

KUBOKAWA, T., (1994). An unified approach to improving equivariant estimators, Ann. Statist., 22, pp. 290–299.

LESAGE, J. P., (1998): Spatial Econometrics, www.spatial-econometrics.com.

LESAGE, J. P, PACE, R. K., (2009). Introduction to spatial econometrics, Boca Raton, FL: CRC, DOI:10.1201/9781420064254.

MARUYAMA, Y., (1998). A unified and broadened class of admissible minimax estimators of a multivariate normal mean, J. Multivariate Anal., 64, 196–205.

MARUYAMA, Y., (1999). Improving on the James-Stein Estimator, Statistics & Decisions, 17, pp. 137–140.

MARUYAMA, Y., (2000). Minimax admissible estimation of a multivariate normal mean and improvement upon the James-Stein estimator, Ph.D. dissertation, Graduate School of Economics, University of Tokyo.

PAL, A., CHATURVEDI, A., DUBEY, A., (2016). Shrinkage estimation in spatial autoregressive model, Journal of Multivariate Analysis, 143, pp. 362–373.

RUBIN, H. (1977). Robust Bayesian estimation, In Statistical Decision Theov and Related Topics II. (S. S. Gupta and D. S. Moore, eds.), Academic Press.

SCHABENBERGER, O., GOTWAY, C. A., (2005): Statistical Methods for Spatial Data Analysis, Chapman and Hall/CRC: Boca Raton, FL.

STEIN, C., (1973). Estimation of the mean of a multivariate normal distribution, In Proc. Prague Symp. Asymptotic Statist., pp. 345–381.

ZELLNER, A., (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions, In: Goel, P. and Zellner, A., Eds., Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, Elsevier Science Publishers, Inc., New York, pp. 233–243.

APPENDIX

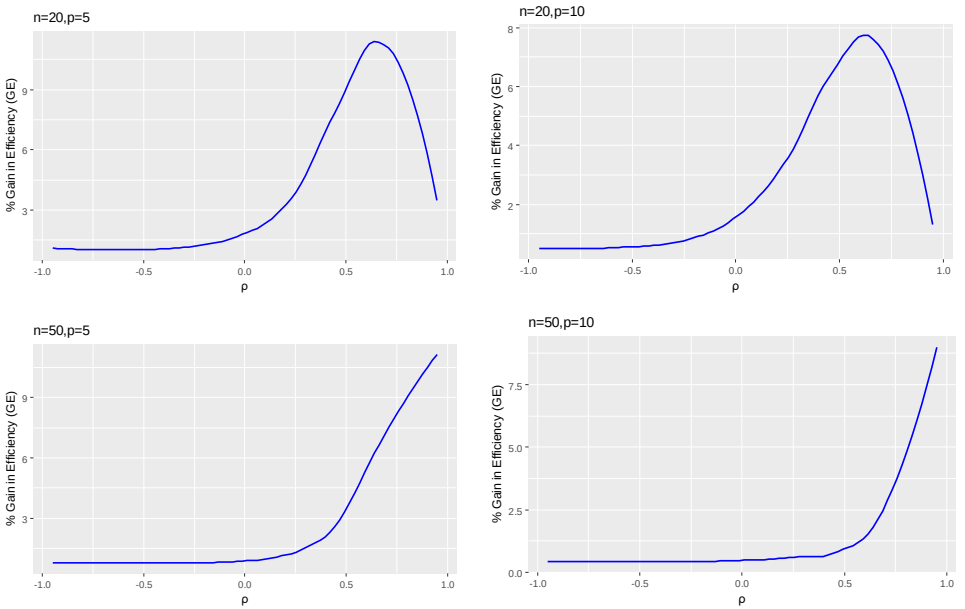

Figure 1. Percentage Gain in efficiency due to change in ρ

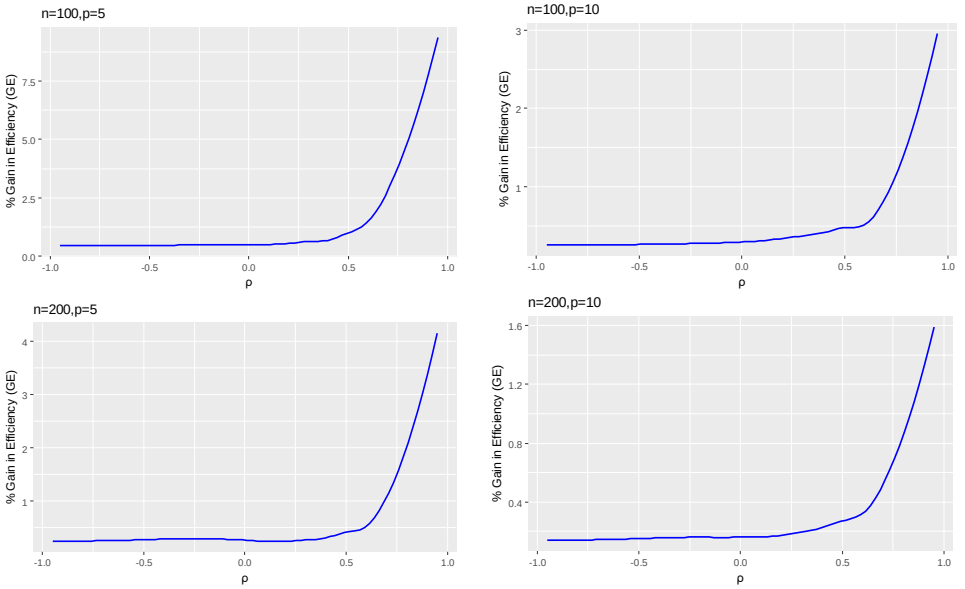

Figure 2. Percentage Gain in efficiency due to change in ρ

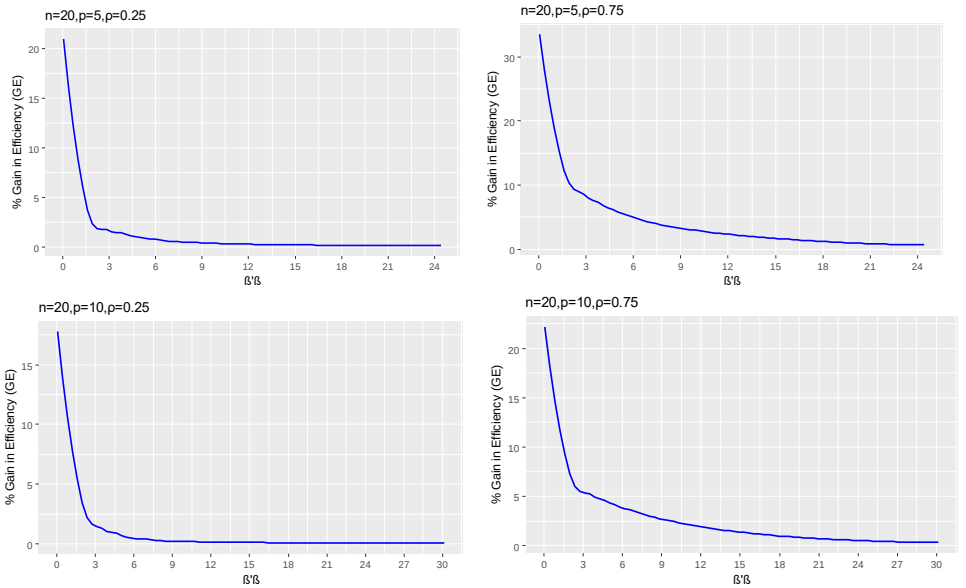

Figure 3. Percentage gain in efficiency due to change in length of parameter β *i.e.* $\beta'\beta$

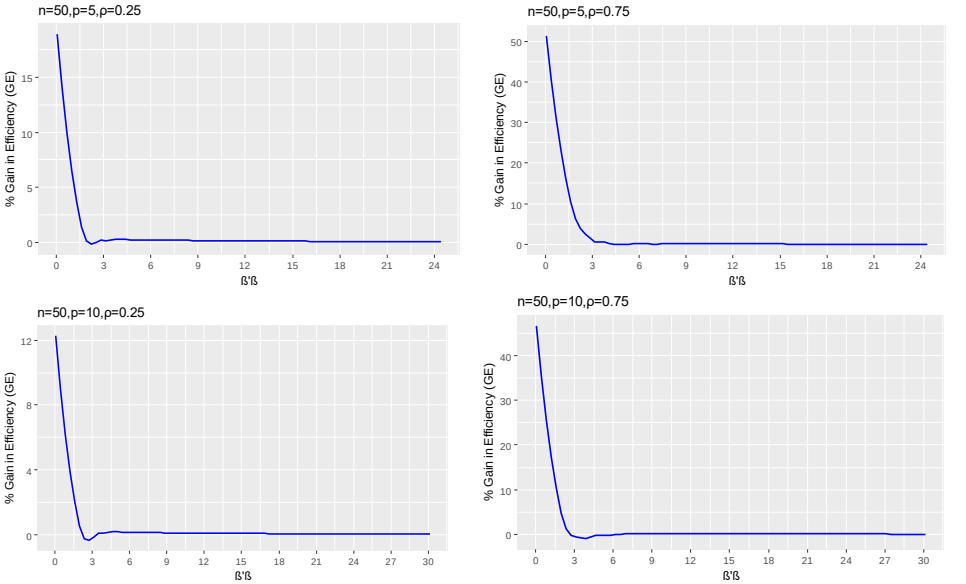

Figure 4. Percentage gain in efficiency due to change in length of parameter β *i.e.* $\beta'\beta$

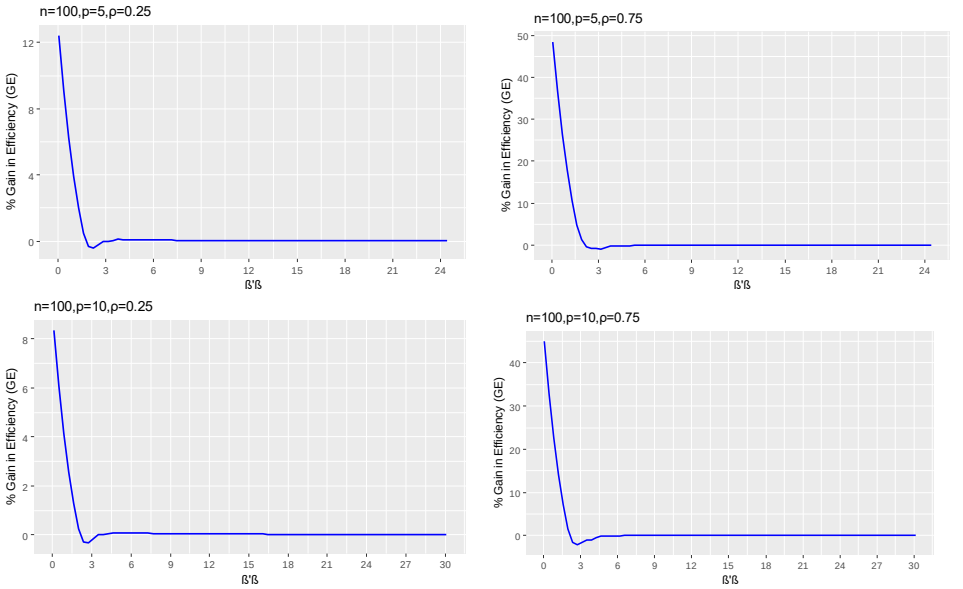

Figure 5. Percentage gain in efficiency due to change in length of parameter β *i.e.* $\beta'\beta$

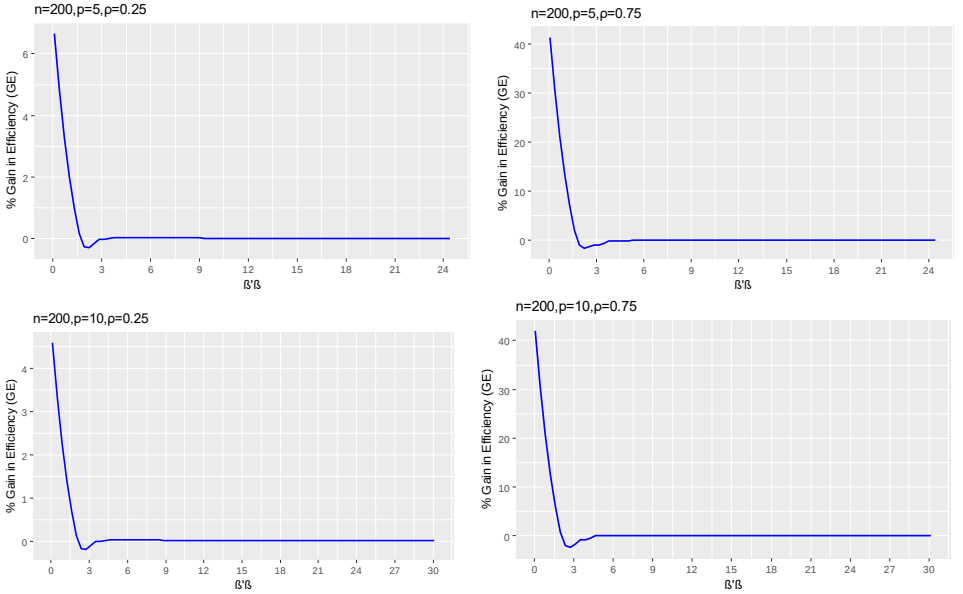

Figure 6. Percentage gain in efficiency due to change in length of parameter β *i.e.* $\beta'\beta$

Table 1. Percentage gain in efficiency due to $\beta'\beta$ for p = 5

ρ	$\beta'\beta$	*n=20*	*n=50*	*n=100*	*n=200*
0.25	0.244	13.5009	11.08661	6.037487	2.837103
	0.976	5.328835	2.463001	1.107283	0.476686
	2.196	2.738421	0.815412	0.357004	0.135803
	3.904	1.454371	0.38847	0.158998	0.048988
	6.1	0.778182	0.226422	0.086162	0.019632
	8.784	0.459649	0.148478	0.052888	0.008067
	11.956	0.303455	0.104995	0.03536	0.002975
	15.616	0.216832	0.078216	0.025173	0.000648
	19.764	0.163445	0.060558	0.01869	-0.00044
	24.4	0.127986	0.048275	0.014404	-0.00647
0.75	0.244	26.1813	42.82635	40.14246	31.70275
	0.976	15.00917	17.25584	8.275401	3.31876
	2.196	10.36237	4.615964	1.276183	0.490065
	3.904	7.243892	1.117806	0.354763	0.144584
	6.1	4.940427	0.386443	0.140805	0.054013
	8.784	3.411094	0.190456	0.064894	0.020024
	11.956	2.426919	0.110726	0.031528	0.004777
	15.616	1.603265	0.071166	0.015088	-0.0027
	19.764	1.035082	0.049153	0.00642	-0.00648
	24.4	0.67636	0.035669	0.001649	-0.00831

Table 2. Percentage gain in efficiency due to $\beta'\beta$ for p = 10

ρ	$\beta'\beta$	*n=20*	*n=50*	*n=100*	*n=200*
0.25	0.301104	12.03765	6.177598	3.646498	1.938245
	1.204416	5.176314	1.081055	0.63606	0.337801
	2.709936	2.276713	0.405873	0.226304	0.110739
	4.817664	0.912911	0.213313	0.112668	0.049924
	7.527600	0.387297	0.131917	0.067129	0.027217
	10.83974	0.213107	0.089933	0.044548	0.016788
	14.7541	0.137561	0.065157	0.031726	0.011300
	19.27066	0.097152	0.049539	0.023766	0.008085
	24.38942	0.072792	0.038788	0.018471	0.006074
	30.1104	0.056718	0.031372	0.014747	0.004733
0.75	0.301104	15.911	38.03073	33.88339	28.86204
	1.204416	8.84946	8.273935	2.498556	1.544655
	2.709936	6.279212	1.14234	0.438372	0.262278
	4.817664	4.539674	0.360881	0.16237	0.092256
	7.5276	3.2402	0.174061	0.079288	0.042532
	10.83974	2.208562	0.102341	0.044992	0.022133
	14.7541	1.409047	0.067498	0.02815	0.012272
	19.27066	0.867744	0.047913	0.018992	0.00696
	24.38942	0.524181	0.035872	0.01348	0.003907
	30.1104	0.299729	0.027958	0.010022	0.002106

Table 3. Percentage gain in efficiency due to ρ

	ρ	*n=20*	*n=50*	*n=100*	*n=200*
p=5	-0.95	1.091354	0.7822986	0.4439748	0.2370162
	-0.75	1.038856	0.7635631	0.4476627	0.2456971
	-0.55	1.030078	0.7567423	0.4588296	0.2617378
	-0.35	1.105928	0.7634651	0.4737263	0.2834117
	0.05	2.060012	0.9269042	0.4977228	0.2475341
	0.25	3.778747	1.3376279	0.5968474	0.2441839
	0.45	7.496059	2.5998401	0.9269096	0.3940382
	0.65	12.3777	6.0657974	1.8282938	0.7344557
	0.75	12.3601	9.2943542	3.0362531	1.1202461
	0.95	2.127767	9.4626555	9.7513806	4.6275375
p=10	-0.95	0.505288	0.4321381	0.2548265	0.1388775
	-0.75	0.509345	0.4270315	0.2572827	0.1424235
	-0.55	0.549603	0.4271182	0.2627093	0.1489208
	-0.35	0.667549	0.4339027	0.270695	0.1577049
	0.05	1.821497	0.5037863	0.3031573	0.1600749
	0.25	3.566544	0.6224854	0.3580659	0.1835203
	0.45	6.001154	0.8986891	0.4700185	0.2592248
	0.65	8.427731	1.6429605	0.6812483	0.4055339
	0.75	7.393491	2.8589676	0.8883084	0.5485187
	0.95	0.880642	9.2330294	3.3604948	1.7665873

Table 4. Total fertility rate, Female literacy rate, Headcount poverty ratio, and Percentage of urban population in major states of India

STATE	TFR	FLIT	PUP	HCPR
A.P.	1.8	50.4	27.3	15.8
ASSAM	2.4	54.6	12.9	19.7
BIHAR	4	33.1	10.5	41.4
CHHATTISGARH	2.6	51.9	20.1	40.9
GUJARAT	2.4	57.8	37.4	16.8
HARYANA	2.7	55.7	28.9	10
H.P.	1.9	67.4	9.8	14
J&K	2.4	43	24.8	5.4
JHARKHAND	3.3	38.9	22.2	40.3
KARNATAKA	2.1	56.9	34	25
KERALA	1.9	87.7	26	15
M.P.	3.1	50.3	26.5	38.3
MAHARASTRA	2.1	67	42.4	30.7
ODISHA	2.4	50.5	15	46.4
PUNJAB	2	63.4	33.9	8.4
RAJASTHAN	3.2	43.9	23.4	22.1
TAMIL NADU	1.8	64.4	44	22.5
U.P.	3.8	42.2	20.8	32.8
UTTARAKHAND	2.6	59.6	25.7	39.6
W.B.	2.3	59.6	28	24.7

Sources: (i) TFR from EPWRF (2010-11) (ii) URBAN and FLIT from Census of India (2001) and (iii) POV from Planning Commission (2011).

Table 5. Predicted TFR

STATE	OBSERVED	PFLS	PGB
A.P.	1.8	2.49576	2.47243
ASSAM	2.4	2.68139	2.65632
BIHAR	4	3.60914	3.5754
CHHATTISGARH	2.6	2.94126	2.91376
GUJARAT	2.4	2.37845	2.35621
HARYANA	2.7	2.3918	2.36944
H.P.	1.9	2.17028	2.14999
J&K	2.4	2.47091	2.44781
JHARKHAND	3.3	3.39959	3.36781
KARNATAKA	2.1	2.07161	2.05225
KERALA	1.9	1.22177	1.21034
M.P.	3.1	2.81993	2.79356
MAHARASTRA	2.1	2.17079	2.1505
ODISHA	2.4	3.04555	3.01708
PUNJAB	2	2.10204	2.08239
RAJASTHAN	3.2	2.81922	2.79286
TAMIL NADU	1.8	1.81505	1.79808
U.P.	3.8	3.11339	3.08428
UTTARAKHAND	2.6	2.55796	2.53404
W.B.	2.3	2.76707	2.7412

Table 6. FLS and GB Estimators of Coefficients

Variable	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{FLS}}$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{GB}}$
Constants	2.42061	2.39798
FLIT	-0.0268	-0.0265
PUP	-0.0041	-0.0041
HCPR	0.00688	0.00682

Table 7. Global Moran's *I* values

Variable	Observed *I*	E[I]
TFR	0.4339011	-0.05263158
FLIT	0.1637386	-0.05263158
PUP	0.1379433	-0.05263158
HCPR	0.3973285	-0.05263158

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 33–47, DOI 10.21307/stattrans-2019-013

THE EFFECT OF BINARY DATA TRANSFORMATION IN CATEGORICAL DATA CLUSTERING

Jana Cibulková[1], **Zdeněk Šulc**[2], **Sergej Sirota**[3], **Hana Řezanková**[4]

ABSTRACT

This paper focuses on hierarchical clustering of categorical data and compares two approaches which can be used for this task. The first one, an extremely common approach, is to perform a binary transformation of the categorical variables into sets of dummy variables and then use the similarity measures suited for binary data. These similarity measures are well examined, and they occur in both commercial and non-commercial software. However, a binary transformation can possibly cause a loss of information in the data or decrease the speed of the computations. The second approach uses similarity measures developed for the categorical data. But these measures are not so well examined as the binary ones and they are not implemented in commercial software. The comparison of these two approaches is performed on generated data sets with categorical variables and the evaluation is done using both the internal and the external evaluation criteria. The purpose of this paper is to show that the binary transformation is not necessary in the process of clustering categorical data since the second approach leads to at least comparably good clustering results as the first approach.

Key words: hierarchical cluster analysis, nominal variable, binary variable, categorical data, similarity measures, evaluation criteria, generated data.

1. Introduction

The practical importance of cluster analysis increases as the volume of collected data in various fields grows. In the paper, distance-based methods (i.e. methods based on distances or dissimilarities between objects) were chosen for the cluster analysis due to their popularity and ease of implementation in a wide variety of scenarios. Also, according to Charu and Chandan (2013), they can be used with almost any data type, as long as an appropriate measure for given data type exists.

In this paper we focus on hierarchical clustering of objects characterized by categorical variables. This type of data is extremely common in real life. It occurs

[1]Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: jana.cibulkova@vse.cz

[2]Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: zdenek.sulc@vse.cz

[3]Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: sergej.sirota@vse.cz

[4]Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: hana.rezankova@vse.cz

often in surveys regarding marketing research (important for a market-oriented economy) and in surveys in the field of official statistics (e.g. surveys of living conditions). However, most of the clustering algorithms in the literature focus solely on clustering of numerical data. When clustering nominal data (categorical data that are not numerical nor inherently comparable in any way), a binary transformation is routinely used. This transformation recodes nominal variables into sets of dummy variables and then they are "treated" as if they were binary variables all along. In the process of hierarchical clustering, the distances between objects are expressed based on measures suited for binary data. They are either dissimilarity measures or similarity measures which are transformed into dissimilarities before clustering. Despite the fact that this approach is regarded as a standard procedure when clustering nominal variables, it could cause a loss of information in the data (since it is not one-to-one transformation and it changes underlying distribution of transformed variables) or decrease the speed of the computations (due to dimensionality increase), as demonstrated by Salem et al. (2017).

This transformation, which often creates a data set with substantially larger amount of binary variables, may not be necessary at all, since similarity measures suitable for clustering nominal data exist and can be used instead, see Boriah, Chandola and Kumar (2008), Šulc (2016). These measures are not as well examined as the binary ones and they are usually not implemented in any commercial software and almost never used. In non-commercial software R (R Core Team, 2018), there is a package *nomclust*, that contains several similarity measures suited for clustering nominal data, see Šulc and Řezanková (2015). This package was used for the purpose of clustering categorical data by Ladds et al. (2018).

The main objective of the paper is to determine whether applying binary transformation to categorical data and then using similarity measures for binary data in the process of hierarchical clustering of categorical data (approach one) leads to better-quality clusters than using similarity measures for nominal data (approach two), which can be applied on a data set with categorical variables in its original state. The secondary objective is to evaluate the cluster quality of hierarchical clustering with similarity measures for nominal data compared to the similarity measures for binary data on data sets with purely binary variables. We perform the analysis on 600 generated data sets, where 300 of them are data sets with nominal data and 300 of them are data sets with binary data. The approaches are evaluated using both the internal and the external evaluation criteria. A language and environment for statistical computing R is used for the calculations and the analysis.

2. Similarity measures and linkage method

In this section the chosen similarity (or distance) measures are presented. One group of similarity measures was developed for nominal data and let us refer to those ones as *nominal data measures* in this paper. The other group of similarity measures is suitable for binary data and let us use a term *binary data measures* for them. At the very end of this section, the chosen linkage method is presented.

2.1. Nominal data measures

Seven nominal data measures were used in the experiment:

- ES measure (Eskin et al., 2002),
- IOF measure and OF measure (Sparck-Jones, 1972),
- LIN measure (Lin, 1998),
- LIN1 measure (Boriah et al., 2008),
- VE measure and VM measure (Šulc, 2016),
- SM measure (Sokal and Michener, 1958),
- G1 measure, G2 measure, G3 measure and G4 measure (Boriah et al., 2008).

Let us denote the categorical data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1,\ldots,n$ and $c = 1,\ldots,m$; n is the total number of objects; m is the total number of variables. The number of categories of the c-th variable is denoted as K_c, absolute frequency as f, relative frequency as p, q is a subset of relative frequencies satisfying a set of conditions.

The overview of formulas can be found in Table 1, where the column $S_c\left(x_{ic} = x_{jc}\right)$ presents similarity computation for matches of categories in the c-th variable for the i-th and j-th objects, and the column $S_c\left(x_{ic} \neq x_{jc}\right)$ corresponds to mismatches of these categories. The third column represents the total similarity $S(\mathbf{x}_i, \mathbf{x}_j)$ between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$.

Table 1. Nominal measures overview

Measure	$S_c\left(x_{ic} = x_{jc}\right)$	$S_c\left(x_{ic} \neq x_{jc}\right)$	$S\left(\mathbf{x}_i, \mathbf{x}_j\right)$
ES	1	$\frac{K_c^2}{K_c^2+2}$	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
IOF	1	$\left(1+\ln f(x_{ic}) \ln f(x_{ic})\right)^{-1}$	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
OF	1	$\left(1+\frac{n}{\ln f(x_{ic})}\frac{n}{\ln f(x_{ic})}\right)^{-1}$	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
LIN	$2\ln p(x_{ic})$	$2\ln\left(p(x_{ic})+p(x_{jc})\right)$	$\frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{\sum_{c=1}^{m}[\ln(p(x_{ic})+p(x_{jc}))]}$
LIN1	$\sum_{q\in Q}\ln p(q)$; [5]	$2\ln\sum_{q\in Q} p(q)$;	$\frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{\sum_{c=1}^{m}[\ln(p(x_{ic})+p(x_{jc}))]}$
VE	$-\frac{1}{\ln K_c}\sum_{u=1}^{K_c} p_u \ln p_u$	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
VM	$\frac{K_c}{K_c-1}\left[1-\sum_{u=1}^{K_c} p_u^2\right]$	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
SM	1	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
G1	$1-\sum_{q\in Q} p^2(q)$; [6]	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
G2	$1-\sum_{q\in Q} p^2(q)$; [7]	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
G3	$1-p^2(x_{ic})$	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
G4	$p^2(x_{ic})$	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$

[5] $Q \subseteq X_c : \forall q, p(x_{ic}) \leq p(q) \leq p(x_{jc})$
[6] $Q \subseteq X_c : \forall q, p(q) \leq p(x_{ic})$
[7] $Q \subseteq X_c : \forall q, p(q) \geq p(x_{ic})$

In order to compute a proximity matrix, the transformation from similarity into dissimilarity between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$ is necessary. According to Šulc (2016, pp. 6–10) transformations of similarity measures ES, IOF, OF, LIN, LIN1 (measures which can exceed the value one) to corresponding dissimilarity measures follow the formula:

$$D(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{S(\mathbf{x}_i, \mathbf{x}_j)} - 1. \tag{1}$$

The similarity measures VE, VM, SM, G1, G2, G3, G4 (measures which take values from zero to one) are transformed into corresponding dissimilarity measures using the following formula:

$$D(\mathbf{x}_i, \mathbf{x}_j) = 1 - S(\mathbf{x}_i, \mathbf{x}_j). \tag{2}$$

2.2. Binary data measures

According to Todeschini (2012) binary data measures are often linearly dependent, and thus the majority of them produce the same clusters. Therefore, the binary data measures used in the study are selected in a way that each measure is based on different principle and in some way represents a whole group of (linearly dependent) measures based on given principle. These five binary data measures were chosen for the experiment:

- SMC measure (Sokal and Michener, 1958) is the simple matching coefficient and it is a basic measure used for comparing the similarity and diversity of sample sets,
- EUC measure is the Euclidean distance that is the base for many similarity measures,
- PRS measure (Pearson, 1900) – the Pearson chi-squared statistic is one of many measures based on the Pearson correlation coefficient,
- YUQ measure (Yule, 1912) – Yule's Q represents similarity measures based on odds ratio,
- JAC measure (Jaccard, 1901) – Jaccard similarity measure represents negative match exclusive measures.

Suppose that two objects, $\mathbf{x}_i$ and $\mathbf{x}_j$, are represented by the binary vector form. Let m be the number of variables. There are symbols used for the numbers of variables with certain combinations of categories for objects presented in the Table 2, inspired by Dunn and Everitt (1982). The symbols are used for definitions of binary distance measures in this paper. In Table 2, a is the number of features where the values of $\mathbf{x}_i$ and $\mathbf{x}_j$ are both equal to 1, meaning "positive matches", b is the number of variables where the value of $\mathbf{x}_i$ and $\mathbf{x}_j$ is $(0,1)$, meaning "$\mathbf{x}_i$ absence mismatches", c is the number of variables where the value of $\mathbf{x}_i$ and $\mathbf{x}_j$ is $(1,0)$, meaning "$\mathbf{x}_j$ absence mismatches", and d is the number of variables where both $\mathbf{x}_i$ and $\mathbf{x}_j$ are 0, meaning "negative matches".

Table 2. Symbols used for the numbers of variables with certain combinations of categories for objects $\mathbf{x}_i$ and $\mathbf{x}_j$

$\mathbf{x}_i \backslash \mathbf{x}_j$	1 (Presence)	0 (Absence)
1 (Presence)	a	b
0 (Absence)	c	d

Table 3 provides the overview of formulas of the binary data measures. Some measures were defined as similarity measures, hence the transformation from similarity measure into dissimilarity measure is necessary in order to be able to calculate a proximity matrix. This transformation follows Choi et al. (2010).

Column $S(\mathbf{x}_i, \mathbf{x}_j)$ in the Table 3 represents the total similarity between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$ if this measure is originally defined as a similarity between objects. $D(\mathbf{x}_i, \mathbf{x}_j)$ in the last column stands for distance between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$.

Table 3. Binary measures overview

Measure	$S(\mathbf{x}_i, \mathbf{x}_j)$	$D(\mathbf{x}_i, \mathbf{x}_j)$
SMC	$\frac{a+d}{a+b+c+d}$	$1 - S(\mathbf{x}_i, \mathbf{x}_j)$
EUC	–	$\sqrt{b+c}$
PRS	$\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$	$\frac{1-S(\mathbf{x}_i, \mathbf{x}_j)}{2}$
YUQ	–	$\frac{2bc}{ad+bc}$
JAC	$\frac{a}{a+b+c}$	$1 - S(\mathbf{x}_i, \mathbf{x}_j)$

2.3. Method of cluster analysis

We applied agglomerative hierarchical cluster analysis (HCA). Its algorithm considers each object to start in its own cluster and at each step the nearest two clusters are combined into a higher-level cluster. This algorithm is usually attributed to Sokal and Michener (1958).

The average linkage method was applied in this analysis since it is a robust method, which is considered a compromise between the single and the complete linkage methods, see (Yim and Ramdeen, 2015). Unlike the single linkage method, the average linkage method is not associated with *chaining phenomenon* and unlike the complete linkage method it is not sensitive to outliers. Also, this method is frequently set as the default one in hierarchical clustering packages. It takes average pairwise dissimilarity between objects in two different clusters. Let us denote $D_{average}(C_k, C_l)$ the distance between cluster C_k and C_l, with the number of objects n_k in the k-th cluster and n_l in the l-th cluster. Then, dissimilarity between two clusters can be expressed by the formula:

$$D_{average}(C_k, C_l) = \frac{\sum_{\mathbf{x}_i \in C_k} \sum_{\mathbf{x}_j \in C_l} D(\mathbf{x}_i, \mathbf{x}_j)}{n_k n_l}. \tag{3}$$

3. Data sets

To achieve the established aims, data sets with nominal and binary variables are generated. In this section, a data generator is introduced and generated data sets are described.

3.1. Generator of nominal data

Data generation is an important part of various research tasks, whether due to lack of real data or, in our case, due to specific requirements given on desired data sets (given number of clusters, variables, variables' categories, ...) that can influence the robustness of the results. Unfortunately, there are not many nominal data generators which can produce data sets with multivariate structure.

In this paper the data generator suitable for the needs of the experiment is used, see (Cibulková and Řezanková, 2018). Each generated data set consists of a given number of clusters, where each cluster corresponds to one sample of a given multivariate distribution. (For the purpose of generating nominal variables, multivariate uniform distribution is desired and multivariate Bernoulli distribution is required in order to generate binary variables.) This idea follows the assumption of finite mixture models from model-based clustering. It is assumed that the population is made up of several distinct clusters, each following a different multivariate probability density distribution, see (Stahl and Sallis, 2012). Hence, the problem of generating data set with given features is reduced to generating samples from given multivariate distributions. To achieve this, NORTA algorithm (Cario and Nelson, 1997) in combination with Cholesky's decomposition (Higham, 2009) is used. Assuming each cluster in the data set is generated from a given multivariate distribution, the generated data set is a mixture of several samples obtained by this approach. This generator allows us to generate numerous data sets with desired features to cover a wide range of data sets "types", making the results of the analysis more robust.

3.2. Data sets with nominal, binary and binarized variables

For the purpose of the analysis, we introduce terms regarding the data sets.

- *Data set with nominal data* is a data set with nominal variables where a number of categories of each variable belongs to the interval $\langle 2, 10 \rangle$. Each column represents one variable.
- *Data set with binary data* is a data set with binary variables, meaning the value of each variable is either 0 or 1. Each column represents one dummy variable.
- *Data set with binarized data* was created by a binary transformation of generated data set with nominal data. Therefore, one variable with K categories from the "original" data set with nominal data transforms into K dummy variables (columns). Hence, this transformation causes that the data set with binarized data contains a lot of zeros and a huge number of columns.

4. Experiment

The experimental part was designed to evaluate two objectives. The first one, connected to the primary aim of the paper, is to determine if better-quality clusters in hierarchical clustering are provided using similarity measures for binary data, which require a binary data transformation, or using similarity measures for nominal data, which can be applied on a data set with nominal data in its original state. The second objective is to evaluate the cluster quality of the similarity measures for nominal data compared to the similarity measures for binary data on data sets with purely binary data. Its outcomes can help to determine if it is meaningful to use nominal data measures on binary data.

4.1. Experiment setting

Using the data generator, which was presented in Section 3, 300 nominal data sets for the main objective and 300 binary data sets for the secondary analysis were generated. A summary of the generated data sets properties is in Table 4.

Table 4. Generated data sets properties

	data sets with nominal data	data sets with binary data
distribution	multivariate uniform distribution	multivariate Bernoulli distribution
number of objects	120–480	120-480
number of categories	2–10	2
number of clusters	4	4
number of variables	10	10
number of replications	300	300

In order to eliminate the influence of the properties which can possibly have effects on the quality of the produced clusters, certain properties were set under control in the performed analyses, while other properties were not set firmly.

The correlation of variables with parameters of multivariate distribution is chosen randomly. The number of objects in a data set varies from 120 to 480 and the number of categories varies randomly from 2 to 10. Data sets with nominal data were generated from multivariate uniform distribution, while multivariate Bernoulli distribution was used for generating data sets with binary data. In both the analyses, the number of clusters was set to four and the number of variables is set to ten to cover typical data set sizes in common clustering tasks. To ensure the robustness of the obtained results, each data set setting combination was replicated 300 times.

4.2. Evaluation criteria

Since the analyses are performed on the generated data sets, and thus objects' cluster memberships are known, the produced clusters can be evaluated using both internal and external evaluation criteria.

For the internal cluster quality evaluation, the variability-based *Pseudo F coefficient based on the mutability* (PSFM) was chosen, see Řezanková et al. (2011).

This coefficient takes into account the within-cluster variability of a data set, which always decreases with the increasing number of clusters. Therefore, the coefficient penalizes an increasing number of clusters. Then, the maximal value indicates the optimal number of clusters. The PSFM criterion can be expressed by the formula

$$PSFM(k) = \frac{(n-k)(WCM(1)-WCM(k))}{(k-1)WCM(k)}, \tag{4}$$

where $WCM(1)$ is the variability in the whole data set with n objects, and $WCM(k)$ the within-cluster variability in the k-cluster solution, which is computed as

$$WCM(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \left(1 - \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g}\right)^2\right),$$

where n_g is the number of objects in the g-th cluster ($g = 1,...,k$), n_{gcu} is the number of objects in the g-th cluster by the c-th variable with the u-th category ($u = 1,...,K_c$).

The external evaluation of the cluster quality was performed using the *Adjusted Rand Index* (ARI), see Hubert and Arabie (1985), which is commonly used for a comparison of two membership partitions. Compared to the standard Rand index, see Rand (1971), it is corrected for a chance. Similarly to the original measure, which takes values from zero to one, where one indicates that the compared cluster partitions are identical, ARI has a similar range of values, but it can also take small negative values if the *Index* is less than the *Expected index*, see

$$ARI = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum \binom{b_j}{2}]/\binom{n}{2}}, \tag{5}$$

where n_{ij} are the joint frequencies of the contingency table created between two compared partitions, a_i are the row marginal frequencies, and b_j are the column marginal frequencies.

4.3. Evaluation methodology

Values of the two criteria used can be compared not only with their values in different cluster solutions of a certain similarity measure but also with their values in a particular cluster solution for different similarity measures. It can be done by averaging the scores of the evaluation criteria over the examined similarity measures (and/or certain data sets' properties). However, the presented approach can be used only by ARI. The values of PSFM must be processed in a different way since this criterion depends on the number of objects in a data set and also on its initial variability, and thus, it is incomparable in an unadjusted form. Therefore, the proposed procedure uses the two-step rank score approach.

In the first step, clusters produced by HCA with all the examined similarity measures are evaluated. The outcome scores are then ranked in a way that the lowest

rank is assigned to the highest value of the coefficient. Then, the rank scores are averaged in the same way as ARI. The resulting mean rank scores and their standard deviations are considered as the main output which can be displayed in the form of an easily interpretable table. The lower is the mean ranked score of a similarity measure, the better is its clustering performance. The lower the value of the standard deviation, the more stable the clustering performance of a given similarity measure.

4.4. Results of the experiment

Interpretation of the results follows the methodology described in Section 4.3. The measures that have a tendency to create "high-quality" clusters are the ones with low ranks of PSFM index, high values of ARI and low values of PSFM rank's standard deviation.

Table 5 and Table 6 show mean ranks of PSFM, standard deviations of PSFM ranks and average ARIs for each measure. The best values are highlighted in bold writing (the highest values of ARI and the lowest values of PSFM). Table 5 provides a summary of evaluation criteria for data sets with nominal data (these data sets were binarized if the binary data measure was used in the clustering process). Table 6 summarizes the same indices for data sets with binary data (these data sets were generated as data sets with purely binary variables). It is possible to distinguish a type of a measure by the column *Type*, where "B" stands for a binary data measure (black colour) and "N" stands for a nominal data measure (red colour).

Figure 1 and Figure 2 give a visualization of Tables 5 and 6. Axes x and y in the graphs reflect the averages from the tables (average PSFM rank and average ARI) and the size of a grey circle changes according to the standard deviation of PSFM rank. The colours of measures in the figures correspond to the colours in the tables. In these figures, the measures at the bottom right lead to the best clustering solutions, while the measures at the top left lead to the worst clustering solutions.

We can see that in the case of clustering of nominal data and also in the case of clustering binary data, the clustering approach without the binary transformation (using nominal data measures) provides at least as good clustering solutions as the standard approach with binary transformation. According to the chosen evaluation criteria, similarity measures for nominal data and similarity measures for binary data perform comparably well when applied on data sets with nominal (binarized) data. Especially measures EUC, SMC, LIN, VE, VM, SM provided good clustering solutions according to the chosen evaluation criteria. Surprisingly, some similarity measures for nominal data (LIN, LIN1, G3) performed even better than all examined measures for binary data on data sets with binary data. The measure for nominal data LIN steadily leads to the above average clustering solutions when applied to data sets with binary and nominal data. The measures PRS and G4 lead to bellow average clustering solutions. The measures for binary data EUC and SMC handled well high dimensional (binarized) data sets with a lot of zeros. However, they were outperformed by several similarity measures when applied to binary data sets.

Table 5. Experiment results (data sets with nominal/binarized data)

Measure	Type	PSFM		ARI
		Mean	SD	Mean
SMC	B	**7.9**	3.94	**0.566**
EUC	B	**7.9**	3.82	**0.565**
PRS	B	12.9	4.78	0.484
YUQ	B	8.1	4.27	0.548
JAC	B	8.0	3.90	0.567
ES	N	9.7	6.04	0.395
IOF	N	8.2	4.60	0.533
OF	N	8.9	4.59	0.600
LIN	N	**7.9**	4.18	**0.564**
LIN1	N	14.1	4.03	0.512
VE	N	**7.8**	3.94	**0.565**
VM	N	**7.8**	3.91	**0.566**
SM	N	**7.8**	3.77	**0.566**
G1	N	8.8	4.52	0.592
G2	N	8.9	4.49	0.585
G3	N	8.5	4.14	0.580
G4	N	10.0	6.04	0.389

Table 6. Experiment results (data sets with binary data)

Measure	Type	PSFM		ARI
		Mean	SD	Mean
SMC	B	8.1	3.82	0.308
EUC	B	8.5	4.20	0.310
PRS	B	16.9	0.53	0.069
YUQ	B	8.2	4.33	0.303
JAC	B	8.5	4.03	0.308
ES	N	8.1	3.98	0.306
IOF	N	7.9	4.13	0.304
OF	N	8.2	3.88	0.307
LIN	N	**6.5**	4.27	**0.329**
LIN1	N	**6.6**	4.35	**0.329**
VE	N	8.3	3.81	0.307
VM	N	8.2	3.88	0.308
SM	N	8.1	3.84	0.307
G1	N	9.1	5.26	**0.336**
G2	N	11.3	4.70	0.275
G3	N	**6.9**	4.81	**0.336**
G4	N	13.7	4.33	0.200

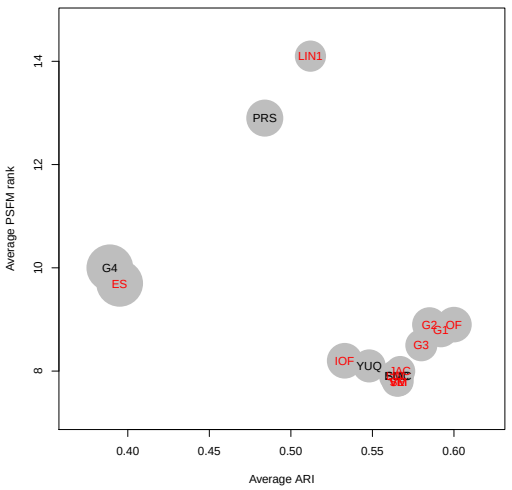

Figure 1: Data sets with nominal (or binarized) data

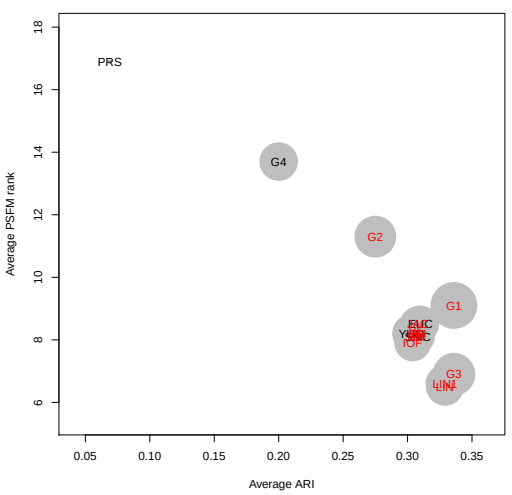

Figure 2: Data sets with binary data

5. Conclusions

In the study we compared two approaches to clustering of categorical data. The first widely used approach performs a binary transformation of the nominal variables into sets of dummy variables and then uses the similarity measures suitable for binary data. The second rarely used approach uses similarity measures developed for the nominal data, hence no data transformation is required. We used internal and external evaluation criteria to determine which of the two approaches creates better quality clusters.

We demonstrated that the binary transformation is not necessary and it is possible to cluster data sets with categorical variables without it. Moreover, according to several internal and external evaluation criteria the approach that uses nominal data measures even leads to "better" clustering results in comparison with clustering solutions obtained by the first approach (clustering data that were transformed by a binary transformation, while using distance measures suitable for binary data) on both types of data sets – data sets with nominal data and data sets with binary data.

Acknowledgements

This work was supported by the University of Economics, Prague under Grant IGA F4/44/2018.

REFERENCES

BORIAH, S., CHANDOLA, V., KUMAR, V., (2008). Similarity measures for categorical data: A comparative evaluation, In Proceedings of the 2008 SIAM International Conference on Data Mining, Society for Industrial, Applied Mathematics, pp. 243–254.

CAIRO, M., NELSON, B., (1997). Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix, Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

CHARU, C. A., CHANDAN, K. R., (2013). Data Clustering: Algorithms and Applications, Chapman & Hall/CRC.

CHOI, S. S., CHA, S. H., TAPPERT, C. C., (2010). A survey of binary similarity and distance measures,Journal of Systemics, Cybernetics and Informatics, 8 (1), pp. 43–48.

CIBULKOVÁ, J., ŘEZANKOVÁ, H., (2018). Categorical data generator, In International Days of Statistics and Economics 2018. T. Löster and T. Pavelka (eds.) Slaný: Melandrium, Libuše Macáková, pp. 288–296.

DUNN, G., EVERITT, B. S., (1982). An Introduction to Mathematical Taxonomy, Cambridge University Press.

ESKIN, E., ARNOLD, A., PRERAU, M., PORTNOY, L., STOLFO, S. V., (2002). A geometric framework for unsupervised anomaly detection, In Applications of Data Mining in Computer Security, D. Barbará and S. Jajodia (eds.) Boston: Springer, pp. 78–100.

HAHSLER, M., BUCHTA, C., GRUEN, B., HORNIK, K., (2015). Arules: Mining Association Rules and Frequent Itemsets. R package version 1.3-1. https://CRAN.R-project.org/package=arules.

HIGHAM, N. J., (2009). Cholesky factorization, Wiley Interdisciplinary Reviews: Computational Statistics, 1 (2), pp. 251–254.

HUBERT, L., ARABIE, P., (1985). Comparing partitions, Journal of Classification, 2 (1), pp. 193–218.

JACCARD, P., (1901). Étude comparative de la distribuition florale dans une portion des Alpes et des Jura, Bulletin de la Societe Vaudoise des Sciences Naturelles, 37 (142), pp. 547–579.

LADDS, M. A., SIBANDA, N., ARNOLD, R., DUNN, M. R., (2018). Creating functional groups of marine fish from categorical traits, PeerJ 6:e5795.

LIN, D., (1998). An information-theoretic definition of similarity. In ICML '98: Proceedings of the 15th International Conference on Machine Learning, San Francisco: Morgan Kaufmann Publishers Inc., pp. 296–304.

PEARSON, K., (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philosophical Magazine, Series 5, 50(302), pp. 157–175.

QIU, W., JOE, H., (2015). clusterGeneration: Random Cluster Generation (with Specified Degree of Separation). R package version 1.3.4. https://CRAN.R-project.org/package=clusterGeneration.

R CORE TEAM (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RAND, W. M., (1971). Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, 66 (336), pp. 846–850.

ŘEZANKOVÁ, H., LÖSTER, T., HÚSEK, D., (2011). Evaluation of Categorical Data Clustering, In Advances in Intelligent Web Mastering – 3, Advances in Intelligent and Soft Computing. E. Mugellini, P. S. Szczepaniak, M. C. Pettenati and M. Sokhn (eds.), vol 86. Berlin:Springer, Heidelberg, pp. 173–182.

SALEM, S. B., NAOUALI, S., SALLAMI, M., (2017). Clustering Categorical Data Using the K-Means Algorithm and the Attribute's Relative Frequency, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 11 (6), pp. 708–713.

SOKAL, R., MICHENER, C., (1958). A statistical method for evaluating systematic relationships, University of Kansas Science Bulletin, 38 (2), pp. 1409–1438.

SPARCK-JONES, K., (1972). A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28 (1), pp. 11–21.

STAHL, D., SALLIS, H., (2012). Model-based cluster analysis, In Wiley Interdisciplinary Reviews: Computational Statistics, 4 (4), pp. 341–358.

ŠULC, Z., (2016). Similarity measures for nominal data in hierarchical clustering. Dissertation thesis, Prague: University of Economics.

ŠULC, Z., ŘEZANKOVÁ, H., (2015). Nomclust: An R package for hierarchical clustering of objects characterized by nominal variables, In International Days of Statistics and Economics 2018. T. Löster and T. Pavelka (eds.) Slaný: Melandrium, pp. 1581–1590.

TODESCHINI, R., CONSONNI, V., XIANG, H., HOLLIDAY, J., BUSCEMA, M., WILLETT, P., (2012). Similarity coefficients for binary chemoinformatics Data: Overview and extended comparison using simulated and real data sets, Journal of Chemical Information and Modeling, 52 (11), pp. 2884–2901.

YULE, G U., (1912). On the methods of measuring association between two attributes, Journal of the Royal Statistical Society, 49 (6), pp. 579–652.

YIM, O., RAMDEEN, K. T., (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data, The Quantitative Methods for Psychology, 11 (1), pp. 8–21.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 49–67, DOI 10.21307/stattrans-2019-014

A COMPARATIVE ANALYSIS OF ECONOMIC EFFICIENCY OF MEDIUM-SIZED MANUFACTURING ENTERPRISES IN DISTRICTS OF WIELKOPOLSKA PROVINCE USING THE HYBRID APPROACH WITH METRIC AND INTERVAL-VALUED DATA

Grażyna Dehnel[1], **Marek Walesiak**[2]

ABSTRACT

The article describes a hybrid approach to evaluating economic efficiency of medium-sized manufacturing enterprises (employing from 50 to 249 people) in districts of Wielkopolska province, using metric and interval-valued data. The hybrid approach combines multidimensional scaling with linear ordering. In the first step, multidimensional scaling is applied to obtain a visual representation of objects in a two-dimensional space. In the next step, a set of objects is ordered linearly based on the distance from the pattern (ideal) object. This approach provides new possibilities for interpreting linearly ordered results of a set of objects. Interval-valued variables characterise the objects of interests more accurately than metric data do. Metric data are atomic, i.e. an observation of each variable is expressed as a single real number. In contrast, an observation of each interval-valued variable is expressed as an interval. The analysis was based on data prepared in a two-stage process. First, a data set of observations was obtained for metric variables describing economic efficiency of medium-sized manufacturing enterprises. These unit-level data were aggregated at district level (LAU 1) and turned into two types of data: metric and interval-valued data. In the analysis of interval-valued data, two approaches are used: symbolic-to-classic, symbolic-to-symbolic. The article describes a comparative analysis of results of the assessment of economic efficiency based on metric and interval-valued data (the results of two approaches). The calculations were made with scripts prepared in the R environment.

Key words: medium-sized enterprise, metric data, interval-valued data, multidimensional scaling, composite measures

JEL: C38, C43, C63, C88, R12

[1] Poznan University of Economics and Business, Department of Statistics, Poznań. E-mail: grazyna.dehnel@ue.poznan.pl. ORCID ID: https://orcid.org/0000-0002-0072-9681.
[2] Wroclaw University of Economics, Department of Econometrics and Computer Science, Jelenia Góra. E-mail: marek.walesiak@ue.wroc.pl. ORCID ID: https://orcid.org/0000-0003-0922-2323.

1. Introduction and motivation

The contribution made to the GDP by small and medium-sized enterprises keeps growing, in contrast to that of large companies. Although the SME sector is dominated by micro enterprises, one cannot ignore the role played by medium-sized companies, employing between 50 and 249 persons (CSO 2017). At present there are nearly 16,000 medium-sized companies in Poland, which accounts for just 0.8% of the entire enterprise sector. This share has remained unchanged for the last 10 years (MED 2017). Medium-sized companies provide more jobs than the small ones (17%). An average medium-sized enterprise employs 104 persons, while the total number of people employed in companies of this category is 1.6 million. Investment outlays in this category account for 33% of the entire enterprise sector, 64% of which are own funds (see Figure 1). Medium-sized enterprises are the most dynamically developing category of companies in terms of the value of exports per one company. They are also characterized by the highest survival rate – 87% of them survive their first year of operation. Medium-sized companies operating for 5 years are likely to survive the next year with a probability of 0.996 (Chaber *et al*. 2017).

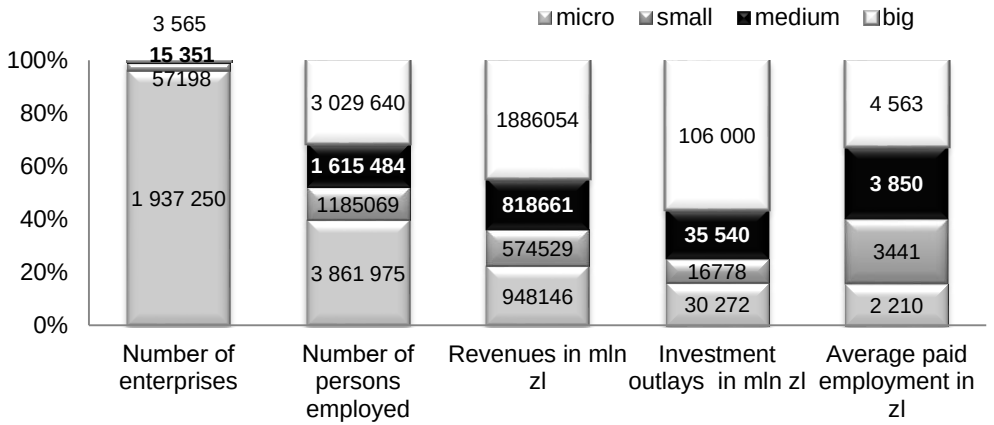

Figure 1. Enterprise characteristics by size class in 2016 (at 31 Dec.)

Source: Based on the CSO study (CSO 2017).

Medium-sized companies are able to compete with large enterprises because they are more flexible and efficient in conducting business activity, are better at controlling costs and take less time to implement innovation and react to changing market requirements.

Taking into account the kind of business activity, one of the most important sections is manufacturing. Looking at the structure of manufacturing companies (see Figure 2), it can be seen that medium-sized enterprises are the smallest group and make up only 3% of all units in this section. People employed by

medium-sized manufacturing enterprises account for about 27% of the workforce working in all manufacturing companies. Revenues earned by medium-sized manufacturing enterprises make up 21% of all revenues generated by companies in the manufacturing section.

The empirical study described below is limited to the group of medium-sized manufacturing enterprises, which includes 42% of all medium-sized companies. Those companies employ 44% of the workforce working in this sector. The share of revenues and wages in this group is similar (CSO 2017).

The main objective of the study was to evaluate the economic efficiency of medium-sized manufacturing enterprises in districts of Wielkopolska province. The study was based on metric and interval-valued data and involved a hybrid approach combining multidimensional scaling and linear ordering (Walesiak 2016; Walesiak, Dehnel 2018). Economic efficiency, defined as a relation between effects and investments, in this case, is measured on an operational level using efficiency ratios to assess the company's performance (Kaplan, Cooper 1998; Kaplan 2008; Koliński 2011). Studies of this kind are usually based on a matrix of metric data. The novelty of the present study is the fact that it was based on a table of interval-valued data. In addition, the authors propose an aggregate measure based on the Euclidean Ichino-Yaguchi distance from the pattern object. Interval-valued variables describe objects of interest more accurately than metric data do, which are atomic, meaning that an observation of each variable is expressed as a single real number. In contrast, an observation of each interval-valued variable is expressed as an interval. The following studies (Gioia, Lauro 2006; Brito *et al*. 2015) include real examples of interval-valued data.

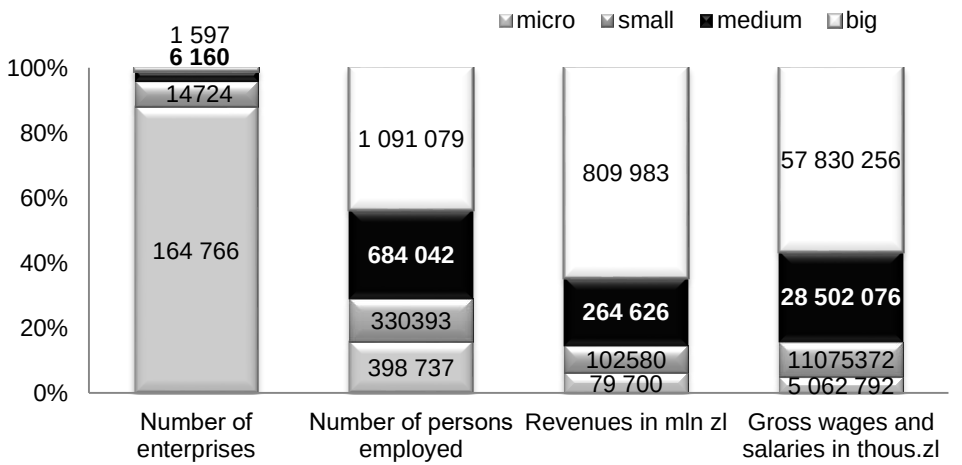

Figure 2. Characteristics of manufacturing enterprises by size class in 2016 (at 31 Dec.)

Source: Based on the CSO study (CSO 2017).

Data for the study were prepared in two steps. The first step involved compiling a set containing metric variables about the economic efficiency of medium-sized manufacturing enterprises; in the second step, the collected data were aggregated at the level of districts, producing metric and interval-valued data. The latter type of data was analysed using two approaches: symbolic-to-classic and symbolic-to-symbolic. Data used in the study come from the DG-1 survey conducted by the Statistical Office in Poznań. The survey is carried out to collect information about basic measures of economic activity in companies (Dehnel 2015). Owing to data availability, the study was conducted for 2012. The official statistics were supplemented by information from the register maintained by the Ministry of Finance.

2. Research methodology

To produce a ranking of medium-sized manufacturing companies operating in districts of Wielkopolska province in terms of economic efficiency, the authors used a hybrid approach, which combines multidimensional scaling (MDS) and linear ordering (Walesiak 2016; Walesiak, Dehnel 2018), which makes it possible to visualize the results of linear ordering. Metric and interval-valued data were used for this purpose. Depending on the type of input and output of multidimensional scaling, three different approaches were used to analyse the data:

a. Classic-to-classic (cc) for metric data,
b. Symbolic-to-classic (sc) for interval-valued data,
c. Symbolic-to-symbolic (ss) for interval-valued data.

The extended analytical procedure (including the above mentioned approaches), accounting separately for metric and interval-valued data, consists of the following steps:

1. Select a complex phenomenon which cannot be measured directly (in this case, it is the economic efficiency of medium-sized manufacturing companies operating in districts of Wielkopolska province).
2. Identify a set of objects of interest and a set of variables that are substantively related to the complex phenomenon. Add a pattern object (upper pole) and an anti-pattern object (lower pole) to the set of objects. Identify preference variables[3] (stimulants, destimulants and nominants).
3. Collect data and construct a data matrix $\mathbf{X} = [x_{ij}]_{nxm}$, (the value of the *j*-th variable for the *i*-th object, $i, k = 1, \ldots, n$, $j = 1, \ldots, m$) for metric data or a data table $\mathbf{X} = [x_{ij}^{l}, x_{ij}^{u}]_{nxm}$ (where $x_{ij}^{l} \leq x_{ij}^{u}$) for interval-valued data. The pattern object includes the most favourable variable values, whereas the anti-pattern – the least favourable values of the preference variables (separately for lower and upper bounds of the interval).
4. Normalize variable values and arrange them in the form of a normalized data matrix $\mathbf{Z} = [z_{ij}]_{nxm}$ for metric data or in the form of a normalized data table $\mathbf{Z} =$

3 The idea of a stimulant and a destimulant was introduced by (Hellwig 1972), while that of a nominant in the work by (Borys 1984, p. 118). Definitions can be found, among others, in (Walesiak 2016).

$[z_{ij}^l, z_{ij}^u]_{nxm}$ (where $z_{ij}^l \leq z_{ij}^u$) for interval-valued data. Normalization is used to ensure comparability of variables. This is achieved by removing dimensional units from measurement results and standardizing their orders of magnitude. Interval-valued data require special normalization treatment. The lower and upper bound of the interval of the *j*-th variable for *n* objects are combined into one vector containing 2*n* observations. This approach makes it possible to apply normalization methods used for classic metric data. Metric data were normalized using the data.Normalization function, while interval-valued data – using interval_normalization function, both available in the clusterSim package (Walesiak, Dudek 2018a).

5. In the classic-to-classic approach, select a measure of distance for metric data (Manhattan, Euclidean, Chebyshev, Squared Euclidean, GDM1[4] – see, e.g. Everitt *et al*. 2011, pp. 49-50), calculate distances and create a distance matrix $\mathbf{\delta} = [\delta_{ik}(\mathbf{Z})]_{nxn}$ $(i, k = 1, \ldots, n)$.

   For interval-valued data (the symbolic-to-classic approach), select a measure of distance (see Table 1), calculate distances and create a distance matrix $\mathbf{\delta} = [\delta_{ik}(\mathbf{Z})]_{nxn}$.

Table 1. Selected distance measures for interval-valued data

Symbol	Name	Distance measure $\delta_{ik}(\mathbf{Z})$
U_2_q1	Ichino-Yaguchi $q = 1, \gamma = 0{,}5$	$\sum_{j=1}^{m} \varphi(z_{ij}, z_{kj})$
U_2_q2	Euclidean Ichino-Yaguchi $q = 2, \gamma = 0{,}5$	$\sqrt{\sum_{j=1}^{m} \varphi(z_{ij}, z_{kj})^2}$
H_q1	Hausdorff $q = 1$	$\sum_{j=1}^{m} [\max(\lvert z_{ij}^l - z_{kj}^l \rvert, \lvert z_{ij}^u - z_{kj}^u \rvert)]$
H_q2	Euclidean Hausdroff $q = 2$	$\left\{\sum_{j=1}^{m} [\max(\lvert z_{ij}^l - z_{kj}^l \rvert, \lvert z_{ij}^u - z_{kj}^u \rvert)]^2\right\}^{1/2}$

$z_{ij} = [z_{ij}^l, z_{ij}^u]$; $\varphi(z_{ij}, z_{kj}) = \lvert z_{ij} \oplus z_{kj} \rvert - \lvert z_{ij} \otimes z_{kj} \rvert + \gamma(2 \cdot \lvert z_{ij} \otimes z_{kj} \rvert - \lvert z_{ij} \rvert - \lvert z_{kj} \rvert)$; $\lvert \ \rvert$ – interval length; $z_{ij} \oplus z_{kj} = z_{ij} \cup z_{kj}$; $z_{ij} \otimes z_{kj} = z_{ij} \cap z_{kj}$.

Source: Based on works by Billard, Diday 2006; Ichino, Yaguchi 1994.

This step does not apply in the symbolic-to-symbolic approach.

6. In the classic-to-classic and symbolic-to-classic approaches conduct multidimensional scaling (MDS): $f: \delta_{ik}(\mathbf{Z}) \rightarrow d_{ik}(\mathbf{V})$ for all pairs (i, k), where *f* denotes distance mapping from *m*-dimensional space $\delta_{ik}(\mathbf{Z})$ into

4 Cf. Jajuga, Walesiak, Bąk 2003.

corresponding distances $d_{ik}(\mathbf{V})$ in q- dimensional space ($q < m$). To enable graphic presentation of results, q is set to 2. Distances $d_{ik}(\mathbf{V})$ are unknown. The iterative procedure, implemented in the **smacof** algorithm and used to find configuration $\mathbf{V}$ (given q dimensions) and calculate distance matrix $d_{ik}(\mathbf{V})$, is presented in (Borg, Groenen 2005, pp. 204–205).

In the classic-to-classic and symbolic-to-classic approaches, after performing MDS, one obtains a data matrix in 2-dimensional space: $\mathbf{V} = [v_{ij}]_{nxq}$ ($q = 2$). Depending on the location of the pattern and anti-pattern object in the dimensional scaling space $\mathbf{V} = [v_{ij}]_{nx2}$ the coordinate system needs to be rotated by an angle of φ according to the formula:

$$[v'_{ij}]_{nx2} = [v_{ij}]_{nx2} \times D, \tag{1}$$

where: $[v'_{ij}]_{nx2}$ – data matrix in 2-dimensional scaling space after rotating the coordinate system by an angle of φ,

$D = \begin{bmatrix} cos\varphi & -sin\varphi \\ sin\varphi & cos\varphi \end{bmatrix}$ – rotation matrix.

The rotation does not change the arrangement of objects relative to one another but makes it possible to position the set axis connecting the pattern and anti-pattern along the identity line, which improves the visualization of results.

In the symbolic-to-symbolic approach, multidimensional scaling needs to be performed using the I-Scal algorithm. The objective of MDS for interval dissimilarities is to represent the lower and upper bounds of the dissimilarities by minimum and maximum distances between rectangles as well as possible distances in the sense of least-squares (Groenen, Winsberg, Rodriguez, Diday 2006).

Under this approach, after performing MDS, one obtains an interval-valued data table in 2-dimensional space $\mathbf{V} = [v_{ij}^l, v_{ij}^u]_{nxq}$ (where $v_{ij}^l \leq v_{ij}^u$; $q = 2$).

A frequent mistake committed while using MDS results is to evaluate stress mechanically (rejecting an MDS solution because its stress seems “too high”). According to Borg, Groenen, Mair (2013, p. 68; 2018, pp. 85-86) “an MDS solution can be robust and replicable, even if its stress value is high” and “Stress, moreover, is a *summative* index for *all* proximities. It does not inform the user how well a *particular* proximity value is represented in the given MDS space”. In addition we should take into account stress per point measure[5] and Shepard diagram[6] (classic-to-classic and symbolic-to-classic approaches) or the I-Stress per box index and the I-dist diagram (the symbolic-to-symbolic approach).

In this study, we used a solution which enables the selection of an optimal MDS procedure for a given normalization method, distance measure and scaling models (in the classic-to-classic and symbolic-to-classic approaches) and, in the case of the symbolic-to-symbolic approach, according to procedures available in the **`mdsOpt`** R package (Walesiak, Dudek 2018b).

[5] Cf. Borg and Mair (2017, pp. 31).
[6] Cf. Mair, Borg and Rusch (2016).

7. In all three approaches, MDS results should be presented graphically in a 2-dimensional space and interpreted.

   In the classic-to-classic and symbolic-to-classic approaches, objects are represented as points. Two points, representing the anti-pattern and pattern, are joined by a straight line to form the so-called set axis in the diagram. Isoquants of development (curves of equal development) are drawn from the pattern point. Objects located between the isoquants represent a similar level of development. The same level can be achieved by objects located at different points along the same isoquant of development (due to a different configuration of variable values).
   In the symbolic-to-symbolic approach, objects are represented in the form of rectangles.
8. In the classic-to-classic and symbolic-to-classic approaches, objects should be ordered linearly according to the values of the aggregate measure d_i based on the Euclidean distance from the pattern object (Hellwig 1981):

$$d_i = 1 - \sqrt{\sum_{j=1}^{2}(v_{ij} - v_{+j})^2} \Big/ \sqrt{\sum_{j=1}^{2}(v_{+j} - v_{-j})^2}, \tag{2}$$

   where: v_{ij} – the *j*-th coordinate for the *i*-th object in the 2-dimensional MDS space, $v_{+j}(v_{-j})$ – the *j*-th coordinate for the pattern (anti-pattern) object in the 2-dimensional MDS space.
   In the symbolic-to-symbolic approach, objects should be ordered according to the values of the aggregate measure d_i based on the Euclidean Ichino-Yaguchi distance (Ichino, Yaguchi 1994) from the pattern object:

$$d_i = 1 - \sqrt{\sum_{j=1}^{2}\varphi\left(v_{ij}, v_{+j}\right)^2} \Big/ \sqrt{\sum_{j=1}^{2}\varphi\left(v_{+j}, v_{-j}\right)^2}, \tag{3}$$

   where: $v_{ij} = \left[v_{ij}^{l}, v_{ij}^{u}\right]$; $v_{+j} = \left[v_{+j}^{l}, v_{+j}^{u}\right]$; $v_{-j} = \left[v_{-j}^{l}, v_{-j}^{u}\right]$;

   v_{ij}^{l} and v_{ij}^{u} – the lower and upper bound of the interval of the *j*-th variable for the *i*-th object in the 2-dimensional MDS space;

   v_{+j}^{l} and v_{+j}^{u} (v_{-j}^{l} i v_{-j}^{u}) – the lower and upper bound of the interval of the *j*-th variable for the pattern (anti-pattern) object in the 2-dimensional MDS space.

   The values of the aggregate measure d_i given by (2) and (3) belong to the interval $[0; 1]$. The higher the value of d_i, the higher the economic efficiency of medium-sized manufacturing enterprises in the objects (districts). The objects are arranged according to the descending values of the aggregate measure d_i.

3. Results of the Empirical Study

The empirical study uses statistical data about the economic efficiency of medium-sized manufacturing enterprises in districts of Wielkopolska province in 2012. The target data set was prepared in two stages. The first step involved selecting three metric variables (x1 and x2 are stimulants and x3 is a destimulant)

describing the economic efficiency of 876 medium-sized manufacturing enterprises:

x1 – return on sales in % (net profit as a percentage of sales revenue).

x2 – sales revenue in thousands PLN per one employee,

x3 – costs in thousands PLN per one employee.

In the second step, the observations were aggregated at the level of districts producing a set of interval-valued data. The economic efficiency of medium-sized manufacturing enterprises operating in 35 districts of Wielkopolska province was measured using three approaches: classic-to-classic, symbolic-to-classic and symbolic-to-symbolic.

In the classic-to-classic approach, the analytical procedure described in the second section was applied to a data matrix containing 35 districts of Wielkopolska province described by the three metric variables. For this purpose, original data for 876 manufacturing enterprises were aggregated at the level of districts by averaging the values of each variable.

In the symbolic-to-classic and symbolic-to-symbolic approaches, the analytical procedure described in the second section was applied to a table containing 35 districts of Wielkopolska province described by the three interval-valued variables. Original data for 876 manufacturing enterprises were aggregated at the level of districts, producing interval-valued data. The lower bound of the interval for each interval-valued variable in each district was given by the first quartile of the entire data set. The upper bound of the interval was obtained by calculating the third quartile.

In the classic-to-classic approach, an optimal scaling procedure was selected after testing combinations of 6 normalization methods (n1, n2, n3, n5, n5a, n12a – see Walesiak, Dudek 2018a), 4 distance measures (Manhattan, Euclidean, Chebyshev, Squared Euclidean, GDM1) and 4 MDS models (ratio, interval, mspline of second and third degree – Borg, Groenen 2005, p. 202) – altogether 120 MDS procedures. As a result of applying the `optSmacofSym_mMDS` function from the **`mdsOpt`** R package (see Walesiak, Dudek 2017; 2018b), the optimal MDS procedure was selected. The procedure uses the normalization method n2 (positional standardization), the mspline 2 scaling model (polynomial of second degree) and the GDM1 distance.

In the symbolic-to-classic approach, an optimal scaling procedure was selected after testing combinations of 6 normalization methods (n1, n2, n3, n5, n5a, n12a), 4 distance measures (Ichino-Yaguchi, Euclidean Ichino-Yaguchi, Hausdorff, Euclidean Hausdorff) and 4 MDS models (ratio, interval, mspline of second and third degree) – altogether 96 MDS procedures. After applying the `optSmacofSymInterval` function from the **`mdsOpt`** R package, the optimal MDS procedure was selected, which involves the normalization method n12a (positional normalization), the mspline 2 scaling model (polynomial of third degree) and the Hausdorff distance.

In the symbolic-to-symbolic approach, an optimal scaling procedure was selected after testing combinations of 6 normalization methods (n1, n2, n3, n5, n5a, n12a) and 2 optimization methods, giving altogether 12 MDS procedures. After applying the `optIscalInterval` function from the **`mdsOpt`** R package, the optimal MDS procedure was selected, which uses the normalization method n1

(standardization) and the MM optimization method (majorization-minimization algorithm).

By taking into account all the three approaches, it was possible to see how assessments of the phenomenon of interest varied when moving from the classic-to-classic approach to more robust ones (symbolic-to-classic, symbolic-to-symbolic). The average value, used in the classic-to-classic approach as the only parameter, which is well known, is strongly affected by outliers. In the other two approaches based on interval-valued data, assessments obtained for districts are not based on average values but account for the variation observed among manufacturing enterprises with respect to the variables of interest. Additional advantage of these approaches is the fact that outliers are excluded from the analysis.

Figures 3, 4 and 5 present MDS results of districts of Wielkopolska province for each approach.

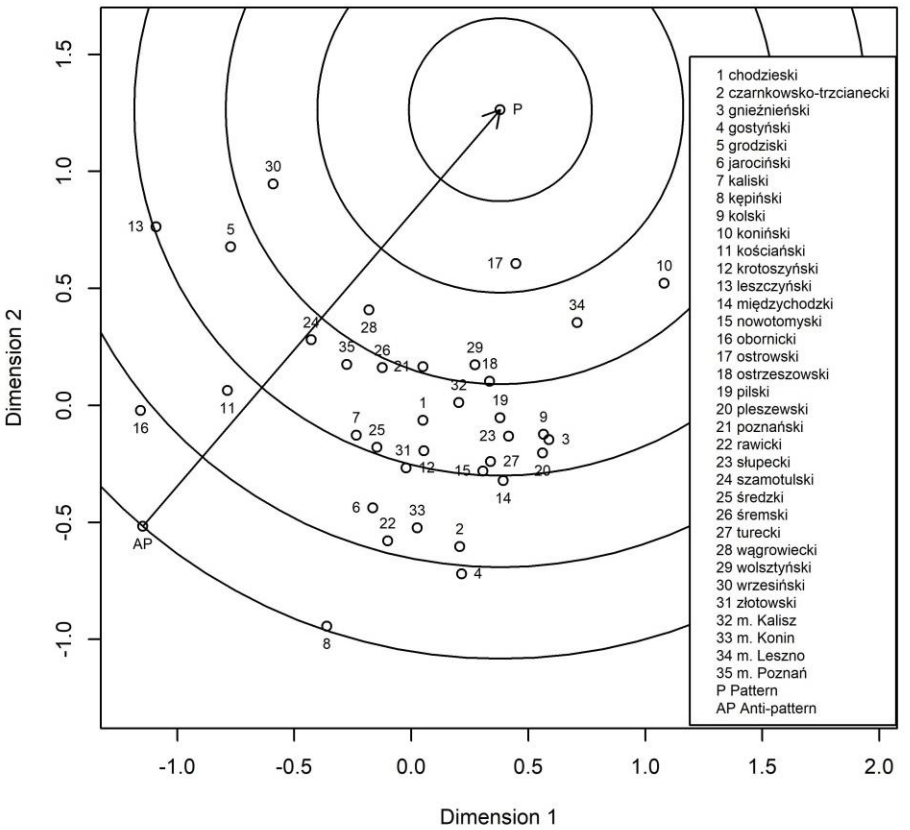

Figure 3. Results of multidimensional scaling of 35 districts of Wielkopolska by economic efficiency of medium-sized manufacturing enterprises in 2012 – the classic-to-classic approach

In the diagram illustrating the classic-to-classic and the classic-to-symbolic approaches, the anti-pattern (AP) and pattern (P) objects were connected by a straight line – the so-called set axis (Figs. 3 and 4). 6 isoquants of development were identified by dividing the set axis into 6 equal parts. The further a given isoquant is located from the pattern object, the less economically efficient are medium-sized companies in districts represented within it.

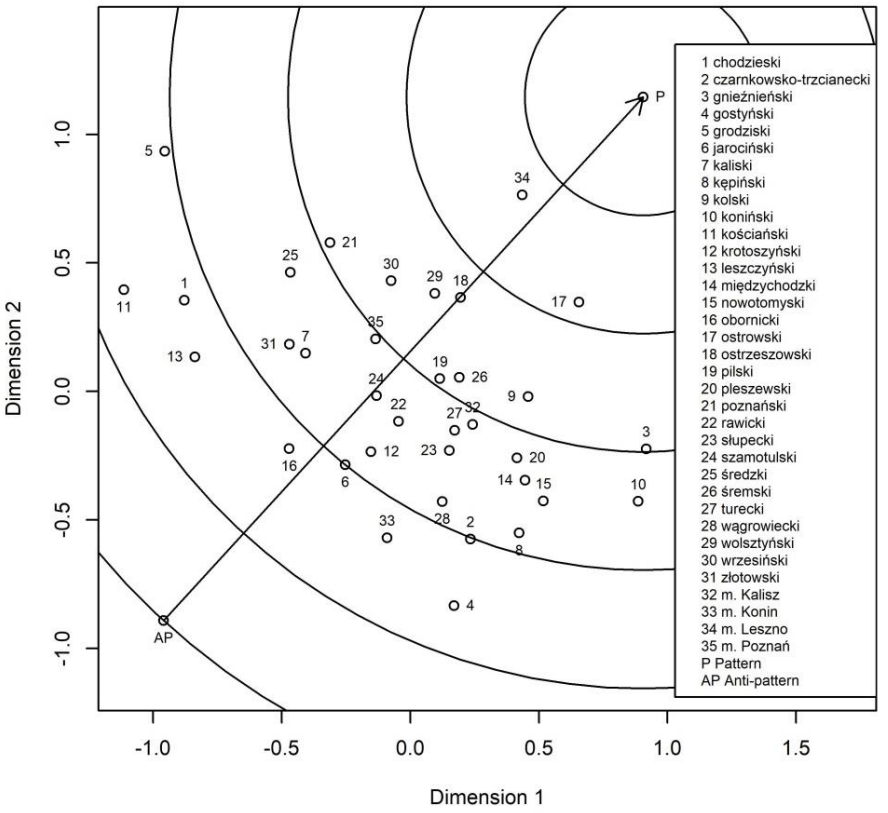

Figure 4. Results of multidimensional scaling of 35 districts of Wielkopolska by economic efficiency of medium-sized manufacturing enterprises in 2012 – the symbolic-to-classic approach

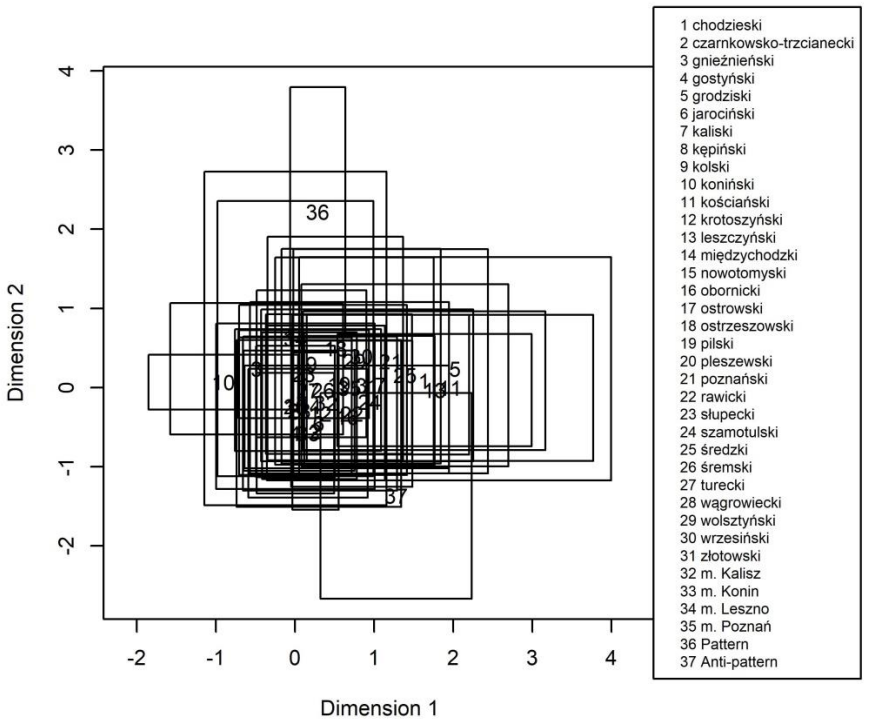

Figure 5. Results of multidimensional scaling of 35 districts of Wielkopolska by economic efficiency of medium-sized manufacturing enterprises in 2012 – the symbolic-to-symbolic approach

By presenting results in this way it is possible to:

- show a graphical ordering of districts in terms of the economic efficiency of manufacturing enterprises measured by three variables according to the values of measure d_i (2),
- distinguish groups of districts with a similar level of economic efficiency (districts between isoquants),
- identify districts characterized by a similar level of economic efficiency, but having a different location on the isoquant of development. Example cases in the classic-to-classic approach include Leszczyński district (13) and Międzychodzki district (14), while in the symbolic-to-classic approach – Grodziski (5) and Jarociński (6) districts. Although the assessment of economic efficiency for these pairs of districts is similar, their respective configurations of values differ.

The visualization of results also reveals that a switch from the classic-to-classic approach to the symbolic-to-classic approach causes a change in the position of objects, and consequently, different assessments of economic efficiency. This is due to the fact that the analysis in the symbolic-to-classic approach is based on the values of the target variables included between the first

and third quartile. At least two directions of changes can be observed in the arrangement of objects. Some objects moved along the set axis or relative to it (an object moving closer to, further away from or crossing the set axis). A large majority of the objects (24) moved towards the pattern (higher values of measure d_i), which, in the symbolic-to-classic approach, represents a higher level of economic efficiency of companies. The group of districts with the highest increase in the value of measure d_i includes those in which companies were assessed as least economically efficient in the classic-to-classic approach: Kępiński, Obornicki, Rawicki (Figures 3, 4 and 6, Table 2). A reverse change, i.e. a shift towards the anti-pattern, was observed for 9 districts, which in the classic-to-classic approach received the highest assessment of economic efficiency of companies: Ostrowski, Wrzesiński, Koniński, Wągrowiecki (Figures 3, 4 and 6, Table 2).

Table 2 shows an ordering of 35 districts of Wielkopolska province depending on the economic efficiency of medium-sized manufacturing enterprises in 2012 obtained under the classic-to-classic, symbolic-to-classic and symbolic-to-symbolic approaches.

Table 2. Ranking of 35 districts of Wielkopolska by economic efficiency of medium-sized manufacturing enterprises in 2012

No.	Districts	d_i^{cc}	Rank	d_i^{sc}	Rank	d_i^{ss}	Rank
1	Chodzieski	0.4153	15	0.2941	31	0.3222	22
2	Czarnkowsko-Trzcianecki	0.1967	31	0.3329	28	0.2553	32
3	Gnieźnieński	0.3901	18	0.5050	10	0.4285	9
4	Gostyński	0.1498	33	0.2358	34	0.2452	34
5	Grodziski	0.4504	13	0.3231	29	0.2921	28
6	Jarociński	0.2367	29	0.3352	27	0.2584	31
7	Kaliski	0.3508	22	0.4030	22	0.3757	11
8	Kępiński	0.0063	35	0.3643	25	0.3185	24
9	Kolski	0.3956	17	0.5480	6	0.4757	5
10	Koniński	0.5640	4	0.4303	19	0.3389	18
11	Kościański	0.2843	28	0.2209	35	0.2530	33
12	Krotoszyński	0.3243	27	0.3682	24	0.2947	27
13	Leszczyński	0.3380	25	0.2712	33	0.2772	29
14	Międzychodzki	0.3303	26	0.4325	18	0.3384	19
15	Nowotomyski	0.3453	23	0.4116	21	0.3331	20
16	Obornicki	0.1459	34	0.2979	30	0.2593	30
17	Ostrowski	0.7186	1	0.6964	2	0.5373	1
18	Ostrzeszowski	0.4999	8	0.6169	3	0.5169	3
19	Pilski	0.4331	14	0.5101	9	0.3959	10
20	Pleszewski	0.3745	19	0.4605	14	0.3174	25

Table 2. Ranking of 35 districts of Wielkopolska by economic efficiency of medium-sized manufacturing enterprises in 2012 (cont.)

No.	Districts	d_i^{cc}	Rank	d_i^{sc}	Rank	d_i^{ss}	Rank
21	Poznański	0.5089	7	0.5124	8	0.4360	7
22	Rawicki	0.1879	32	0.4301	20	0.3053	26
23	Słupecki	0.4028	16	0.4340	17	0.4345	8
24	Szamotulski	0.4589	12	0.4356	16	0.3315	21
25	Średzki	0.3435	24	0.4479	15	0.3715	13
26	Śremski	0.4822	9	0.5270	7	0.3660	14
27	Turecki	0.3600	21	0.4608	13	0.3755	12
28	Wągrowiecki	0.5600	5	0.3635	26	0.3186	23
29	Wolsztyński	0.5276	6	0.5971	4	0.4639	6
30	Wrzesiński	0.5642	3	0.5618	5	0.4794	4
31	Złotowski	0.3626	20	0.3926	23	0.3535	16
32	m. Kalisz	0.4599	11	0.4819	12	0.3425	17
33	m. Konin	0.2204	30	0.2823	32	0.2411	35
34	m. Leszno	0.5921	2	0.7804	1	0.5190	2
35	m. Poznań	0.4607	10	0.4927	11	0.3551	15
Parameters		Value		Value		Value	
Mean		0.3840	X	0.4359	X	0.3579	X
Standard deviation		0.1450	X	0.1234	X	0.0821	X
Median		0.3901	X	0.4325	X	0.3389	X
Median absolute deviation		0.1047	X	0.1150	X	0.0694	X

d_i^{cc} – value of measure (2) in the classic-to-classic approach,
d_i^{sc} – value of measure (2) in the symbolic-to-classic approach,
d_i^{ss} – value of measure (3) in the symbolic-to-symbolic approach,

Source: Calculations performed in the R program (R Core Team 2018) and the clusterSim package (Walesiak, Dudek 2018a).

It can be seen that the application of the robust approaches (symbolic-to-classic and symbolic-to-symbolic) results in a different dispersion of objects. The range of d_i values changed from $[0.0063; 0.7186]$ in the classic-to-classic approach to $[0.2411; 0.5373]$ in the symbolic-to-symbolic approach, while the spread of districts expressed in terms of the standard deviation of measure d_i decreased from $S_{d_i} = 0.1450$ in the classic-to-classic approach to $S_{d_i} = 0.0821$ in the symbolic-to-symbolic approach.

The degree of correlation between the values of measure d_i for 35 districts obtained under each approach was measured by the Pearson correlation coefficient. The consistency of rank orders was measured by the Kendall rank correlation coefficient. The results are shown in Table 3.

Table 3. Correlation coefficients (Pearson's *r* and Kendall's *tau*) between the values of measures (2) and (3) obtained under the three approaches

Pearson correlation coefficient			Kendall rank correlation coefficient		
	cc	sc	ss	cc	sc	ss
cc	1.000	0.692	0.702	1.000	0.546	0.543
sc	0.692	1.000	0.911	0.546	1.000	0.741
ss	0.702	0.911	1.000	0.543	0.741	1.000

cc – classic-to-classic approach,
sc – symbolic-to-classic approach,
ss – symbolic-to-symbolic approach.

The highest degree of similarity between rankings of districts (measured by the Kendall rank correlation coefficient) and correlation between districts (measured by the Pearson correlation coefficient) depending on the values of measure d_i is observed for the approaches based on interval-valued data (symbolic-to-classic and symbolic-to-symbolic). The results based on metric data are considerably different from those obtained using interval-valued data. The latter ones are more reliable (since districts were assessed on the basis of intervals of variable values with the exclusion of outliers) than those based on metric data (where districts were assessed on the basis of the mean values of the target variables).

The results of multidimensional scaling of districts of Wielkopolska province obtained under each approach along with the geographical location are presented in a map chart (Figure 6). One can clearly see the impact of Poznań on the neighbouring districts – it functions as a pole of growth (Isard 1960). Districts located further away from Poznań tend to appear lower in the ranking based on measure d_i. The only exceptions are Ostrowski and Ostrzeszowski districts, which, despite their relatively large distance from Poznań, are characterised by very high values of measure d_i regardless of the approach adopted ($d_{17}^{cc} = 0.7186$, $d_{17}^{sc} = 0.6964$, $d_{17}^{ss} = 0.5373$, and $d_{18}^{cc} = 0.4999$, $d_{18}^{sc} = 0.6169$, $d_{18}^{ss} = 0.5169$, respectively). It should be noted that these districts are part of the Kalisko-Ostrowski Industrial District and are important centres of electromechanical and construction industry. Another factor which may be contributing to the high economic efficiency of companies operating in these districts is that fact that they are located in a special economic zone (Kamiennogórska Subzone).

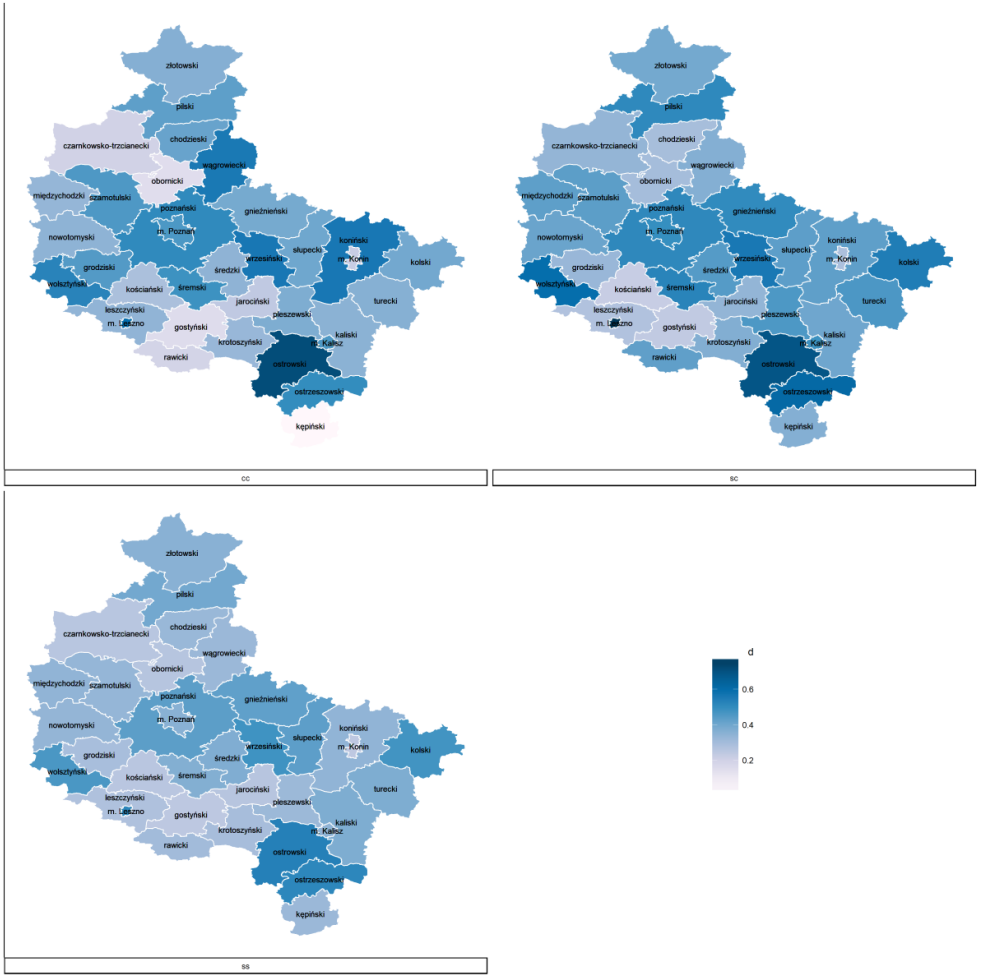

Figure 6. Assessment of districts in terms of economic efficiency of medium-sized manufacturing companies in 2012 in the classic-to-classic (cc), symbolic-to-classic (sc) and symbolic-to-symbolic (ss) approaches

Source: Calculations performed in the R program.

4. Conclusions

The aim of the study was to compare districts of Wielkopolska province in terms of the economic efficiency of medium-sized manufacturing companies, which operated in them in 2012. Variables used in the study are typically used in the financial analysis of economic entities. Assessments were obtained using a hybrid approach combining multidimensional scaling and linear ordering and performed for three types of data set-ups: classic-to-classic, symbolic-to-classic and symbolic-to-symbolic. Thanks to this methodology, it was possible to obtain

a graphic presentation of economic efficiency, which is a multidimensional phenomenon, in a 2-dimensional space. In addition, the districts could be ranked according to the economic efficiency of medium-sized manufacturing companies.

By comparing results obtained under three different data set-ups, it was possible to identify changes caused by switching from the classic-to-classic approach to the interval-based approach (interval-valued data). In the two modified approaches, assessments were not based only on the mean values of the target variables describing companies in each district but accounted for the observed variation. Moreover, companies showing outlying values of the financial variables were excluded from the analysis.

The results were used to identify groups of districts with similar levels of economic efficiency and particular districts within the groups with similar and different values of the target variables. The analysis confirmed the impact of Poznań as a pole of growth on the neighbouring districts.

The authors are aware of the limitations resulting from the selected set of variables. However, the main purpose of the study was to present a new methodological approach.

Acknowledgements

The project is financed by the Polish National Science Centre DEC-2015/17/B/HS4/00905.

REFERENCES

BILLARD, L., DIDAY, E., (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining, John Wiley, Chichester, ISBN: 978-0-470-09016-9.

BORG, I., GROENEN, P. J. F., (2005). Modern Multidimensional Scaling. Theory and Applications. 2nd Edition, Springer Science+Business Media, New York, ISBN: 978-0387-25150-9, URL http://www.springeronline.com/0-387-25150-2.

BORG, I., GROENEN, P. J. F., MAIR, P., (2013). Applied Multidimensional Scaling, Springer, Heidelberg, New York, Dordrecht, London. ISBN 978-3-642-31847-4, URL http://dx.doi.org/10.1007/978-3-642-31848-1.

BORG, I., GROENEN, P. J. F., MAIR, P., (2018). Applied Multidimensional Scaling and Unfolding, Springer, Heidelberg, New York, Dordrecht, London. ISBN 978-3-319-73470-5, URL https://doi.org/10.1007/978-3-319-73471-2.

BORG, I., MAIR, P., (2017). The Choice of initial configurations in multidimensional scaling: local minima, fit, and interpretability, Austrian Journal of Statistics, 46 (2), pp. 19–32,
URL https://doi.org/10.17713/ajs.v46i2.561.

BORYS, T., (1984), Kategoria jakości w statystycznej analizie porównawczej, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, nr 284, Seria:

Monografie i Opracowania nr 23, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław, ISBN: 83-7011-000-0.

BRITO, P., NOIRHOMME-FRAITURE, M., ARROYO, J., (2015). Editorial for special issue on symbolic data analysis, Advances in Data Analysis and Classification, Vol. 9, Issue 1, pp. 1–4,
URL https://dx.doi.org/10.1007/s11634-015-0202-1.

CHABER, P., ŁAPIŃSKI, J., NIEĆ, M., ORŁOWSKA, J., ZAKRZEWSKI, R., WIDŁA-DOMARADZKI, Ł., DOMARADZKA, A., (2017). Raport o stanie sektora małych i średnich przedsiębiorstw w Polsce, Polska Agencja Rozwoju Przedsiębiorczości, Warszawa, URL https://badania.parp.gov.pl/raport-o-stanie-sektora-msp/stan-sektora-msp-w-polsce.

CSO, (2017). Działalność przedsiębiorstw niefinansowych w 2016 r. (Activity of Non-financial Enterprises in 2016), Central Statistical Office of Poland, Warszawa. URL http://stat.gov.pl/obszary-tematyczne/podmioty-gospodarcze-wyniki-finansowe/przedsiebiorstwa-niefinansowe/dzialalnosc-przedsiebiorstw-niefinansowych-w-2016-r-,2,12.html [Accessed 17 July 2018].

DEHNEL, G., (2015). Robust regression in monthly business survey, Statistics in Transition – new series, Vol. 16, No. 1, pp. 1–16.

EVERITT, B.S., LANDAU, S., LEESE, M., STAHL, D., (2011). Cluster Analysis, Wiley, Chichester, ISBN: 978-0-470-74991-3.

GIOIA, F., LAURO, C. N., (2006). Principal component analysis on interval data, Computational Statistics, 21 (2), pp. 343–363,
URL https://doi.org/10.1007/s00180-006-0267-6.

GROENEN, P.J.F. WINSBERG, S., RODRIGUEZ, O., DIDAY, E., (2006), I-Scal: multidimensional scaling of interval dissimilarities, Computational Statistics & Data Analysis, 51 (1), pp. 360–378,
URL http://dx.doi.org/10.1016/j.csda.2006.04.003.

HELLWIG, Z., (1972). Procedure of Evaluating High-Level Manpower Data and Typology of Countries by Means of the Taxonomic Method, [In:] Gostkowski Z. (ed.), Towards a system of Human Resources Indicators for Less Developed Countries, Papers Prepared for UNESCO Research Project, Ossolineum, The Polish Academy of Sciences Press, Wrocław, pp. 115–134.

HELLWIG, Z., (1981). Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych. In: Welfe, W. (ed.), Metody i modele ekonomiczno-matematyczne w doskonaleniu zarządzania gospodarką socjalistyczną, PWE, Warszawa, pp. 46–68, ISBN 83-208-0042-0.

ICHINO, M., YAGUCHI, H., (1994). Generalized Minkowski metrics for mixed feature-type data analysis, IEEE Transactions on Systems, Man, and Cybernetics, 24 (4), pp. 698–708, URL http://dx.doi.org/10.1109/21.286391.

ISARD, W., (1960). Methods of Regional Analysis: An Introduction to Regional Science. Cambridge, Massachusetts: The M.I.T. Press.

JAJUGA, K., WALESIAK, M., (2000). Standardisation of Data Set under Different Measurement Scales, In: Decker, R., Gaul, W., (Eds.), Classification and Information Processing at the Turn of the Millennium, pp. 105–112, Springer-Verlag, Berlin, Heidelberg,
URL http://dx.doi.org/10.1007/978-3-642-57280-7_11.

JAJUGA, K., WALESIAK, M., BĄK, A., (2003). On the General Distance Measure, in Schwaiger, M., Opitz, O., (Eds.), Exploratory Data Analysis in Empirical Research. Berlin, Heidelberg: Springer-Verlag, pp. 104–109,
URL http://dx.doi.org/10.1007/978-3-642-55721-7_12.

KAPLAN, R. S., COOPER R., (1998). Cost & Effect: Using Integrated Cost Systems to Drive Profitability and Performance, Harvard Business School Press, ISBN: 978-0875847887.

KAPLAN, R. S., (2008). Conceptual foundations of the balanced scorecard. In: C. Chapman, A. Hopwood, M. Shields (Eds.), Handbook of Management Accounting Research, Vol. 3, Elsevier, ISBN: 9780080554501.

KOLIŃSKI, A., (2011). Przegląd metod i technik oceny efektywności procesu produkcyjnego, Logistyka, 5, pp. 1083–1091.

MAIR, P., BORG, I., RUSCH, T., (2016), Goodness-of-fit assessment in multidimensional scaling and unfolding, Multivariate Behavioral Research, Vol. 51, No. 6, pp. 772–789,
URL http://dx.doi.org/10.1080/00273171.2016.1235966.

MAIR, P., DE LEEUW, J., BORG, I., GROENEN, P. J. F., (2018). smacof: Multidimensional Scaling. R package ver. 1.10-8,
URL https://CRAN.R-project.org/package=smacof.

MED, (2017). Entrepreneurship in Poland, Ministry of Economic Development, Warsaw, URL https://www.mpit.gov.pl/strony/zadania/analiza-i-ocena-polskiej-gospodarki/przedsiebiorczosc/.

R CORE TEAM, (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org.

WALESIAK, M., (2016). Visualization of linear ordering results for metric data with the application of multidimensional scaling, Ekonometria [Econometrics], 2 (52), pp. 9–21, URL http://dx.doi.org/10.15611/ekt.2016.2.01.

WALESIAK, M., DEHNEL, G., (2018). Evaluation of Economic Efficiency of Small Manufacturing Enterprises in Districts of Wielkopolska Province Using Interval-Valued Symbolic Data and the Hybrid Approach. In M. Papież and S. Śmiech (Eds.), The 12th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings, Foundation of the Cracow University of Economics, Cracow, pp. 563-572,
URL http://dx.doi.org/10.14659/SEMF.2018.01.57.

WALESIAK, M., DUDEK, A., (2017). Selecting the optimal multidimensional scaling procedure for metric data with R environment, Statistics in Transition – new series, 18 (3), pp. 521–540,
URL http://dx.doi.org/10.21307/stattrans-2016-084.

WALESIAK, M., DUDEK, A., (2018a). clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package, version 0.47-2,
URL https://CRAN.R-project.org/package=clusterSim.

WALESIAK, M., DUDEK, A., (2018b). mdsOpt: Searching for Optimal MDS Procedure for Metric and Interval-valued Symbolic Data, R package, version 0.3-2, URL https://CRAN.R-project.org/package=mdsOpt.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 69–84, DOI 10.21307/stattrans-2019-015

ECONOMIC GROWTH AND ITS DETERMINANTS: A CROSS-COUNTRY EVIDENCE

Adedayo A. Adepoju[1], **Tayo P. Ogundunmade**[2]

ABSTRACT

Empirical evidence from a panel of 126 countries, over the time period of 2010 to 2014, indicates that economic growth is dependent on various factors. This paper finds that government expenditure control, reduced inflation and increased trade openness are the factors that boost the economic growth of a country. Significant evidence is seen for government consumption, fiscal policy and trade openness. No significant relationship has been observed between exchange rate and economic growth, whereas unemployment influences output for African countries. The cross regional analysis of Asian, European, African, Caribbean, and American countries gives specific determinants for these regions. Economic growth is also analysed in developing, developed, least developed, Muslim and petroleum exporting and emerging countries.

The results of this study validate the dependence of economic growth on various factors. Fiscal balance has shown a consistent positive relationship with economic growth throughout the analyses. Fiscal balance and unemployment rate played their role in the growth of African countries. Inflation rates and increased openness were significant for some regions. Exchange rate did not return significant coefficients for any of the sub-regions. Government consumption, trade openness, policy interest rate and industrial production rate showed significant effect for different regions of the world.

Key words: economic growth, panel data analysis, growth determinants.

1. Introduction

Economic growth, no doubt, is the backbone of an economy's development and its enhancement remains one of the major strategic and policy issues for the policymakers. Researchers, over the years, have analysed the economic growth and its development; special emphasis has been laid upon the factors that influence the economic growth. A vast body of economic literature has, empirically and Bayesian researched the economic growth and its determinants (Kormendi and Meguire 1985; Barro, 1990, 1995, 1996, 1997; Sachs and Warner 1997). These studies have identified several factors, having empirical and

[1] Department of Statistics, University of Ibadan, Oyo State, Nigeria. E-mail: pojuday@yahoo.com. ORCID ID: https://orcid.org/0000-0003-2368-4313.
[2] Department of Statistics, University of Ibadan, Oyo State, Nigeria. E-mail: ogundunmadetayo@yahoo.com. ORCID ID: https://orcid.org/0000-0002-0160-3896.

Bayesian backing, which impact economic growth of a country. Many researchers consider that the most promising approach to accounting for model uncertainty is to employ model averaging techniques. This approach allows constructing parameter estimates that formally address the dependence of model specific estimates on a given model. The studies relating to economic growth have used cross-sectional, time-series and panel data models for their analyses. This study has focused on panel/longitudinal (cross-sectional time-series) data to investigate the relationship. This study utilizes panel data for 126 countries over the time period of 5 years in order to determine the impact of fiscal policy, government consumption, inflation, trade openness, policy interest rate, industrial production, unemployment and public debt on the economic growth.

Should cross-country growth evidence be discounted? Are there no growth determinants that are robust to variable selection? Carmen Fernandez, Eduardo Ley, and Mark F. Steel (2001b) and Xavier Sala-i-Martin, Gernot Doppelhofer, and Ronald I. Miller (2004) propose to answer these questions using Bayesian model averaging.

Literature vastly contains evidence on the relationship between economic growth and the factors influencing it. Barro (1996b) identified various factors which enhance the real per capita GDP growth rate. These factors include low government consumption, low inflation and rule of law. Various other factors which influence growth are greater life expectancy level (indicator for health), higher schooling levels (indicator for human capital) and better trade terms. Drury, Krieckhaus and Lusztig (2006) found insignificant relationship between economic growth and population growth; and between economic growth and life expectancy. Barro (1996a) found significant effects of rule of law, openness, less government consumption and increased human capital; in growth determination.

Literature vastly contains evidence on the relationship between economic growth and the factors influencing it. Barro (1996b) identified various factors which enhance the real per capita GDP growth rate. These factors include low government consumption, low inflation and rule of law. Various other factors which influence growth are greater life expectancy level (indicator for health), higher schooling levels (indicator for human capital) and better trade terms. Drury et al (2006) found insignificant relationship between economic growth and population growth; and between economic growth and life expectancy. Barro (1996a) found significant effects of rule of law, openness, less government consumption and increased human capital in growth determination. Kormendi and Meguire (1985) found a negative relation between inflation and growth rate but the explanatory power becomes insignificant when investment rate is also included, indicating inflation directly affects investment and may be less relevant in the capital growth. Cozier and Selody (1992) also estimated that the effect of inflation on income is negative for OECD. Barro (1995, 1996) has also obtained similar results for inflation, a negative long-run effect of inflation on growth.

Dewan and Hussein (2001) used a sample of 41 middle-income developing countries to develop an empirical model for growth. The study also presents a wide-ranging examination of both theoretical and empirical evidence on the many ways macroeconomic policies affect growth. The results suggest that apart from

growth in the labour force, investment in both physical and human capital, as well as low inflation and open trade policies are necessary for economic growth. Furthermore, the ability to adopt technological changes in order to increase efficiency is also important. Since many developing countries have a large agricultural sector, adverse supply shocks in this sector was found to have a negative impact on growth. Yanikkaya (2003) notes that there are different measures of measuring trade openness that can be found in literature. Many researchers have used the simple measure of trade openness (exports plus imports divided by GDP), whereas others have used different other available measures. Using the simple measure, Harrison (1996) reports that researchers have found robust positive relationship between trades share in GDP and economic growth. Machi (2011) empirically test the determinants of economic growth in Nigeria using time series data ranging from 1970 to 2008 and adopting the Johansens method of co-integration-regression analysis. The findings showed that policies that encourage investments in physical capital, human capital, man power development, training, research and technological development would boost both short run and long term growth of the economy. Hence policy tools such as fiscal, monetary and income-price policies should be used by the government to achieve economic growth in Nigeria. Sabir and Tahir (2012) study the impact of different macroeconomic variables on the welfare of the poor in Pakistan, through annual time series data which spanned between 1981 and 2010. Using multiple regression technique to detect the relation between macroeconomic variables and poverty, the findings revealed that GDP growth rate per capita income, major crops, minor crops and livestock had negative impact while inflation and population growth rate had positive impact on poverty and concluded that reduction in poverty in Pakistan is driven by changes in the macroeconomic variables.

Zafar and Zahid (2013) examined the effects of some of the key macroeconomic variables on economic growth. Employing multiple regression framework and time series data over the period 1959-60 to 1996-97. The quantitative evidence shows that primary education is an important precondition for accelerating growth. Similarly, increasing the stock of physical capital and openness of the economy contribute to growth. The empirical results also suggested that budget deficit and external debt is negatively related to economic growth, suggesting that relying on domestic resources is the best alternative to finance growth and reinforce the importance of sensible long-run growth-oriented policies to obtain sustainable growth. The objective of this study is to identify the determinants of economic growth of selected countries using panel data regression approach. Also, to identify the variables that contribute to economic growth in the Developed, Developing, Least developed, Asian, Carribean, Tropical, Petroleum Exporting, Emerging market, European, American and Muslim countries (Abdulalh 2012).

The rest of the paper is structured as follows: Section 2 explains the model specification. Section 3 shows the empirical analysis of the study. Section 4 concludes the paper.

2. The model

Time series cross-sectional (panel) data of 126 countries has been used in the analysis. The annual time period ranges from 2010 to 2014.

Consider a linear regression model with a constant term, and k potential explanatory variables

$Y_{it} = (X^{j}{}_{it})\ \beta + e_{it}$

i = 1, 2, 3… N; t = 1, 2, 3… T where,

Y_{it} = GDP per capital

X_{it} = the value of the jth explanatory variables for unit i at time t. There are k explanatory variables indexed by j = 1... k. The variables considered are Policy Interest Rate, Industrial Production, Trade Openness, Unemployment Rate, Exchange Rate, Public Debt, Fiscal Balance and Inflation rate.

Real GDP growth, representing economic growth, is the dependent variable. Data for all the variables used were obtained from the World Bank World Development Indicators (WDI) database. GDP per capita stands for Gross Domestic Product (GDP) per capita (per person), and it is derived from a straightforward division of total GDP by the population. Per capita GDP is typically expressed in local current currency, local constant currency or a standard unit of currency in international markets, such as the U.S. dollar (USD). Initial real GDP for a particular year was also used as an independent variable.

GDP per capita is an important indicator of economic performance and a useful unit to make cross-county comparisons of average living standards and economic wellbeing. However, GDP per capita is not a measure of personal income and using it for cross-country comparisons also has some known weaknesses. In particular, GDP per capita does not take into account income distribution in a country. In addition, cross-country comparisons based on the U.S. dollar can be distorted by exchange rate fluctuations and often do not reflect the purchasing power in the countries being compared.

Industrial production measures the output of the industrial sector, which typically comprises mining, manufacturing, utilities and, in some cases, construction. The industrial production indicator is generally provided as an index in volume terms.

Inflation refers to an overall increase in the Consumer Price Index (CPI), which is a weighted average of prices for different goods. The policy interest rate is an interest rate that the monetary authority (i.e. the central bank) sets in order to influence the evolution of the main monetary variables in the economy (e.g. consumer prices, exchange rate or credit expansion, among others).

The policy interest rate is an interest rate that the monetary authority (i.e. the central bank) sets in order to influence the evolution of the main monetary variables in the economy (e.g. consumer prices, exchange rate or credit expansion, among others).

Other factors that contribute to economic growth is Public debt, sometimes also referred to as government debt, represents the total outstanding debt (bonds and other securities) of a country's central government. It is often expressed as a ratio of Gross Domestic Product (GDP).

Private consumption also referred to as personal consumption, consumer expenditure, or also referred to as personal consumption, consumer expenditure, or personal consumption expenditures (PCE), measures consumer spending on goods and services.

Fiscal balance, sometimes also referred to as government budget balance, is calculated as the difference between a governments revenues (taxes and proceeds from asset sales) and its expenditures. It is often expressed as a ratio of Gross Domestic Product (GDP). If the balance is positive, the government has a surplus (it spends less than it receives). If the balance is negative, the government has a deficit (it spends more than it receives). Fiscal balance as a percentage of GDP is used as an instrument to measure a government's ability to meet its financing needs and to ensure good management of public finances.

The unemployment rate is defined as the percentage of unemployed workers in the total labour force. Workers are considered unemployed if they currently do not work, despite the fact that they are able and willing to do so. The total labour force consists of all employed and unemployed people within an economy. The unemployment rate provides insights into the economy's spare capacity and unused resources. Unemployment tends to be cyclical and decreases when the economy expands as company's contract more workers to meet growing demand. Unemployment usually increases as economic activity slows.

3. Analysis

Panel data for 126 countries over the period of 2010– 2014 yielding a panel with $N = 126$ and $T = 5$. Clearly, our pool is cross-sectional dominant ($N > T$). Separate regressions were run for the complete sample and then for sub-samples consisting of developing, developed, least developed, petrol exporting, emerging, Caribbean, Asian, European, American region, African and Muslim countries; in order to get an insight into the relevant determinants of economic growth for these sub-samples. A panel regression analysis was used to run the separate regressions as it better used over ordinary least squares when the data is cross-sectional and time dominant.

Table I gives the descriptive statistics of the variables under consideration. It gives the minimum, 1st quartile, median, mean, 3rd quartile and the maximum values of each of the variables across the countries under consideration.

Table 1. Summary Statistics

Variables	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
GDP per capital	306	3360	8062	18370	25810	125700
Consumption rate	-8.72663	0.05884	0.025030	0.90396	0.89000	35.8000
Exchange rate	0.010	1.380	7.535	816.819	106.175	28050.00
Fiscal balance	-32.300	-4.475	-2.600	-2.318	-0.800	34.500
Industrial production	-15.264	0.01402	0.18078	0.96809	0.84002	58.59155
Inflation rate	-35.100	-1.900	3.400	3.777	8.500	62.200
Public debt	0.10	25.25	38.40	47.22	60.65	212.00
Unemployment rate	0.300	4.425	7.100	8.931	11.400	40.600

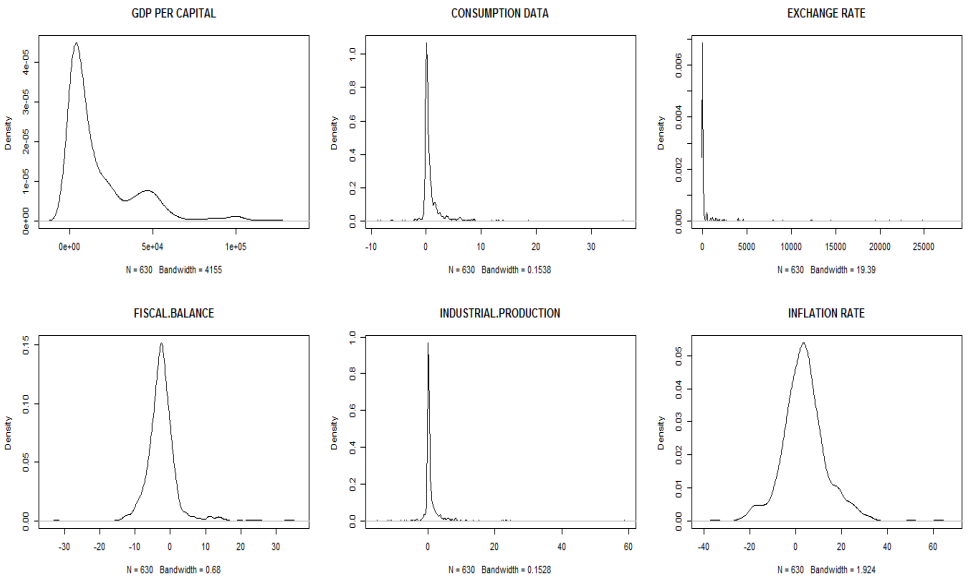

Figure 1. Density plot of the variables

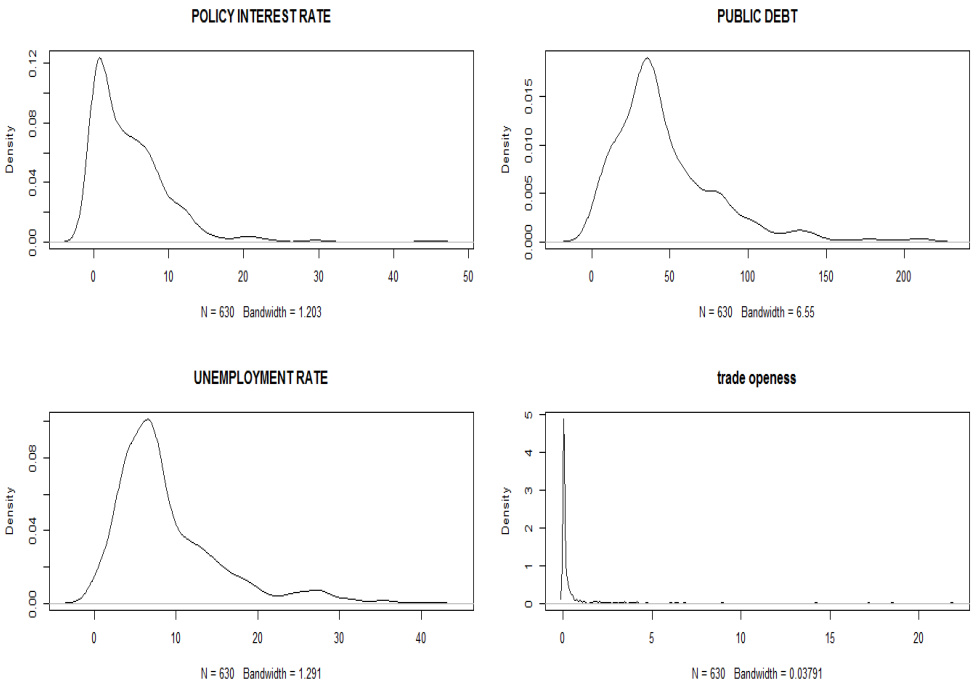

Figure 2. Density plot of the variables

Figures 1 and 2 show the density plot of the variables used in the analysis. Each of the variable was plotted to know the shape of the data.

Table 2. Results for Economic Growth and its Determinants

Variables	(a)	(b)	(c)	(d)	(e)	(f)
Consumption rate	4.47e-05*	0.9693	0.99708	0.1425	0.006*	0.018*
Exchange rate	0.57873	0.3002	0.3963	0.4468	0.4062	0.881
Fiscal balance	7.93e-10*	1.58e-05*	0.0446*	0.002*	0.6648	0.359
Industrial production	0.002*	0.9839	0.3731	0.7225	0.1833	0.441
Inflation rate	0.1245	0.2458	0.5649	0.4967	0.0836	0.86
Policy interest rate	0.1078	0.1681	0.5884	0.8617	0.0388*	0.831
Public debt	0.4348	0.6643	0.7787	0.7357	0.4348	0.905
Unemployment rate	0.6467	0.3272	0.593	0.3766	0.007*	0.248
Trade openness	0.00256*	0.0277*	0.003*	0.9891	0.4453	0.268
R^2	0.34	0.24	0.32	0.31	0.28	0.23

Table 2. Results for Economic Growth and its Determinants (cont.)

Variables	(g)	(h)	(i)	(j)	(k)	(l)	(m)
Consumption rate	5.9e-11*	0.00238*	0.95	0.0069*	0.9217	0.9356	0.9287
Exchange rate	0.1923	0.6291	0.763	0.6353	0.668	0.3638	0.8676
Fiscal balance	0.0005*	4.65e-07*	0.0446*	1.38e-05*	9.15e-10*	0.9314	4.28E-10*
Industrial production	0.3835	0.01762*	0.304	0.3918	0.6701	0.413	0.6159
Inflation rate	0.5779	0.89085	0.298	0.4441	0.4173	0.8191	0.7873
Policy interest rate	0.7822	0.37666	0.164	0.3717	0.8164	0.4226	0.1559
Public debt	0.4364	0.70152	0.179	0.593	0.9516	0.7786	0.4479
Unemployment rate	0.8032	0.12014	0.16	0.2214	0.1549	0.342	0.5642
Trade openness	0.0235*	0.08373	0.032*	0.3599	0.0007*	0.9026	0.0261*
R^2	0.13	0.15	0.28	0.21	0.21	0.19	0.15

(*) implies significance

Table 2 above contains the significance value of economic variables. Columns (a) represents the analysis for the complete panel; Columns (b),(c),(d), (e) and (f),(g), (h),(i),(j), (k) and (l) shows analyses for developed countries, developing countries, least developed countries, African countries and American countries, Emerging market, European countries, Asian countries, Tropical countries, Petroleum countries, Caribbean countries and Muslim Countries respectively.

3.1. Economic growth in the complete panel

Table 2 (Column a) gives the results of the regression for the complete panel of 126 countries. The R^2 statistic, 0.34 (34%), is not very strong. Low R^2 values have also been reported by Drury (2006) and Abdullah (2012) for the similar analysis. Government consumption (significant), Fiscal balance (significant), industrial production (significant) and trade openness (significant) return positive coefficients except government consumption(significant); which indicates that all these variables have positive impact on economic growth. The significance of the fiscal balance, government consumption rate, openness, inflation and industrial production variables contribute to economic growth.

An increase in the level of government consumption rate of a country, economic growth tends to be affected; complementing the results of Mauro (1995). Therefore, in order to boost a country's economic growth, government consumption rate should be minimized.

Government expenditure is also seen to impact the economic growth. Barro (1990) mentioned the dependence of long run growth on the structure of government expenditure. Barro (1997) mentioned that government consumption retards growth; our analysis also indicated that government consumption, affects growth.

Openness is found to have a positive impact on a country's economic growth. Harrison

(1996) has also observed a similar positive relation between openness and growth. Open international market boost a country's economic growth and open economies tend to grow more rapidly as compared to those whose trade have restrictions.

Industrial production casts a positive effect on the economic growth; complementing the results of Kormendi and Meguire (1985), Cozier and Selody (1992) and Barro (1995, 1996). Fiscal balance also impacts economic growth indicating higher fiscal policy which boost economic growth; accepting the view of Matsuyama (1992).

Generalizing the results, it can be concluded that decreased government consumption rate, increased openness, increased industrial production and a moderate trade openness will enhance the economic growth of a country.

3.2. Economic growth in developed countries

The analysis of developed countries in Table 2 (Column b) shows that only fiscal balance and trade openness relate to the economic growth. The R^2 statistic for the regression is 0.24 (24%). Fiscal balance yields a positive coefficient, as expected; whereas trade openness returns a positive coefficient. Drury, Krieckhaus and Lusztig (2006) found an insignificant relationship between trade openness and economic growth; however for our study of developed countries, the relationship is significant. In general, for developed countries, high fiscal policy trade openness enhance economic growth.

3.3. Economic growth in developing countries

Similar method, to the one presented for the complete panel, is used to analyze the economic growth in the developing countries in Table 2 (Column c).

The R^2 statistic for the regression is 0.32 (32%). Fiscal balance and trade openness give positive coefficients, as is the case in the complete panel. The only difference in the results for the complete panel and developing countries is that fiscal balance, although having a negative coefficient, is found to be significant.

Generalizing the results for developing countries; we conclude that increased openness to trade and fiscal balance contribute significantly, towards the economic growth in a developing country.

3.4. Economic growth in least developed countries

The analysis for least developed countries is presented in Table 2 (Column d) only fiscal balance returns significant. The R^2 statistic for the regression is 0.31 (31%).

Most of the least developed countries are mostly dependent on the manpower for output.

So, the positive relationship between fiscal balance and economic growth, as indicated by Barro (1996), shows that high fiscal balance can lead to higher economic growth.

3.5. Economic growth in African countries (East, West and Central African Countries)

The result for African countries (Table 2 Column e) show significant relationship between government consumption (negative), unemployment rate (negative), policy interest rate and economic growth. The R^2 statistic for the regression is 0.28 (28%). This implies that unemployment rate has adverse effect on economic growth in African countries.

In general, low unemployment rate and low government consumption rate will contribute to a higher economic growth in African countries.

3.6. Economic growth in American region countries (North, South and Central American Countries)

The analysis for the American region countries (Table 2 Column f) only returns negative government consumption. The R^2 statistic for the regression is 0.23 (23%). Hence, for American region countries, low government consumption will route towards a successful economic growth.

3.7. Economic growth in emerging markets/countries

For emerging countries (Table 2 Column g); the analysis shows a positive coefficient for trade openness, government consumption rate and a positive coefficient for fiscal balance. The positive coefficient for fiscal balance is consistent with our findings for other regions. The positive coefficient for trade openness is also consistent with the findings of Barro (1996, 1997) and consistent with those of Aschauer (1990). Government may allocate resources to the effective and required sectors in order to boost up economic growth. The R^2 statistic for the regression is 0.13 (13%). So, in the case of emerging markets/countries; high fiscal policy and high trade openness lead to better economic growth.

3.8. Economic growth in European countries

Analysis for European countries (Table 2 Column h) shows that Government consumption rate, openness and industrial production significantly impact economic growth. The positive fiscal balance, positive coefficient for openness and a negative coefficient for government consumption are all aligned with the literature on economic growth. The R^2 statistic for the regression is 0.15 (15%). Generalizing the results for European countries; reduced government consumption, increased openness and increase industrial production contribute to a sound economic growth.

3.9. Economic growth in Asian countries

For Asian countries (Table 2 Column i); fiscal balance (positive coefficient) and openness (positive coefficient) give significant coefficients in the regression analysis. The R^2 statistic for the regression is 0.28 (28%). The positive coefficient for openness and positive coefficient for fiscal balance are consistent with the theory on these coefficients; as mentioned in the above analyses. So, high fiscal balance and increased trade openness contribute to a high economic growth in Asian countries.

3.10. Economic growth in Tropical countries

For Tropical countries (Table 2 Column j), the regression analysis with PCSEs (Panel Corrected Standard Error) returns two significant variables. Fiscal balance (positive coefficient) and government consumption rate show significant result to economic growth. The positive coefficient shows high trading in this region and this leads to a higher economic growth in Tropical countries. The R^2 statistic for the regression is 0.21 (21%).

Generalizing the results for Tropical countries; high government consumption and high trade openness lead to high economic growth.

3.11. Economic growth in petroleum exporting countries

For petroleum exporting countries (Table 2 Column k), the regression analysis with PCSEs returns two significant variables. Fiscal balance (positive coefficient) and trade openness (positive coefficient) show significant result to economic growth. The positive coefficient for trade openness indicates that higher trade openness leads to a higher economic growth in petroleum exporting countries. The R^2 statistic for the regression is 0.21 (21%).

Generalizing the results for petroleum exporting countries; high fiscal policy and high trade openness lead to high economic growth.

3.12. Economic growth in Caribbean countries

For Caribbean countries (Table 2 Column l), no variable return significant. The R^2 statistic for the regression is 0.19 (19%). Generalizing the results for Caribbean countries, no variable contributes to economic growth.

3.13. Economic growth in Muslim countries

Regression analysis for Muslim countries (Table 2 Column m) returns significant coefficients for trade openness (positive coefficient) fiscal balance (positive coefficient).

The R^2 statistics for the regression is 0.15(15%). Positive coefficient for trade openness indicates the need for a trade-free environment to prevail in the Muslim countries in order to attain a higher economic growth. Positive relationship of fiscal balance with economic growth shows that more fiscal measures will yield higher levels of economic growth.

In general, for Muslim countries; high fiscal balance and high trade openness will bring higher economic growth.

4. Conclusion

For a broad panel of 126 countries, this paper investigated the relationship between economic growth and various variables which have strong theoretical support of affecting economic growth of a country. Thirteen separate regression analyses were conducted to check the impact of the variables on economic growth in different regions, cultures and classifications of the world.

Fiscal balance, throughout our analysis, returned positive coefficients; indicating that fiscal balance positively affects the economic growth of a country, irrespective of the location and status of the country. Unemployment rate only showed its significant coefficient for African countries, indicating the fact that unemployment rate will have better prospects of affecting economic growth in a country. Policy interest rate was also seen to positively impact the economic growth for least developed countries, showing its contribution in the country to boost up the economic output. Government consumption, fiscal balance, trade openness variables led to a mixed relationship with economic growth, positive for some of the regions whereas negative for other regions. Trade openness positively impacted economic growth for most of the regions, indicating that a country with open access to its trade is expected to have higher economic growth. Inflation, on the other hand, returned negative coefficients for most of our analyses.

This study makes several contributions to the existing knowledge on economic growth.

First, a very wide panel of 126 countries is used for the analysis. Second, separate regression analysis for developing countries, developed countries, least developed countries, petroleum exporting countries, emerging markets/countries, Caribbean countries, Asian countries, European countries, American region countries, African countries and Muslim countries were run. This gives an understanding of the relationship of economic growth and the variables under consideration for different regions and classifications of the world. Third, we have employed a variety of variables which had strong theoretical backing based on existing literature. Fourth, our results may help policy makers to focus on the specified areas that support the economic growth in a country or a region.

The results of the study present important implications for policy makers. Economists and relevant policymakers can use the analysis to have an insight

into the economic growth factors prevailing in the whole world (referring to the complete sample) and the ones having vital influence for the sub-samples analysis (referring to the regional analysis). The empirical results of the study can be essential for the direction of policies towards relevant factors that play significant roles in the enhancement and the development of the economy.

Future research should consider other relevant explanatory variables like labour force and investment (gross capital formation) and income inequality. Also, a causality analysis may be conducted for understanding the relationship between economic growth and its significant determinants.

REFERENCES

ACHEN, C. H., (2000). Why Lagged Dependent Variables can Suppress the Explanatory Power of other Independent Variables, Paper presented at the annual meeting of the Political Methodology Section of the American Political Science Association, Los Angeles, CA, July 20–22.

ASCHAUER, D. A., (1990). Is Government Spending Stimulative?, Contemporary Economic Policy, 8 (4), pp. 30–46.

AZARIADIS, C., LAHIRI, A., (1997). Do Rich Countries Choose Better Governments?, Working paper, Department of Economics, UCLA.

BARRO, R. J., (1990). Government Spending in a Simple Model of Endogenous Growth, Journal of Political Economy, 98, pp. 103–125.

BARRO, R. J., (1995). Inflation and Economic Growth, Bank of England Economic Bulletin 35, pp. 1–11.

BARRO, R. J., (1996a). Democracy and Growth, Journal of Economic Growth 1, pp. 1–27.

BARRO, R. J., (1996b). Determinants of Economic Growth: A Cross-Country Empirical Study, NBER Working Paper, No. 5698, Cambridge, Mass.: National Bureau of Economic Research.

BARRO, R. J., (1997). Determinants of Economic Growth: A Cross-Country Empirical Study, Cambridge, MA: MIT Press.

BENHABIB, J., SPIEGEL, M. S., (1994). The Role of Human Capital in Economic Development: Evidence from Aggregate Cross-Country Data, Journal of Monetary Economics, 34, pp. 143–173.

BRUMM, H. J., (1997). Military Spending, Government Disarray, and Economic Growth: A Cross-Country Empirical Analysis, Journal of Macroeconomics, 19, pp. 827–838.

CHOWDHURY, A. R., (1991). A Causal Analysis of Defense Spending and Economic Growth, Journal of Conflict Resolution, 35, pp. 80–97.

COZIER, B., SELODY, J., (1992). Inflation and Macroeconomic Performance: Some Cross-Country Evidence, Working Paper, No. 92–06, Ottawa: Bank of Canada, Department of Monetary and Financial Analysis.

DEVARAJAN, S., SWAROOP, V., ZOU, H-F., (1996). The Composition of Public Expenditure and Economic Growth, Journal of Monetary Economics, 37, pp. 313–344.

DRURY, A. C, KRIECKHAUS, J., LUSZTIG, M., (2006). Corruption, Democracy and Economic Growth, International Political Science Review, 27 (2), pp. 121–136.

HARRISON, A., (1996). Openness and Growth: A Time Series, Cross-Country Analysis for Developing Countries, Journal of Development Economics, 48, pp. 419–447.

HSIEH, E., LAI, K. S., (1994). Government Spending and Economic Growth: The G-7 Experience, Applied Economics, 26, pp. 535–542.

KAUFMANN, D., AART, K., MASTRUZZI, M., (2003). Governance Matters III: Governance Indicators for 1996–2002, World Bank Policy Research Working Paper, No. 3106, Washington, D.C.

KIM, H., (1996). Trade-offs between Military Spending, Quality of Life and Economic Growth, Comparative Economic Studies, 38, pp. 69–84.

KLEIN, T., (2004). Military Expenditure and Economic Growth: Peru 1970–1996, Journal of Defense and Peace Economics, 15, pp. 275–287.

KORMENDI, R. C., MEGUIRE, P. G., (1985). Macroeconomic Determinants of Growth: Cross-Country Evidence, Journal of Monetary Economics, 16 (2), pp. 141–163.

LIM, D., (1983). Another Look at Growth and Defense in Less Developed Countries, Economic Development and Cultural Change, 31, pp. 377–384.

LIPSET, S. M., (1959). Some Social Requisites of Democracy, American Political Science Review, 53, pp. 69–105.

LIPSET, S. M., (1960). Political Man: The Social Bases of Politics, New York: Doubleday.

MATSUYAMA, K., (1992). Agricultural Productivity, Comparative Advantage and Economic Growth, Journal of Economic Theory, 58 (2), pp. 317–334.

MAURO, P., (1995). Corruption and Growth, Quarterly Journal of Economics, 110, pp. 681–712.

SACHS, J. D., (2001). Tropical Underdevelopment, NBER Working Paper, No. W8119.

SACHS, J. D., WARNER, A. M., (1995). Natural Resource Abundance and Economic Growth, NBER Working Paper, No. W5398.

SACHS, J. D., WARNER, A. M., (1997). Sources of Slow Growth in African Economies, Journal of African Economies, 6, pp. 335–376.

YANIKKAYA, H., (2003). Trade Openness and Economic Growth: A Cross-Country Empirical Investigation, Journal of Development Economics, 72, pp. 57–89.

APPENDIX

Complete Panel of 126 Countries:
(D, D* and LD represent countries used in the analysis as developed countries, developing countries and least developed countries)
Albania(D*), Algeria(D*), Angola(LD), Argentina(D*), Armenia(D*), Australia(D), Austria(D), Azerbaijan(D*), Bahrain(D), Bangladesh(LD), Barbados(D*), Belarus(D*), Belgium(D), Belize(D*), Bolivia(D*), Bosnia and Herzegovina(D*), Botswana(D*), Brazil(D*), Brunei(D*), Bulgaria(D*), Cambodia(LD), Cameroon(D*), Canada(D), Chile(D*), China(D*), Colombia(D*), Costa Rica(D*), Cote d'Ivoire(D*), Croatia(D*), Cyprus(D), Czech Republic(D), Democratic Republic of Congo(D*), Denmark(D), Dominican Republic(D*), Ecuador(D*), Egypt(D*), El Salvador(D*), Estonia(D), Ethiopia(LD), Finland(D), France(D),Georgia(D*), Germany(D), Ghana(D*), Greece(D), Grenada(D*), Guatemala(D*), Guinea(LD), Haiti(LD), Honduras(D*), Hungary(D*), Iceland(D), India(D*), Indonesia(D*), Iran(D*), Iraq(D*), Ireland(D), Israel(D), Italy(D), Jamaica(D*), Japan(D), Jordan(D*), Kazakhstan(D*), Kenya(D*), Kosovo(D*), Kuwait(D), Kyrgyzstan(D*), Latvia(D*), Lebanon(D*), Lesotho(LD), Liberia(LD), Libya(D*), Lithuania(D*), Luxembourg(D), Macedonia(D*), Malaysia(D*), Mali(LD), Malta(D), Mexico(D*), Moldova(D*), Mongolia(D*), Montenegro(D*), Morocco(D*), Mozambique(LD), Netherlands(D), New Zealand(D), Nicaragua(D*), Niger(LD), Nigeria(D*), Norway(D), Oman(D*), Pakistan(D*), Panama(D*), Paraguay(D*), Peru(D*), Poland(D*), Portugal(D), Puerto Rico(D), Qatar(D), Russia(D*), Saudi Arabia(D*), Serbia(D*), Singapore(D), Slovakia(D), Slovenia(D), South Africa(D*), South Korea(D), Spain(D), Sri Lanka(D*), Swaziland(D*), Sweden(D), Switzerland(D), Tajikistan(D*), Tanzania(LD), Thailand(D*),Togo(LD), Trinidad and Tobago(D*), Tunisia(D*), Turkey(D*), Turkmenistan(D*), Uganda(LD), Ukraine(D*), United Arab Emirates, United Kingdom(D), United States(D), Uruguay(D*), Uzbekistan(D*), Venezuela(D*), Vietnam(D*), Yemen(LD), Zambia(LD).

List of Petroleum Exporting Countries:
Algeria, Angola, Australia, Bahrain, Brunei, Canada, China, Colombia, Gabon, Indonesia, Iran, Iraq, Kuwait, Malaysia, Mexico, Nigeria, Oman, Qatar, Russia, Saudi Arabia, Trinidad and Tobago, United Arab Emirates, Venezuela, Yemen.

List of Emerging Market / Countries:
Brazil, Chile, China, Colombia, Czech Republic, Egypt, Hungary, India, Indonesia, Korea, Malaysia, Mexico, Morocco, Peru, Philippines, Poland, Russia, South Africa, Taiwan, Thailand, Turkey.

List of Caribbean Countries:
Belize, Dominican Republic, Haiti, Jamaica, Puerto Rico, Trinidad and Tobago.

List of Asian Countries:
Armenia, Azerbaijan, Bahrain, Bangladesh, Brunei, Cambodia, China, India, Indonesia, Iran, Iraq, Israel, Japan, Jordon, Kazakhstan, Kuwait, Kyrgyzstan, Lebanon, Malaysia, Mongolia, Nepal, Oman, Pakistan, Philippines, Qatar, Russia,

Saudi Arabia, Singapore, Sri Lanka, Tajikistan, Thailand, Turkey, Turkmenistan, United Arab Emirates, Uzbekistan, Vietnam, Yemen.

List of European Countries:
Albania, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Macedonia, Malta, Moldova, Montenegro, Netherlands, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, United Kingdom.

List of American Region Countries:
Argentina, Barbados, Belize, Bolivia, Brazil, Canada, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Trinidad and Tobago, United States, Uruguay, Venezuela.

List of African Countries:
Algeria, Angola, Cameroon, Cote d'Ivoire, Ethiopia, Ghana, Kenya, Madagascar, Mauritania, Mozambique, Nigeria, Rwanda, Tanzania, Togo, Uganda, Zambia, Zimbabwe.

List of Muslim Countries:
Albania, Algeria, Bahrain, Bangladesh, Cote d'Ivoire, Egypt, Ethiopia, Gambia, Indonesia, Iran, Iraq, Kuwait, Lebanon, Libya, Malaysia, Morocco, Niger, Nigeria, Oman, Pakistan, Qatar, Saudi Arabia, Tanzania, Tunisia, Turkey, United Arab Emirates, Yemen.

List of Tropical Countries:
Angola, Belize, Bolivia, Botswana, Brazil, Brunei, Burundi, Cameroon, Colombia, Cote d'Ivoire, Ecuador, El Salvador, Ethiopia, Gabon, Gambia, Ghana, Grenada, Guatemala, Guyana, Haiti, Honduras, India, Indonesia, Jamaica, Kenya, Kiribati, Liberia, Macau, Madagascar, Malawi, Malaysia, Mali, Mexico, Mozambique, Nicaragua, Niger, Nigeria, Oman, Panama, Papua New Guinea, Peru, Puerto Rico, Singapore, Sri Lanka, Tanzania, Thailand, Trinidad and Tobago, Uganda, Venezuela, Vietnam, Yemen, Zambia, Algeria, Australia, Bangladesh, Chile, China, Egypt, Paraguay, Saudi Arabia, United Arab Emirates.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 85–106, DOI 10.21307/stattrans-2019-016

APPLICATION OF THE STRATEGY COMBINING MONETARY UNIT SAMPLING AND THE HORVITZ–THOMPSON ESTIMATOR OF ERROR AMOUNT IN AUDITING – RESULTS OF A SIMULATION STUDY

Bartłomiej Janusz[1]

ABSTRACT

Auditors need information on the performance of different statistical methods when applied to audit populations. The aim of the study was to examine the reliability and efficiency of a strategy combining systematic Monetary Unit Sampling and confidence intervals for the total error based on the Horvitz-Thompson estimator with normality assumption. This strategy is a possible alternative for testing audit populations with high error rates. Using real and simulated data sets, for the majority of populations, the interval coverage rate was lower than the assumed confidence level. In most cases confidence intervals were too wide to be of practical use to auditors. Confidence intervals tended to become wider as the observed error rate increased. Tests disclosed the distribution of the Horvitz-Thompson estimator was not normal. A detailed analysis of the distributions of the error amount in the examined real audit populations is also given.

Key words: audit sampling, Monetary Unit Sampling, Horvitz-Thompson estimator, error distribution.

1. Introduction

Audit tests are often based on samples. Auditors can use statistical sampling methods in order to estimate or test hypotheses about the error or the correct (audit) value. In practice, when using statistical sampling, auditors usually estimate confidence intervals for the total error amount or the total audit value of the tested account. This practice allows one to control inference precision as well as sampling risk. Moreover, by comparing confidence bounds with tolerable error or category's value auditors can test hypothesis about the error amount or the correct value of a category.

Systematic Monetary Unit Sampling (MUS) scheme combined with confidence intervals based on the Horvitz-Thompson point estimator of the total error and an assumption of the estimator's asymptotic normality is one of the sampling strategies proposed in the audit literature (Statistical Models and Analysis

[1] TAURON Polska Energia S.A., Operational Audit Unit. E-mail: bartlomiej.janusz@tauron.pl. ORCID ID: https://orcid.org/0000-0002-9472-9833.

in Auditing, 1989). We will refer to this strategy as MUS HT strategy. In the literature MUS scheme with the Horvitz-Thompson estimator is often called MUS with the mean-per-unit estimator. Auditors select items with probabilities proportional to their book amount mainly because it is assumed that the risk of big error is higher for line items with a higher book value than for line items with smaller values (Arens and Loebbecke, 1981). Systematic sampling is used because of its simplicity and low sampling costs. Confidence intervals based on the Horvitz-Thompson estimator may be applied especially to audit populations characterized by non-trivial error rate, in the case of which other popular audit sampling strategies based on attribute methods do not yield useful outcomes. However, the performance of intervals based on the Horvitz-Thompson statistic for populations with low error rates, which dominate in audit practice, is questionable (Statistical Models and Analysis in Auditing, 1989). Another problem is the convergence to normality of the distribution of the Horvitz-Thompson estimator when systematic MUS is used.

The aim of the conducted simulation study was to verify reliability and effectiveness of the MUS HT strategy. The simulation was based on real data sets containing annual inventory results. The error rate in analysed sets was higher than for usual audit populations described in the literature. In the case of such populations, the examined sampling strategy should perform better than other popular audit sampling strategies, for example strategies based on attribute sampling. Additionally, we analysed the performance of MUS HT for generated populations with low error rates. Reliability was measured by comparing actual confidence levels to nominal confidence levels. Effectiveness was evaluated by comparing intervals' length to the population total book amount. We conducted our study for samples of size 50 and 100. For these sample sizes we examined normality of point estimator distribution.

The results of the simulation conducted on real audit data may be useful for audit practitioners as well as those who are occupied with applications of statistical methods in auditing. It can contribute to the identification when the strategy can yield positive outcomes and when it should not be used.

We also give an analysis of the distributions of the error amount in the examined populations. The majority of such analysis described in the literature is based on audit samples due to unavailability of accounting data. As our data sets come from a full study the results may also help auditors in better understanding the error amount distributions in accounting populations.

2. Error distribution in audit populations

The efficiency and reliability of the applied sampling strategy can be strongly influenced by characteristics of audited populations, especially by the distribution of errors. Arens and Loebbecke (1981) indicate that typical distribution of an absolute error value is characterized by high rate of error free (error value equal to 0) elements. Due to high rate of error free elements it is often assumed in the literature (for example (Statistical Models and Analysis in Auditing (1989))) that the error amount distribution can be modelled as a nonstandard mixture of the

distributions. The error amount for i^{th} element is treated as random variable of the following type:

$$D_i = \begin{cases} D_i', & \text{with probability } \nu, \\ 0, & \text{with probability } (1-\nu), \end{cases} \tag{1}$$

where:

D_i – the error amount for i^{th} element – random variable,

$D_i = X_i - Y_i$,

X_i – the book amount for i^{th} element – random variable,

Y_i – the audit amount for i^{th} element – random variable,

D'_i – random variable different than zero and representing error value,

$\nu = \sum_{i=1}^{N} \frac{d_{0/1i}}{N}$ – error rate in the population,

$d_{0/1i}$ – dummy variable equal to 1 in the case i^{th} element contains an error, and 0 in the case i^{th} element is error free,

N – population size (line items).

Johnson, Leith, and Neter (1981) studied distribution of errors for accounts receivable and inventories of companies in the United States of America. The authors analysed audit files for 55 companies in the case of accounts receivable and 26 companies in the case of inventories. Their results show high variability of the error rate. Furthermore, the error rate increases with an increase in category value and with an increase in mean value of line items, which is contrary to the general auditors' beliefs that line items with high book amount are precisely verified and thus error probability should be lower. The median of error rate in their study equalled 0.024 for accounts receivable and 0.154 for inventories.

For accounts receivable overstatements (errors for which book value is higher than correct value) dominated significantly. In the case of inventories the number of overstatements and understatements (errors for which book value is lower than correct value) was similar.

Distribution of the error amount in the examined populations differed from normal distribution. High concentration around mean was observed by the authors. Moreover, the distributions were positively skewed and a big number of high value overstatements (exceeding value: mean + 3 x standard deviation) was observed.

A similar study was carried out by Ham, Losell, and Smielauskas (1985) who examined accounts receivable, inventories, accounts payable, sales and purchases. The data used by authors came from the audit files of 5 annual audits for each of 20 companies selected by an audit firm. The median error rate for different categories varied from 0.011 to 0.188 and in the case of inventories it equalled 0.041.

The authors showed that for accounts receivable and sales overstatements prevailed. Understatements dominated for accounts payable and purchases. In the case of inventories the number of overstatements and understatements was

similar. For the majority of cases the error amount distribution was not normal for accounts receivable, accounts payable and inventories.

Allen and Elder (2005) analysed 435 sampling applications collected from inventory and accounts receivable during 1994 and 1999. Authors found that in 49% of sampling applications from year 1994 and 46% from year 1999 auditors detected errors.

Durney, Elder, and Glover (2014) indicate that introduction of Sarbanes-Oxley Act in 2002 caused a decrease in error rates and error magnitudes in accounting data in the United States of America. Authors analysed data set of 160 audit sampling applications from audits conducted by a large auditing firm after SOX implementation. The mean misstatement rate (the sum of absolute values of difference between audit and the book amount for line items tested by auditors divided by the sum of the book amount for line items tested by auditors) across all sampling applications in their study was 0.002. In the case of 0.581 examined sampling applications misstatement rate was 0.

The presented representative results regarding the error distributions indicate potential problems with interval estimation mainly due to non-normality and rare error occurrence. Further on we present the results of our simulation study on interval estimation efficiency and reliability. The examination was based on annual inventory results conducted in the plants of international corporation as well as additional sets generated from real populations, characterized by lower frequency of errors.

3. Description of populations being basis for the simulation study

The basis for the examination were sets containing annual inventory results conducted in the 13 warehouses of 6 manufacturing plants of an international corporation. Our populations are not based on results of sample tests but come from a full study of all inventory items in a particular warehouse. The origin of populations is the reason for special character of errors – they were caused only by incorrect registration of stock quantity. The errors were detected by employees of a plant who conducted stock counting and were corrected before conducting audit procedures by auditors.

Stock taking results were in form of files and consisted of records that for each stock item (for example a specific type of springs or specific type of pipes) – line item – contained the following data: stock item description, warehouse, manufacturing plant, quantity according to inventory results, inventory correction, unit cost. For the purpose of the study we made an assumption that the quantities registered after stock taking were correct. We gave a denomination for each population according to the following convention: “plant_type of warehouse”. Plants were numbered from 1 to 6. The type of warehouse was coded using the following capital letters: M – warehouse of materials, PT – warehouse of work in progress, PG – finished goods, Z – all warehouses in a plant.

Distributions of the book amount are similar for all the examined populations. They are highly skewed right and contain outliers – a small number of stock items of a very big book amount. Observations of zero value occur – these are stock items for which quantity registered before inventory equalled zero but during stock

taking they were identified in a warehouse. Moreover, distributions are strongly concentrated around values smaller than the mean book amount for stock items.

In the case of the examined warehouses the percentage of stock items containing errors varied from 0.428 to even 0.980 and were very high compared to the studies described in the literature. The conducted analysis did not reveal any relationship between the error rate and either plant or warehouse type. The error rate in the studied populations did not depend on warehouse book amount. No relationship between the book amount of the stock item and the error rate was observed.

In Table 1 we present characteristics of the error distribution in the studied populations.

Table 1. Characteristics of error distribution in studied populations.

Population	Number of stock items	Error rate	Mean error (euro)	Total error amount (euro)	Coefficient of variation	Moment coefficient of skewness	total error amount / total book amount
1_M	2 370	0.931	12.54	29 713.44	4.81	-4.43	0.0103
1_PG	462	0.755	234.33	108 258.69	271.32	3.27	0.1274
1_Z	2 934	0.890	6.07	17 795.75	122.21	-7.44	0.0041
2_M	256	0.946	0.10	26.36	12.53	1.58	0.0000
2_PT	360	0.975	1.29	463.11	13.62	14.88	0.0003
2_Z	648	0.935	0.74	479.25	9.60	19.73	0.0001
3_PT	518	0.894	0.97	499.76	13.94	20.93	0.0003
4_PT	747	0.980	1 606.75	1 200 244.55	7.57	4.63	0.1228
5_M	501	0.603	-531.83	-266 448.65	-213.06	-6.22	-0.0240
5_PT	410	0.751	181.00	74 209.13	-15.76	2.87	0.0286
5_Z	2 615	0.428	-16.49	-43 129.91	15.70	-13.34	-0.0029
6_PT	430	0.837	16.19	6 962.08	58.21	-2.35	0.0050
6_Z	925	0.533	34.26	31 686.49	169.66	-2.54	0.0035

In the case of 11 populations the total error was positive and only in 2 cases it was negative. The number of overstatements exceeded the number of understatements for all the analysed sets. The rate of overstated line items among all items in the error ranged from 0.571 to 0.722. High coefficient of variation together with big differences between the minimum and maximum value of errors indicate high variability of the error amount.

In accordance with audit methodology, auditors are interested in material errors, i.e. errors that can influence the economic decisions taken on the basis of the financial statements. The materiality level, in a simplified way, can be established as a percentage, ranging usually from 0.5% to 2%, of the category's book amount. For all populations we calculated a ratio of the total error amount divided by the total book amount. The absolute value of this ratio in 6 cases exceeded 0.005 and in 4 cases exceeded 0.02. It means that only in the case of 4 populations the error would be material if auditors established the materiality level in a way described above and 2% multiplying factor was used. If auditors used 0.5% multiplying factor 6 populations would be assumed significantly in an error. Only for 2 out of 13 sets the absolute value of the total error divided by the total

book amount ratio was higher than 5%. In the case of 4 populations the total error was lower than 500 euro while the book amount exceeded 1.5 million euro.

The conducted analysis did not reveal any significant relationship between the error amount and the book amount of the stock item.

Figures 1 and 2 present typical distributions of the error amount for the examined populations. For better presentation the outliers – stock items with a very big or a very small error amount were not included in the figures.

The distribution of the error amount was strongly concentrated around zero. The number of errors decreased with increasing absolute error amount. This property is typical for all the examined populations. In the case of sets for which the error rates were lower (for example warehouse 5_Z) the observed distribution of the error amount can be described as nonstandard mixture of the distributions given by Eq. (1). An example of such a distribution may be seen in Figure 1. Concentration of the error amount around zero caused low values of mean and median of the error amount. At the same time outliers – stock items with a very big absolute error amount – caused high level of variability of error value measured by standard deviation. Due to low mean values, strong concentration around zero and high values of standard deviations, 87% to 99% of observations laid in the interval: mean +/- standard deviation. In the case of 7 populations the error distribution was positively skewed and in the case of 6 populations it was negatively skewed. The absolute value of moment coefficient of skewness was high for all populations.

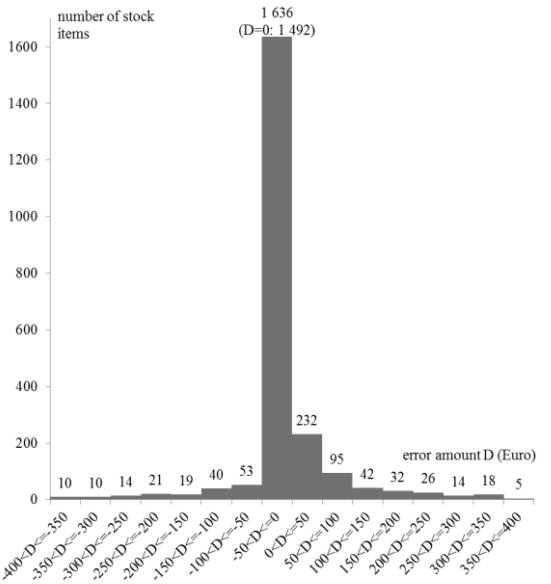

Figure 1. Exemplary distribution of error amount (D) – population 5_Z (outliers* excluded).

* Outliers include 154 items containing errors ranging from - 110 980.64 euro to - 400.14 euro and 194 items containing errors ranging from 406.40 euro to 34 957.03 euro.

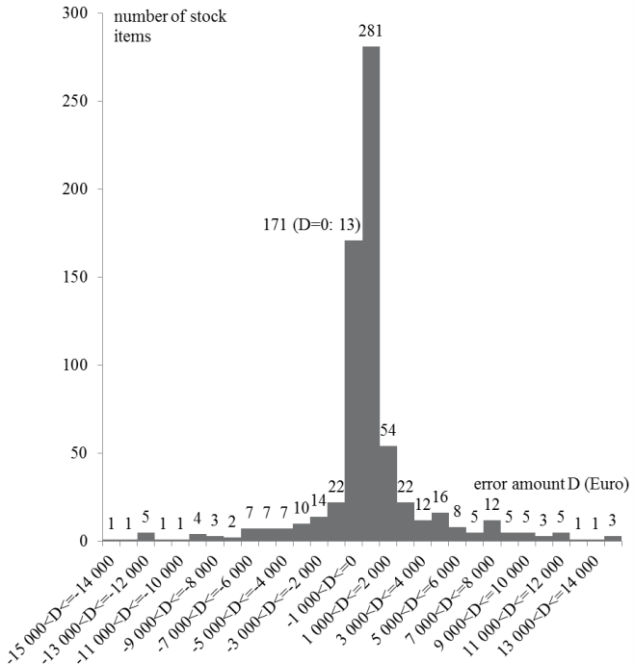

Figure 2. Exemplary distribution of error amount (D) – population 4_PT (outliers* excluded).

* Outliers include 20 items containing errors ranging from - 43 121.09 euro to - 15 321.31 euro and 38 items containing errors ranging from 15 119.99 euro to 127 057.86 euro.

The error rates for all the examined populations were much higher than the error rates identified in other studies on the distribution of errors in accounting populations. In order to test the performance of the sampling strategy for populations with low error rates we generated additional data sets based on selected original populations.

In order to choose populations for creating new data sets we determined the selection criteria taking into account to what extent these criteria allocate the original sets into homogenous, representative groups and what impact they may have on estimation results. We determined five selection criteria:

- variability of the book amount measured with coefficient of variation,
- negative or positive sign of the total error value,
- variability of the error amount measured with coefficient of variation,
- materiality of the total error,
- distribution characteristics of taints (stock item's error amount divided by its book amount).

Using these criteria we chose four populations as the basis for generating additional sets: 1_PG, 3_PT, 5_PT, 5_Z.

On the Basing on results described in the literature we selected three different target error rates: 0.02; 0.07; and 0.15. For each chosen original population we created three sets with different target rates.

For generating purposes we assumed that during stock taking all errors were identified and thus after making the adjustments all data is correct. We generated additional populations in such a way as if stock counting was not fully effective, that means, for randomly selected line items in error no adjustments resulting from stock count were made. Stock items for which errors were not corrected were chosen in such a way that for each erroneous item a number from 0 to 1 was randomly drawn with equal probability. If the number was smaller or equal to the value of quotient of target and the original (value in real set) error rate, the adjustment was not made and an error equal to the original error amount was assigned to line item. The original book value and the audit value were assigned to line item. If the randomly drawn number was higher than the value of quotient of target and the original (value in real set) error rate, the error was corrected and the line item's book value in generated set equalled audit value in the original population.

In such a way we generated 12 additional populations. For each population we used the following notation convention: base population and numbers 2, 7, or 15 depending on the target error rate. For example 1_PG_2 stands for the population generated from set 1_PG with the assumed target error rate equal to 0.02.

In the case of additional populations generated based on set 5_Z, for which the total error amount was negative, total book value was higher than in the case of the original population. For all other generated sets the total book amount decreased comparing to the original populations. Variability of the book amount in generated populations was similar to variability in base sets. For all populations the distribution of the book amount was highly skewed right. No relationship between the book amount of the stock item and the error rate was observed for new populations.

Due to random selection of stock items remaining in the error, the real error rates in the generated populations differed from the target rates. We present the error rates and characteristics of the error amount in Table 2.

Table 2. Characteristics of error in generated populations.

Population	Number of stock items	Error rate	Mean error (euro)	Total error amount (euro)	Coefficient of variation	Moment coefficient of skewness	total error amount / total book amount
1_PG_2	462	0.024	6.18	2 853.26	28.32	18.90	0.0038
1_PG_7	462	0.065	15.32	7 077.14	38.43	14.00	0.0095
1_PG_15	462	0.147	54.63	25 237.29	32.07	11.64	0.0329
3_PT_2	518	0.017	0.00	0.82	31.32	8.32	0.0000
3_PT_7	518	0.077	0.02	11.04	12.93	6.35	0.0000
3_PT_15	518	0.160	0.57	295.32	21.99	22.41	0.0002
5_PT_2	410	0.012	1.56	639.23	-714.10	18.59	0.0003
5_PT_7	410	0.071	59.77	24 504.89	11.96	8.43	0.0096
5_PT_15	410	0.180	169.71	69 581.19	8.26	15.31	0.0269
5_Z_2	2 615	0.022	14.34	37 486.48	23.08	27.92	0.0025
5_Z_7	2 615	0.065	-1.46	-3 830.84	345.56	-20.77	-0.0003
5_Z_15	2 615	0.137	3.81	9 972.03	11.24	-10.52	0.0007

The absolute total error for the majority of new populations decreased significantly in comparison with the base sets. For populations 5_Z_2 and 5_Z_15 the total error was positive in contrary to the original set 5_Z for which it was negative. The only population with the negative total error was 5_Z_7. Despite a significant decrease in the total errors, in the case of 4 generated sets the absolute value of the ratio: the total error divided by the total book amount was higher than 0.005 and for 2 of them it was higher than 0.02. Lower total errors caused lower mean errors and higher coefficient of variation. No significant relationship between the error amount and the book amount of the stock item was observed.

Due to applied generation method, the distribution of the error amount in additional populations can be described as nonstandard mixture of the distributions given by Eq (1). In Figure 3 we present a typical distribution of the error amount for the generated populations. For better presentation outliers were excluded.

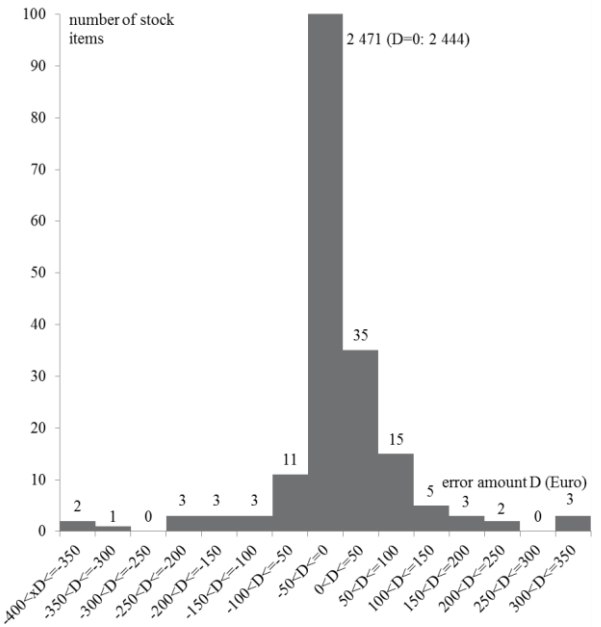

Figure 3. Exemplary distribution of error amount (D) – population 5_ Z_7 (outliers* excluded).

* Outliers include 22 items containing errors ranging from -41 257.73 euro to -485.64 euro and 36 items containing errors ranging from 406.40 euro to 15 031.86 euro.

In addition to a very high number of error free stock items, relatively numerous elements with low error amount occurred. The majority of these errors were overstatements. In the case of 10 populations the number of overstatements was higher than the number of understatements. For the majority of the generated

sets the distribution of the error amount was highly positively skewed. For two populations the distribution of the error amount was highly skewed left.

4. Description of the simulation study

The subject of the simulation study was the efficiency and reliability of MUS HT strategy. It is believed that this strategy may work well when the error rate in the audited population is high as it is in the case of our data sets. The actual confidence levels compared to the nominal confidence levels as well as the average length of confidence intervals compared to population the total book amount were used as the main evaluation criteria. Auditors use the total book amount to determine the materiality levels and thus it is useful to compare interval's length with the total book amount for judging estimation efficiency. We calculated the actual confidence levels as a ratio of the number of intervals that contained the total error divided by the number of all estimated intervals.

We conducted the study for samples of size 50 and 100. For each sample size 1,000 samples, from each population (original and additionally generated), were drawn. For each sample confidence interval was calculated.

When conducting substantive testing, that is testing the accuracy of the registration of transactions in book of accounts and correctness of book balances of accounts, auditors usually know book values of line items that make up the account, population of transactions. It allows one to apply such random sampling designs that elements are selected with probabilities proportional to their book amount.

This way of selecting a sample is very useful to auditors because line items with bigger values have bigger probabilities of being selected. Auditors want to test such elements for two reasons. First of all, one of the evaluation criteria of audit works is the ratio of the value of the tested elements to the value of all elements. The higher the value of this ratio, the more complete and accurate the audit is concerned to be.

Furthermore, it is assumed, that the risk of a big error is higher for line items with the higher value than for line items with the small value (Arens and Loebbecke (1981)). Even if this relationship does not hold it was showed (for example Jonhson, Leitch, and Neter (1981) and Neter, Jonhson, and Leitch (1985)) that in the case of audit populations the variability of the error value measured by standard deviation increases for line items with bigger book values. It seems to justify the concentration of works on elements with big book values.

A review of different sampling designs in the case of which first-order inclusion probabilities are proportional or approximately proportional to the book amount can be found in (Tillé (2006), Wywiał (2016)). Because of simplicity and low selection costs auditors often use a systematic sampling scheme, without replacement, proposed by Madow (Sarndal, Swensson, and Wretman (1992)), (Arens and Loebbecke (1981)). We will refer to this sampling scheme described below as systematic MUS (Monetary Unit Sampling). Let population elements be listed in a random order,

$T_{x0}=0,$

$T_{xi}=T_{x(i-1)}+x_i,$

where:

x_i – the book amount of the i^{th} line item, we assume that $x_i>0$ for each $i=1,...,N$;

T_{xi} – the sum of the book amount of line items numbered from 1 to i;

N – the number of elements in a population (line items).

Let the sample size be equal n line items, then sampling interval a equals: $a=T_x/n$, where $T_x = T_{xN}$ – population the total book amount. Number b is drawn with equal probability from the interval (0,a). Sample s consists of the following line items:

$$s = \{i: T_{x(i-1)} < b + (k-1)a <= T_{xi} \text{ for } k = 1,...,n\}. \quad (2)$$

If for each $i=1,...,N$ $x_i< T_x/n=a$, then we obtain a sample scheme without replacement. Because N is usually significantly bigger than n, and elements with big book values are usually tested separately with probability equal to 1, the condition $x_i< T_x/n$ is assumed to be easily fulfilled in the case of audit populations.

For systematic MUS the sampling interval does not have to be an integer. That is why the number of elements in a sample is fixed and equals n. First-order inclusion probability for the i^{th} element equals (Wolter (1985)):

$$\pi_i = \frac{nx_i}{T_x}. \quad (3)$$

One can see that the second-order inclusion probabilities depend not only on the book value of elements but also on their order in the population. One of the ways to solve this problem is to replace exact values of second-order inclusion probabilities with their approximations. Wolter (1985) proposes, after Hartley and Rao (1962), the following approximation of second-order inclusion probabilities:

$$\boldsymbol{\pi}'_{ij} = \frac{n-1}{n}\boldsymbol{\pi}_i\boldsymbol{\pi}_j + \frac{n-1}{n^2}(\boldsymbol{\pi}_i^2\boldsymbol{\pi}_j + \boldsymbol{\pi}_i\boldsymbol{\pi}_j^2) - \frac{n-1}{n^3}\boldsymbol{\pi}_i\boldsymbol{\pi}_j\sum_{k=1}^{N}\boldsymbol{\pi}_k^2. \quad (4)$$

The approximation is correct to order $O(N^{-3})$, if the following two conditions hold:

(i) elements are listed in random order; (ii) π_i is order $O(N^{-1})$, (Wolter (1985)).

The (ii) condition holds as

$$\pi_i = \frac{nx_i}{T_x} \le \frac{nx_i}{Nx_{\min}} \le \frac{1}{N}\frac{nx_{\max}}{x_{\min}} \quad \text{and} \quad \frac{nx_{\max}}{x_{\min}} = const \quad (5)$$

where:

$x_{min} = min(x_i, \text{ for } k = 1,...,n)$ - minimum book amount over all population elements,

$x_{max} = max(x_i, \text{ for } k = 1,...,n)$ - maximum book amount over all population elements.

In audit practice, unless contrary evidence exists, the natural order of the population is accepted as one possible random ordering. We assumed that the auditor would not randomize the analysed populations before sampling as there is no evidence that a relationship between the population original order and the

error amount exists. Thus, the sets were not randomized before the sample selection. Such an approach may significantly reduce the sampling space. Sampling distribution for systematic selection without randomization differs from distributions used for evaluation purposes, which may lead to differences between actual and the assumed confidence level (Hoogduin, Hall, Tsay, Pierce (2015)).

We used the following confidence interval based on an assumption of asymptotic normality of point estimator proposed by Horvitz and Thompson (1952):

$$\left(t_{d\pi} - u_{\alpha/2} v^{1/2}\left(t_{d\pi}\right); t_{d\pi} + u_{\alpha/2} v^{1/2}\left(t_{d\pi}\right)\right) \tag{6}$$

where:

$t_{d\pi} = \sum_{i \in s} \frac{d_i}{\pi_i}$ – the Horvitz-Thompson estimator of the total error amount (T_d), (7)

d_i – the error amount for i^{th} element, $d_i = y_i - x_i$; y_i - audit amount for i^{th} element,

$v\left(t_{d\pi}\right) = \frac{1}{n-1} \sum_{i}^{n} \sum_{i<j}^{n} \left(1 - \pi_i - \pi_j + \sum_{k=1}^{N} \frac{\pi_k^2}{n}\right)\left(\frac{d_i}{\pi_i} - \frac{d_j}{\pi_j}\right)^2$ – estimator of variance of $t_{d\pi}$, (8)

$u_{\alpha/2}$ – number for which: $\Phi(u_{\alpha/2}) = 1 - \alpha/2$,

Φ – cumulative distribution function of standardized normal distribution N(0,1).

The estimator given by Eq. (8) was obtained by Wolter (1985) from the well-known estimator of variance of the Horvitz-Thompson total value estimator due to Yates and Grundy (1953):

$$v_{wk}\left(t_{d\pi}\right) = -\frac{1}{2} \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{d_i}{\pi_i} - \frac{d_j}{\pi_j}\right)^2, \tag{9}$$

by replacing π_{ij} with their approximations given by Eq. (4). Tillé (2006) proposes several variance estimators requiring knowledge of only first inclusion probabilities. These variance estimators are constructed on the basis of variance approximations.

The study conducted by Christensen, Elder, and Glover (2015) revealed that the confidence level required in the case of substantive tests ranges from 30% to 96%. For all firms the high end of confidence range was consistently at or near 95%. In order to check how the examined strategy performs in the most demanding situation, we applied 95% nominal confidence level for all confidence intervals computed in the simulation study.

In the case of systematic MUS the scheme elements with 0 book amount have zero inclusion probability and are excluded from sampling population. Furthermore, as mentioned above, elements with “very big book amount”, which means elements for which $x_i \geq T_x / n$ are rejected from the sampled population

and usually form a stratum which is being subject to full testing. Rejection of elements with "very big book amount" causes a decrease in the total book amount T_x of sampled population and thus further elements may not fulfil the "new" condition $x_i < T_x / n$ and must be rejected from sampled population. It did not allow to apply systematic MUS in the case of the following 14 populations:

- sample size 50 – 2_PT
- sample size 100 – 1_PG, 1_PG_2, 1_PG_7, 1_PG_15, 2M, 2_PT, 2_Z, 5_M, 5_PT, 5_PT_2, 5_PT_7, 5_PT_15, 6_PT.

The subject of estimation in the simulation study was the total error amount of the "modified" populations obtained after excluding elements with 0 book amount and elements with "very big book amount".

When selecting samples from populations with low error rates it is possible that for some samples no error occurs. For such samples the length of interval given by Eq (6) equals zero. It is perceived by auditors as a serious drawback of the analysed strategy. For all populations and sample sizes the rate of intervals with zero length was calculated. We took into consideration such intervals while computing coverage percentage and excluded them for calculation of mean length of confidence interval.

5. Simulation study results

In Table 3 we present the true level of confidence, mean length of confidence interval, mean distance between point estimator and the total error value and the rate of zero length intervals (corresponding to the rate of error free samples) for the sample size 50.

For 10 out of 24 populations coverage percentage was higher than the nominal value (95%). For 1 population (1_PG_7) coverage percentage equalled 1. The lowest coverage ratio amounted to 0.035 in the case of population 3_PT_15.

Only for 6 populations the mean length of confidence interval was lower than 0.005 of the book amount. In the case of 6 sets the mean length of interval was higher than the population book amount. Such wide confidence intervals are in practice useless for auditors.

The value of the mean distance between the point estimator and the total error value divided by the population book amount only in the case of 8 sets was lower than 0.02, among these, in the case of 6 sets it was lower than 0.005. For 7 populations the value of this quotient exceeded 0.1. In the case of all the examined sets the value of the mean distance between the point estimator and the total error value was lower than the mean length of the confidence interval.

In the case of 8 populations zero length intervals occurred, which means that samples with no errors were present. All cases related to the generated populations with lower error rates. The highest percentage of such intervals – 0.769 – occurred for set 1_PG_2. No zero length intervals occurred for populations with the non-trivial error rate.

Table 3. True level of confidence, mean length of confidence interval, mean distance between point estimator and total error value, rate of zero length intervals – sample size 50.

Population	Coverage percentage – true level of confidence	mean length of confidence interval / population book amount	mean distance (total error value – point estimator value) / population book amount	Rate of intervals length = 0
1_M	0.920	0.316	0.080	0.000
1_PG	0.438	2.011	0.374	0.000
1_PG_2	0.231	0.122	0.045	0.769
1_PG_7	1.000	0.824	0.156	0.000
1_PG_15	0.999	1.146	0.183	0.000
1_Z	0.927	0.313	0.068	0.000
2_M	0.998	0.002	0.000	0.000
2_Z	0.728	0.001	0.000	0.000
3_PT	0.355	0.001	0.000	0.000
3_PT_2	0.597	0.000	0.000	0.370
3_PT_7	0.621	0.000	0.000	0.077
3_PT_15	0.035	0.001	0.000	0.002
4_PT	0.884	0.308	0.060	0.000
5_M	0.997	0.727	0.111	0.000
5_PT	0.589	3.830	1.117	0.000
5_PT_2	0.892	0.209	0.031	0.108
5_PT_7	0.998	1.258	0.197	0.001
5_PT_15	0.999	1.727	0.212	0.000
5_Z	0.987	0.322	0.073	0.000
5_Z_2	0.752	0.040	0.008	0.203
5_Z_7	0.919	0.086	0.019	0.002
5_Z_15	0.991	0.148	0.033	0.000
6_PT	0.995	3.942	0.555	0.000
6_Z	0.989	0.195	0.033	0.000

Taking into account both evaluation criteria: the actual confidence level and the mean length of interval, the only population for which the analysed strategy performed well was set 2_M. It should be also noted that in the case of this population there were no samples free of error. For all other data sets the performance of MUS HT strategy was unacceptable because of either too low coverage ratio or too long intervals.

As discussed above, the reasons for applying sampling schemes for which selection probability is proportional to the book amount are among others alleged growth of risk of a big error as well as an increase in variability of the value of errors with an increase in the book amount of line items.

For the examined populations no relationship between the error amount and the book amount of the stock item was observed. We carried out an additional analysis in order to verify if the variability of the value of errors increases for

elements with bigger book values. For this purpose, we ordered stock items in each examined population with growing book amount. We divided population size (N) by 10 and rounded the obtained result to the nearest integer ($N/10_{rounded}$). Next, we divided the set into 10 strata in such a way that to the first stratum $N/10_{rounded}$ stock items with highest book amount were assigned, to the second stratum next $N/10_{rounded}$ stock items with highest book amount were allocated and so on until the ninth stratum. The tenth stratum contained elements with lowest book amount that were not assigned to previous nine strata. For each stratum we calculated standard deviation of the error amount.

For all sets for which the coverage percentage was greater than or equal to the nominal confidence level, the variability of the error amount generally increases with an increase in the book amount. However, in some of these populations this trend is not so obvious. Furthermore, in the case of, 4 out of 5 populations for which the true confidence level was lower than 0.5 such relationship did not occur. For 2 data sets for which true level of confidence was lower than the assumed level of confidence - populations 5_PT and 5_Z_2 - the variability of the error amount generally increases with an increase in the book amount.

5.1. Increase in sample size effect

In Table 4 we present the true level of confidence, the mean length of confidence interval, the mean distance between point estimator and the total error value and the rate of zero length intervals for samples size 100.

Table 4. True level of confidence, mean length of confidence interval, mean distance between point estimator and total error value, rate of zero length intervals – sample size 100.

Population	Coverage percentage – true level of confidence	mean length of confidence interval / population book amount	mean distance (total error value – point estimator value) / population book amount	Rate of intervals length = 0
1_M	0.922	0.275	0.059	0.000
1_Z	0.929	0.276	0.055	0.000
3_PT	0.936	0.017	0.004	0.000
3_PT_2	0.926	0.000	0.000	0.035
3_PT_7	1.000	0.001	0.000	0.000
3_PT_15	0.180	0.015	0.003	0.000
4_PT	0.914	0.976	0.625	0.000
5_Z	0.997	0.587	0.121	0.000
5_Z_2	0.924	0.104	0.015	0.045
5_Z_7	0.999	0.282	0.048	0.000
5_Z_15	1.000	0.318	0.046	0.000
6_Z	0.991	0.676	0.096	0.000

An increase in sample size from 50 to 100 caused a higher coverage percentage for all the studied populations. Only for one set the true confidence level was lower than 0.9 and amounted to 0.180 (3_PT_15). For this population no increase in the error amount variability with an increase in the book amount was observed. In the case of 5 sets coverage ratio was greater than or equal to nominal confidence level.

An increase in sample size caused a decrease in estimator variance but still "very long" intervals occurred. For 3 populations mean length of confidence interval was higher than 0.5 of the population book amount. In the case of 4 out of 12 sets the mean length of interval was lower than 0.02 of population the total book amount, among these, in the case of 2 sets it was lower than 0.005 of population total book value.

For 7 out of 12 populations an increase in the sample size caused an increase in the ratio: the mean distance between point estimator and the total error divided by the population book amount. In the case of 5 sets this quotient was lower than 0.02, among these, in the case of 4 populations it was lower than 0.005. For all the examined sets from which 100 item samples were drawn, the value of mean distance between the point estimator and the total error value was lower than the mean length of the confidence interval.

In the case of 2 populations: 3_PT_2 and 5_Z_2, zero length intervals occurred.

Taking into account both evaluation criteria: actual confidence level and mean length of interval, the only population for which the analysed strategy performed well was set 3_PT_7. It should be also noted that in the case of this population there were no error free samples.

5.2. Error rate effect

We did not observe a relationship between the error rate and the true confidence level. For groups of populations 3_PT_50 and 1_PG_50 a sharp decrease in the coverage percentage with an increase in the nominal error rate from 0.07 to 0.15 can be observed. Furthermore, in the case of group of populations 5_PT_50 the true confidence level for the original set is the lowest, for group of populations 1_PG_50 it is significantly lower than for sets with the nominal error rate 0.07 and 0.15. Finally, in the case of group of populations 3_PT_50 the coverage percentage for the original set is much lower than for sets with the nominal error rate 0.02 and 0.07. Figure 4 presents changes in coverage percentage with change of error rates for these groups of populations.

For group 3_PT_100, not presented in Figure 4, changes in actual confidence level with changes in the error rate had the same pattern as in the case of group 3_PT_50. An upward trend of the true confidence level with an increase in the error rate occurs only in the case of 2 groups of populations: 5_Z_50 and 5_Z_100 – groups not presented in Figure 4.

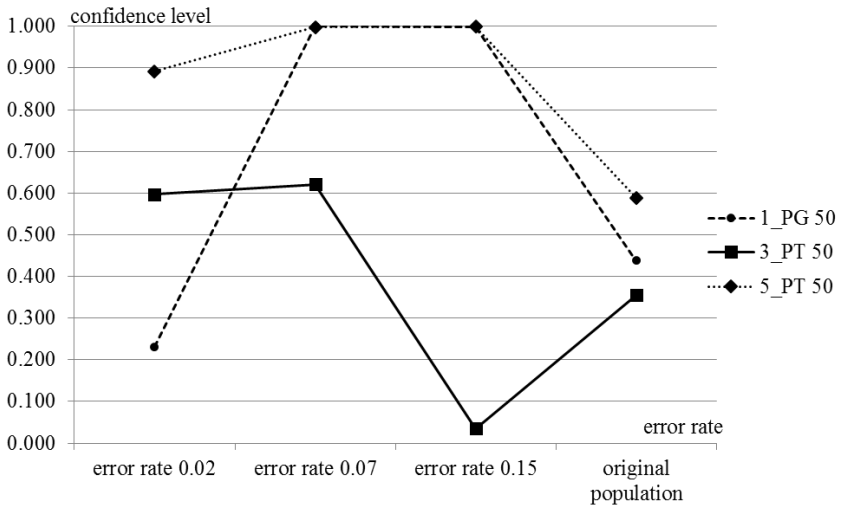

Figure 4. Relationship between error rate and true confidence level.

In contrast, it can be observed that together with the growth of the error rate the mean length of confidence interval increases. This relationship is presented in Figure 5.

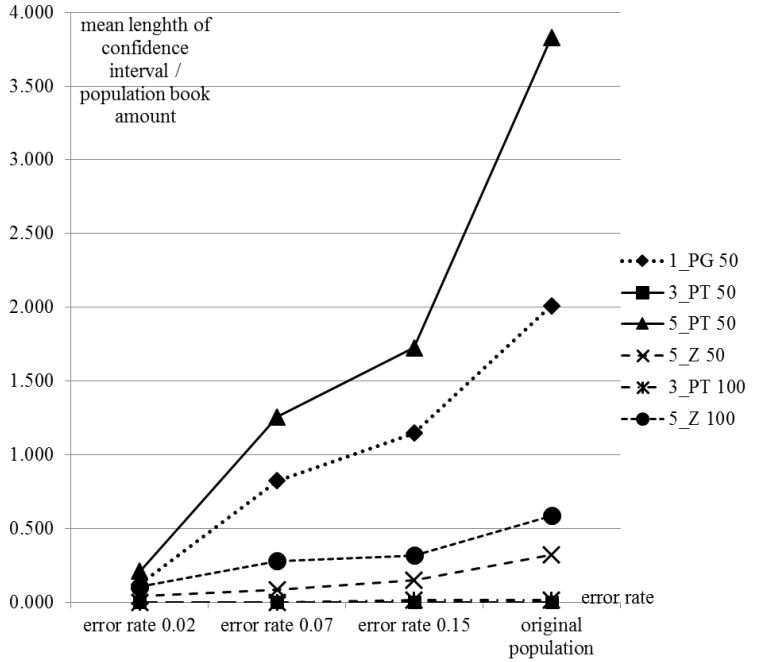

Figure 5. Relationship between error rate and length of confidence interval.

5.3. Normality of the distribution of point estimator

Examined confidence intervals were based on an assumption of asymptotic normality of the Horvitz-Thompson point estimator. Hájek (1964) derived conditions for asymptotic normality of the Horvitz-Thompson estimator in the case of rejective sampling under assumptions that $n \to \infty$ and $N - n \to \infty$. The author proposed two estimators of the Horvitz-Thompson estimator variance. Berger (1998) showed that if divergence between a given sampling design and rejective sampling design goes to zero then the Horvitz-Thompson estimator has an asymptotic normal distribution. The author gave also the rate of convergence of the Horvitz-Thompson estimator for any kind of sampling.

On the basis of the simulation results we tested if the assumption of normality of the Horvitz-Thompson estimator holds for statistic

$$st_{d\pi} = \frac{t_{d\pi} - T_d}{v^{1/2}\left(t_{d\pi}\right)}. \tag{10}$$

We did not conduct evaluation of the necessary sample size to ensure the sufficient convergence of the estimator to the normality. Such evaluation was discussed by Wywiał (2016). We verified normality of statistic given by Eq (10) with the Shapiro – Wilk W-test. It should be taken into account that this test has very high power.

In order to calculate the test statistic W we used approximation for the required coefficients proposed by Royston (1992). According to the author the approximation is accurate to ±1 in the fourth decimal place. Furthermore, we applied Royston's (1992) normalizing transformation for the W statistic.

Only in the case of population 5_Z_100 p – the value was higher than 0.05. For the remaining populations p – the values were very low: only in 2 cases they differed substantially from zero but were much lower than 0.05. One reason for this may be the observed high skewness of the analysed populations (Statistical Models and Analysis in Auditing (1989)). The conducted analysis showed that for different sets $st_{d\pi}$ statistic had very different distributions that cannot be attributed to one or specified group of the distribution types. One way of solving the problem of non-normality may be applying the bootstrap procedure.

6. Results of other studies

Our results are similar to other studies' results. Sampling strategy using systematic MUS scheme and the confidence interval based on the Horvitz-Thompson point estimator was subject of the simulation conducted by Neter and Loebbecke (description and results of the study are given following (Statistical Models and Analysis in Auditing (1989))). Four audit populations, for which the error rate was 0.3 were used in this study. Additionally, based on the original sets, 16 populations with lowered error rates from 0.3 to: 0.005, 0.01, 0.05, and 0.1 were generated. The examination was conducted for 14 of these sets. Six hundred samples of size 100 were drawn from each examined population. The true confidence level was calculated as percentage of intervals that covered the total error.

For none of the populations the true confidence level reached nominal value of 95.4%. The lowest coverage percentage equalled 5.2% for one of sets with the error rate 0.005. For two populations with error rates: 0.1 and 0.3 the true confidence level reached the highest value equal to 94.5.

These results are consistent with results obtained by Dworin and Grimlund (1984), who compared the reliability of the proposed new method of interval estimation called moment bound with the mean-per-unit estimator combined with MUS scheme. The performance of one-sided confidence interval (upper bound) for 128 inventory populations was analysed in the study. Their results show that only in the case of 13 populations the coverage rate for the mean-per-unit method reached or exceeded 95% nominal confidence level. The lowest coverage rate equalled 72.6%.

Kim Neter and Godfrey (1987) analysed reliability and efficiency of upper bound based on the mean-per-unit estimator and MUS Cell Sampling - a two-stage sampling scheme. According to the authors the MUS Cell Sampling scheme can be treated as an easy to use alternative to systematic sampling of monetary units. The reliability was measured by coverage ratio while efficiency was measured by relative tightness - mean bound in the replications expressed relative to the total error amount in a sampled population. The average coverage over all 64 study populations for bound based on the mean-per-unit estimator and MUS Cell Sampling equalled 76.7% while the minimum coverage was 23% and the maximum 94.2%. Average relative tightness equalled 1.75. The minimum and maximum value of this measure was 1.29 and 2.71 respectively.

Marazzi and Tillé (2017) conducted a simulation study in which MUS with the Horvitz-Thompson point estimator was compared with other sampling strategies. Authors did not analysed the confidence intervals but only the estimators' mean standard error. Their results show a relatively high empirical mean standard error in the case of MUS with the Horvitz-Thompson point estimator strategy.

7. Conclusion

The purpose of the simulation study was to examine the efficiency and reliability of interval estimation for the MUS HT sampling strategy. The main evaluation criteria included actual confidence levels compared to nominal confidence levels as well as the average length of confidence intervals compared to the population total book amount. The basis for the examination were sets containing annual inventory results as well as additional, generated populations with lower error rates.

For the majority of populations the percentage of intervals that covered the total error amount was lower than the nominal confidence level. The obtained results show that for all populations for which the coverage percentage was greater than or equal to the nominal confidence level, the variability of the error amount increases with an increase in the elements' book value. The sample size growth had a positive effect on the coverage percentage. No relationship between the error rate and the actual confidence level was found. The observed non-normality of the Horvitz-Thomson point estimator standardized by its estimator of standard deviation, for applied sample sizes, may be one of the reasons for lower than the assumed true confidence levels.

For most cases, the length of the btained confidence intervals make them useless for auditors. In the case of some populations, the mean length of interval was higher than the population book amount. An increase in sample size caused a decrease in estimator variance but still “very long” intervals occurred. We observed that together with growth of the error rate the mean length of confidence interval increased.

Taking into consideration both evaluation criteria: the actual confidence level and the mean length of interval, the MUS HT strategy performed well only in two cases. Taking into account that for the majority of the used populations the error rates were very high, our results are in contrary to the belief that the analysed sampling strategy may be useful for populations with high error rates. The fact that the length of intervals increases with growth of the error rate seems to strengthen this conclusion.

The obtained results are consistent with results of other simulation studies on MUS HT strategy.

It must be, however, stressed that the applied approach, consistent with the typical auditors’ way of using systematic MUS, assuming lack of randomization of populations before sample selection might significantly reduce the sampling space and thus might have a substantial impact on the obtained results.

One disadvantage of the systematic MUS revealed by the study is inability to apply this scheme to populations composed of elements with “very big book amount”. Rejection of elements with “very big book amount” causes a decrease in the total book amount of the sampled population and thus further elements must be rejected because their book value is higher than the “new” sampling interval. In the case of 14 populations it did not allow for the application of the systematic MUS scheme.

The analysis of real accounting data sets showed that the distribution of the book amount is strongly concentrated around values smaller than the mean book amount, highly skewed right and contains outliers.

We did not observe a significant relationship between either the book amount of the stock item and the error rate or the book amount and the error amount of the stock item. The distribution of the error amount was strongly concentrated around zero. With increasing absolute error amount the number of errors decreased. The outliers caused a high level of variability of the error value measured by standard deviation. The number of overstatements exceeded the number of understatements for all the analysed sets. The absolute value of moment coefficient of skewness was high for all populations.

REFERENCES

ALLEN, R. D., ELDER, R. J., (2005). A Longitudinal Investigation of Auditor Error Projection Decisions, Auditing: A Journal of Practice & Theory, 24 (2), pp. 69–84.

ARENS, A. A., LOEBBECKE, J. K., (1981). Applications of Statistical Sampling to Auditing, Englewood Cliffs: Prentice – Hall, Inc.

BERGER, Y. G., (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. Journal of Statistical Planning and Inference, 67 (2), pp. 209–226.

CHRISTENSEN, ELDER, GLOVER, (2015). Behind the Numbers: Insights into Large Audit Firm Sampling Policies, Accounting Horizons, 29 (1), pp. 61–81.

DURNEY, M., ELDER, R. J., GLOVER, S. M., (2014). Field Data on Accounting Error Rates and Audit Sampling, Auditing: A Journal of Practice & Theory, 33 (2), pp. 79–110.

DWORIN, L., GRIMLUND, R. A., (1984). Dollar Unit Sampling for Accounts Receivable and Inventory, The Accounting Review, LIX(2), pp. 218–241.

HAM, J., LOSELL, D., SMIELIAUSKAS, W., (1985). An Empirical Study of Error Characteristics in Accounting Populations, The Accounting Review, LX (3), pp. 387–406.

HARTLEY, H. O., RAO, J. N. K., (1962). Sampling with Unequal Probabilities and without Replacement, The Annals of Mathematical Statistics, 33 (2), pp. 350–374.

HÁJEK, J., (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population, The Annals of Mathematical Statistics, 35 (4), pp. 1491–1523.

HOOGDUIN, L. A., HALL, T. W., TSAY, J. J., PIERCE, B. J., (2015). Does Systematic Selection Lead to Unreliable Risk Assessments in Monetary – Unit Sampling Applications? Auditing: A Journal of Practice & Theory, 34 (4), pp. 85–107.

HORVITZ, D. G., THOMPSON, D. J., (1952). A Generalization of Sampling Without Replacement From a Finite Universe, Journal of the American Statistical Association, 47 (260), pp. 663–685.

JOHNSON, J. R., LEITCH, R. A., NETER J., (1981). Characteristics of Errors in Accounts Receivable and Inventory Audits, The Accounting Review, LVI (2), pp. 270–293.

KIM, H. S., NETER, J., GODFREY, J. T., (1987). Behavior of Statistical Estimators in Multilocation Audit Sampling, Auditing: A Journal of Practice & Theory, 6 (2), pp. 40–58.

MARAZZI, A., TILLÉ, Y., (2017). Using past experience to optimize audit sampling design, Review of Quantitative Finance and Accounting, 49 (2), pp. 435–462.

NETER, J.,. JOHNSON, J. R, LEITCH, R. A., (1985). Characteristics of Dollar – Unit Taints and Error Rates in Accounts Receivable and Inventory, The Accounting Review, LX (3), pp. 488–499.

ROYSTON, P., (1992). Approximating the Shapiro – Wilk W-test for non-normality, Statistics and Computing, 2, pp. 117–119.

SARNDAL, C. E., SWENSSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling, Springer – Verlag New York, Inc.

STATISTICAL MODELS AND ANALYSIS IN AUDITING, (1989). Panel on Nonstandard Mixtures of Distributions, Statistical Science, 4 (1), pp. 2–33.

TILLÉ, Y., (2006). Sampling Algorithms, Springer Science+Business Media, Inc.

WOLTER, K. M., (1985). Introduction to Variance Estimation, Springer – Verlag New York, Inc.

WYWIAŁ, J. L., (2016). Contributions to Testing Statistical Hypotheses in Auditing, Warszawa: Wydawnictwo Naukowe PWN SA.

YATES, F., GRUNDY, P. M., (1953). Selection without Replacement from Within Strata with Probability Proportional to Size, Journal of the Royal Statistical Society. Series B (Methodological), 15 (2), pp. 253–261.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 107–119, DOI 10.21307/stattrans-2019-017

ESTIMATION OF ENERGY INTENSITY IN INDIAN IRON AND STEEL SECTOR: A PANEL DATA ANALYSIS

Anukriti Sharma[1], **Hiranmoy Roy**[2], **Narendra Nath Dalei**[3]

ABSTRACT

India holds the third position in the world as energy consumer of fossil fuels (BP SRWE, 2016). The total primary energy consumption in India in 2015 was 107 mtoe (The Economics Times, January 27, 2017). The industrial sector in India consumed about 30 percent (185 Mtoe) of aggregate final energy consumption of around 527 Mtoe in 2013. (India Energy Outlook, IEA, 2015). One of the most energy-intensive sectors is the Iron and Steel sector which consumes 25 percent of the total energy consumption. The energy consumption in Indian Iron and Steel sector is on the declining trend. It has declined from 10 GCal / tcs in 1990 to 6.9 GCal / tcs in 2010-11. About 20-40 percent of the total production in steel industry is energy cost. Therefore, energy cost share is important in deciding price of steel. Energy Conservation Act, 2001 (ECA) and formulation of Bureau of Energy Efficiency is an important initiative taken up by government in order to reduce energy consumption by various sectors in the Indian economy. Another important initiative is launching of first of its kind market-based mechanism, Perform, Achieve and Trade (PAT) mechanism in 2010 particularly targeting the energy consumption by industrial sector of the economy. Phase-I for PAT ran from 2012-2015 including eight most energy-intensive sectors under Indian Industrial sector of which Iron and Steel sector being a prominent sector. The objective of this paper is to empirically estimate the energy intensity of Indian Iron and Steel sector, also accounting for impact of ECA and PAT Phase-I in dummy variable form. The results indicate that the decline in energy consumption till 2011 by this sector can also be attributed to Energy Conservation Act implemented in the year 2001 along with other factors. This is empirically confirmed by our results that ECA has a significant impact over reduction of energy intensity of the steel firms. PAT doesn't seem to have much impact over energy intensity alone but the years where both PAT and ECA are prevalent, i.e., from 2012 to 2015 there seems to be a significant impact of around 0.050 reduction in energy intensity as accounted by different models in this paper. There is one more observation from the empirical results, that profit margin intensity was indirectly related to energy intensity implying more profitable firms invest more in energy efficiency.

Key words: energy intensity, Indian Iron and Steel sector, Energy Conservation Act, Perform-Achieve-Trade Mechanism, panel data.

[1] Doctoral Research Fellow, Department of Economics, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India. E-mail: anukriti1807@gmail.com. ORCID ID: https://orcid.org/0000-0002-2860-9227

[2] Associate Professor, Department of Economics, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India.

[3] Assistant Professor, Department of Economics, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India.

1. Introduction

Energy is the most important constituent that is necessary for all development in the economy. In fact the relation between the two is a prominent one, for a country to develop energy is required. Energy consumption in India has been steadily increasing. According to BP Energy Outlook 2017, India's energy consumption growth rate is 4.2 percent a year which is faster than all major countries in the world and will overtake China. Among Asian countries, India is the second largest energy consumer since 2008, surpassing Japan as the world's third largest oil consumer behind US and China.

India holds the third position in the world as primary energy consumer which includes fossil fuels like coal, oil, etc. (BP SRWE, 2016). Total primary energy consumption in India in 2015 was 107 mtoe (The Economics Times, January 27, 2017). The industrial sector in India consumed about 30 percent (185 Mtoe) of the total final energy consumption of around 527 Mtoe in 2013. (India Energy Outlook, IEA, 2015). In the list of GHG emitters in the world, India holds third rank after China and U.S. in 2016, with its greenhouse gas emissions increasing at a high rate of 4.7 percent in comparison to the last year (Netherlands Environmental Assessment Agency, September 29, 2017). Industries contribute approximately one fourth of India's total GHG emissions (Gupta et al. 2017). One of the most energy-intensive sectors is the Iron and Steel sector which consumes 25% of the total energy consumption (IEA, 2012). Energy consumption in Indian steel plants is high in comparison to world average as well mainly due to obsolete technology but it is gradually improving (Ministry of Steel, 2017). The Indian Iron and Steel sector contributed to about 28 percent of the emissions by the industrial sector in 2007. (Krishnan et al., 2013)

As per Worldsteel Association, in 2016, India ranked third in terms of steel production after China and Japan. The steel sector contribution to India's GDP is approximately 2 percent in 2015-16 (Ministry of Steel, GoI, 2016).

In order to reduce nation's energy intensity and emission intensity, Energy efficiency and low carbon growth have become apparent pathways.

In order to reduce energy consumption and promote energy efficiency in the country, Ministry of Power introduced the Energy Conservation Act in 2001. The Act proposed adherence energy norms for energy consumption for heavy consumers, developed Energy Conservation Building Code for new buildings, standards for performance in energy efficiency and also display of labels on appliances indicating their energy consumption. Under this Act, Bureau of Energy Efficiency (BEE) was formulated to implement provisions defined by the Act. The strengthening and amendments in Act was done in 2010 (Tata Strategic, 2014).

In addition to this, National Action Plan on Climate Change (NAPCC) was launched in 2008. Under this, National Mission for Enhanced Energy Efficiency (NMEEE) came into picture.

One of the important initiatives promulgated under NMEEE is Perform Achieve and Trade scheme, under which most energy intensive units such as Thermal power plants, Steel, Cement, Aluminium, Chlor Alkali, Textiles, Pulp & Paper, Fertilizers (known as Designated Consumers) has been assigned energy efficiency improvement targets. This created Tradable Energy Savings Certificates (ESCerts) under PAT scheme. Companies not able to meet their

target buy tradable energy saving certificates from those over achieving the target, creating an Energy Savings market in India.

PAT is a cost-effective mix of regulation in terms of mandatory energy saving targets along with formation of market for trading of these energy saving white certificates.

Total 67 plants of iron and steel are assigned energy reduction targets. For Iron and steel sector the threshold limit of energy consumption per annum is 30000 TOE (BEE, 2017). By the end of PAT Cycle-I, energy savings equivalent of 2.10 Million tonne of oil equivalent annually has been achieved, which is around 41% higher than the savings targets from 67 number of notified DCs. (Oak, 2017).

2. Literature review

Kumar (2003) and Sahu and Narayan (2009) has conducted a study to find out the factors affecting energy intensity of manufacturing industries. They used multiple regression technique to carry out their analysis. Kumar used eight years data for 1342 firms for their analysis whereas 2350 firms data for the year 2008 for their analysis. In 2017, Oak published a paper on factors affecting energy intensity of firms in Indian cement industry and also quantifying the effect of Perform, Achieve and Trade effect using panel data fixed effect model and difference-in-difference technique. The source of data for all these studies is CMIE Prowess database. Most of the explanatory variables used in these studies are similar such as firm size, age of the firm, technology import intensity, ownership. According to the authors, ownership (foreign or domestic), firm size, age are the important determinants of energy intensity in Indian manufacturing industry. Oak (2017) found cement firms having higher energy intensity to be covered under PAT Cycle-I (2012-15) are correctly identified by Government of India though cement industry remains highly energy intensive sector.

Bhandari and Shrimali (2017) studied the effectiveness of PAT by carrying out semi-structured interviews of designated consumers, BEE and EESL and also used PAT booklet as secondary source of information. According to them, the set targets of PAT are not strict enough to cause any energy savings more than business-as-usual, may not causing any long-term investment in energy efficiency and PAT market may not form, it's too early to assess transaction costs. Amendment needs to cater these issues to make PAT more effective.

Teng (2012) carried out the similar analysis taking into account indigenous Research and Development to study the effect on energy intensity of Chinese industries. Mukherjee (2008) accounted for inter-state heterogeneity and carried out the similar analysis for the period of 1998-2003 using Data Envelopment Analysis for Indian industries.

3. Methodology

Our objective is to determine various factors affecting energy intensity of Indian Iron and Steel industries. Coal, Electricity and Natural Gas are the principal energy inputs used by Indian Iron and Steel sector and this makes it highly energy intensive. The minimum energy consumption by the DCs for this sector is

30,000 toe. By the end of first PAT Phase-I, energy savings equivalent of 2.10 Million tonne of oil equivalent annually has been achieved, which is around 41% higher than the saving targets from 67 Nos. of notified DCs. Since we want to determine the impact of PAT Phase-I as well on energy intensity of this industry we have particularly chosen those 18 firms which are included under PAT Phase-I for reducing their specific energy consumption and 7 other firms which are not included in PAT but belongs to size decile 1 category of Indian Steel sector as per CMIE ProwessIQ.

In PAT Phase-I (2012-2015), there are 67 DCs (plants) which are included, out of which we have selected 18 firms for our analysis as listed below:

Table 3.1. List of PAT Phase-I firms included in the study

S.No.	Firm
1	Bhushan Steel Ltd
2	ESSAR Steel
3	Rashtriya Ispat Nigam Ltd.
4	Steel Authority Of India Ltd.
5	Tata Sponge Iron Ltd.
6	Tata Steel Ltd.
7	Welspun Corp Ltd.
8	Aarti Steels Ltd.
9	Balasore Alloys Ltd.
10	Hira Ferro Alloys Ltd.
11	J S W Ispat Steel Ltd. [Merged]
12	Monnet Ispat & Energy Ltd.
13	Orissa Sponge Iron & Steel Ltd.
14	Sunflag Iron & Steel Co. Ltd.
15	Usha Martin Ltd.
16	Bhilai Engineering Corpn. Ltd.
17	Mukand Ltd.
18	Sharda Ispat Ltd.

Table 3.2. List of Non-PAT Phase-I firms included in the study

S.No.	Firm
1	Kalyani Steels Ltd.
2	Modern Steels Ltd.
3	Vardhman Industries Ltd.
4	Mahindra Ugine Steel Co. Ltd.(Merged)
5	Pennar Industries Ltd.
6	Tulsyan NEC Ltd.
7	Uttam Value Steels Ltd.

The data source for the study is CMIE ProwessIQ Version 1.80. The time period for the study has been taken from 1995-2015. Since we want to study the impact of both Energy Conservation Act, 2001 and Perform, Achieve and Trade (Phase-I), 2012-15 we have particularly taken the time span of 20 years. The names of the designated consumer of Iron and Steel industry have been taken from The Ministry of Power report published in July 2012.

In this study, dependent variable is Energy Intensity (EI) which is defined as the ratio of Power and Fuel expenses (in Billion) to Sales (in Billion). Due to data unavailability on energy consumption & output in physical units we have taken Power and Fuel Expenses (Rs. Billion) and Sales (Rs. Billion) to define Energy Intensity.

Table 3.3. The variables are defined as follows

Variable	Defined as (all values in Rs. Million)	Expected Relationship
Energy Intensity	Power and Fuel Expenses to Sales	
Profit Margin Intensity (PMI)	Profit After Tax to Sales	positive
Labor intensity	Salaries and Wages to Sales	negative
Capital intensity	Ratio of Net Fixed Assets to Sales	negative
Firm Size	Sales and Assets in three years (current year plus last two years)	negative

Table 3.3. The variables are defined as follows (cont.)

Variable	Defined as (all values in Rs. Million)	Expected Relationship
Technology Import intensity	Ratio of the sum (of the foreign exchange spending on the capital goods, raw materials and the foreign exchange spending on royalties, technical knowhow paid by the firm to foreign collaborations) to Sales	negative
Repairs Intensity	Ratio of total expenses on repairs of plants and Machineries to Sales	positive
Age	Current year minus year of incorporation	positive/negative
PAT dummy (pat)	This is a dummy variable capturing the effect of PAT Phase-I on energy intensity of firms defined as pat = 1 for the years 2012-15 and 0 otherwise.	negative
ECA dummy (eca)	This is a dummy variable capturing the effect of Energy Conservation Act, 2001 on energy intensity of firms defined as eca = 1 for the years 2001-2015 and 0 otherwise	negative
_Ipat_eca_1	This is a dummy variable capturing the effect of Energy Conservation Act, 2001 on energy intensity of firms defined as _Ipat_eca_1 = 1 for the years 2001-2015 and 0 otherwise	negative
_Ipat_eca_2	This is a dummy variable capturing the impact of both PAT and ECA simultaneously on energy intensity of firms defined as _Ipat_eca_2 = 1 for the years 2012-2015 and 0 otherwise	negative

All the variables are first corrected for inflation using Index numbers and then converted into natural log form. In this paper we have used Fixed Effect Model to estimate the impact of above factors on Energy Intensity of Steel firms.

Following is the suggestive Fixed Effect equation for the model:

$$\ln EI_{it} = \beta_0 + \beta_1 \ln A_{it} + \beta_2 \ln PMI_{it} + \beta_3 \ln LI_{it} + \beta_4 \ln RI_{it} + \beta_5 \ln SI_{it} + \beta_6 \ln CI_{it} + \beta_7 \ln TMI_{it} + \beta_8 ECA + \beta_9 PAT + \beta_{10}(\_Ipat\_eca\_1) + \beta_{11}(\_Ipat\_eca\_2) + \varepsilon_{it}$$

Where, the variables are described in the following table:

Table 3.4.

Model	Dependent Variable	Independent Variable
Model-1	Energy Intensity (EI)	Age of the firm (A) Profit Margin Intensity (PMI) Labour intensity (LI) Repairs Intensity (RI) Size of the Firm (SI) Capital intensity (CI) Technology Import Intensity (TMI) PAT {1 = 2012 to 2015, 0 = otherwise} ECA {1 = 2001 to 2015, 0 = otherwise}
Model-II (with PAT)	Energy Intensity (EI)	Age of the firm (A) Profit Margin Intensity (PMI) Labour intensity (LI) Repairs Intensity (RI) Size of the Firm (SI) Capital intensity (CI) Technology Import Intensity (TMI) PAT {1 = 2012 to 2015, 0 = otherwise}
Model-III (with ECA)	Energy Intensity (EI)	Age of the firm (A) Profit Margin Intensity (PMI) Labour intensity (LI) Repairs Intensity (RI) Size of the Firm (SI) Capital intensity (CI) Technology Import Intensity (TMI) ECA {1 = 2001 to 2015, 0 = otherwise}
Model-IV (with PAT and ECA)	Energy Intensity (EI)	Age of the firm (A) Profit Margin Intensity (PMI) Labour intensity (LI) Repairs Intensity (RI) Size of the Firm (SI) Capital intensity (CI) Technology Import Intensity (TMI) _lpat_eca_1{1 = 2001 to 2015, 0 = otherwise} _lpat_eca_2{1 = 2012 to 2015, 0 = otherwise}
Model-V (Tobit Regression with PAT and ECA)	Energy Intensity (EI)	Age of the firm (A) Profit Margin Intensity (PMI) Labour intensity (LI) Repairs Intensity (RI) Size of the Firm (SI) Capital intensity (CI) Technology Import Intensity (TMI) _lpat_eca_1{1 = 2001 to 2015, 0 = otherwise} _lpat_eca_2 {1 = 2012 to 2015, 0 = otherwise

4. Analysis

Table 4.1. Panel unit root tests

	LLC (Levin-Lin-Chu) Test		Breitung Test		HT (Harris-Tzavalis) Test	
Variables	**Level (Adjusted t*)**	**First Difference (Adjusted t*)**	**Level (lambda)**	**First Difference (lambda)**	**Level (rho)**	**First Difference (rho)**
Include Trend (Panel Means and Time Trend included)						
lnA	-29.4846***	-33.6975***	8.6887	7.0427(1.0000)	0.6874	0.6655(0.5606)
lnPMI	1.6640	-6.5752***	0.8454	-2.2194**	0.6026	0.1340***
lnEI	-1.3215	-7.2656***	-0.7045	-4.7738***	0.5801**	-0.0569***
lnLI	-6.0000***	-9.5107***	-2.3595***	-4.8624***	0.4971***	-0.1074***
lnRI	-3.2267***	-8.2206***	0.0501	-3.1826***	0.4553***	-0.1456***
lnSI	-23.2976***	-57.5117***	0.1171	-0.1638(0.4349)	0.6285	0.0782***
lnCI	0.1428	-7.9140***	1.8007	-4.4809***	0.5948**	-0.1465***
lnTMI	-1.8461**	-5.3417***	-1.7593**	-5.9932***	0.3406***	-0.3071***

Note - Level of Significance 5% - **, 10% - *, 1% - ***

4.1 Panel unit root tests

When we have a panel dataset our first step is to test for stationarity of all variables included in the study. For this, panel unit root test is conducted for all variables individually. A number of tests exists to test the stationarity of unit root. We have selected the two out of these namely, Levin–Lin–Chu (LLC) test, Breitung Test and Harris-Tzavalis (HT) test to enhance the robustness of the results. There is a problem of serial correlation with LLC test which cannot be completed removed therefore it has low power when we have small sample to test but it accounts for heterogeneity in various sections. The null hypothesis and alternate hypothesis of these unit root tests are there exist unit root implying that the variables are non-stationary and the alternative hypothesis is that there is no unit root implying that the variables are stationary. Table 4.1 shows the results of each variable for panel unit root tests. It can be seen from Table 4.1 that the variables lnA, lnLI, lnRI, lnSI and lnTMI in level form are statistically significant under the LLC test and the variables lnEI, lnLI, lnRI, lnCI and lnTMI in level form are statistically significant under HT test. Also, the variables lnLI and lnTMI at level are statistically significant under Breitung Test. The level of lnPMI is

statistically insignificant under all three panel unit root tests. However, after first-order differencing, it is found that all the variables become stationary. Therefore, we may conclude that each variable is integrated of order one, i.e. I(1).

Table 4.2. Panel Data Analysis

	Model-I	**Model-II**	**Model-III**	**Model-IV**	**Model-V**
	Fixed Effect (d.lnei)	**Fixed Effect (d.lnei) with PAT**	**Fixed Effect (d.lnei) with ECA**	**Regression (d.lnei) with PAT and ECA**	**Random Effects Tobit Regression (d.lnei) with PAT and ECA**
d.lna	.0430493 (.0236734)	.0420724 (.0236985)	.020151 (.0262972)	.0117013 (.0231499)	-.0251003 (.0519315)
d.lnpmi	-.029305** (.0126731)	-.0294189** (.0126285)	-.0277864** (.0124823)	-.0264096*** (.0082738)	.0102759 (.0144196)
d.lnli	.1217567 (.064475)	.1215724 (.0647464)	.0925055 (.0565193)	.0918982 (.0780218)	.0242819 (.1357845)
d.lnri	.1674165 (.1140994)	.1688265 (.1173236)	.1533531 (.1156504)	.157753 (.1717628)	-.1073084 (.2978768)
d.lnsi	-.0022371 (.0014945)	-.0022522 (.0014738)	-.0022353 (.0013608)	-.0021529 (.0019842)	-.0039505 (.003526)
d.lnci	.017824 (.0111022)	.0179308 (.0111871)	.0171829 (.0113085)	.0165042** (.0082753)	.0064548 (.0144504)
d.lntmi	-.0254261 (.0161362)	-.0256924 (.0161611)	-.0240584 (.0163596)	-.0236041 (.0184868)	.0198583 (.0322479)
d.lnei					
_cons	-.0038066*** (.0010709)	-.0036542*** (.0011662)	.0021938 (.0029389)	.0028769 (.0030286)	.1159452*** (.0132042)
eca			-.0065094** (.002748)		
pat		-.0005471 (.0015633)			
_lpat_eca_1				-.0072334** (.0031461)	-.0318942*** (.0055608)
_lpat_eca_2				-.0059626 (.0038211)	-.0496313*** (.0067687)
Number of obs.	500	500	500	500	500
Number of groups	25	25	25		25
F	F(7,24)=2.51	F(8,24)= 2.25	F(8,24)=3.40	F(9, 490) = 4.14	Wald chi2(9) = 65.42
Prob > F	0.0440	0.0595	0.0095	0.0000	Prob > chi2 = 0.0000

Our objective is to empirically estimate energy intensity of the Iron and Steel industry using various factors affecting it and also evaluating the impact of Energy Conservation Act, 2001 (ECA) and PAT Cycle-I on the Energy intensity of Iron and Steel Sector in India by accounting for these two in dummy variable form.

The results Table 4.2, (similar to Sahu and Narayan (2009)) indicate a positive relation of age with energy intensity in Model I, II, III and IV. Model V indicates a negative relation of age with energy intensity.

Profit margin intensity is found to be significant in almost all the regressions with a negative relation with energy intensity implying if profit margin intensity will increase energy intensity will decline. This may be interpreted as if profits are increased then industry will be able to invest more in energy efficiency thereby reducing energy consumption.

The coefficient of labor intensity was insignificant i.e. labor intensity does not seem to be affecting energy intensity of the firms in Steel sector. But as the results suggest there seems to be a positive relationship between energy intensity implying as higher the labor intensive firms higher will be the energy intensity of the production process.

As reported by most of the Models, there is a +ve relationship between repairs intensity and energy intensity implying as firms are spending more on repairs of plant and machinery their energy intensity is also high. Though the coefficient for this variable is not significant the positive relation is at par with findings of Sahu and Narayan (2009), an analysis of energy intensity of Indian Manufacturing.

As the size of industry increases, it will lead to decline in energy intensity as stated by the results of all the regressions. This is in line with the results of Kumar (2003) but in opposition of findings by Sahu and Narayan (2009) stating an inverted U-shaped relation between firm size and energy intensity. The negative relation can be interpreted as growth of industry will lead to more resources for investment in energy intensity and thereby reducing energy consumption means if the industry produce at large-scale its per unit energy consumption will decline.

As reported by all the regressions, capital intensity is +ve related with energy intensity implying more capital-intensive firms are more energy-intensive. Though this variable is found to be significant only in Model IV. This result is in line with Papadogonas et al. (2007) and Sahu and Narayan (2009), found similar result for Hellenic and Indian manufacturing sector respectively.

Though the coefficient of technological import intensity is not found to be significant in any of the Models, but there seems to a negative relation of this variable with energy intensity. This implies that as the firm spends more on technological imports from abroad it will lead to advancement and thereby reduce energy intensity of firms.

The ECA dummy capturing the impact of Energy Conservation Act, 2001 (ECA) on energy intensity of Steel companies has a significant and negative impact as depicted by Model III. The same result is also depicted by _Ipat_eca_1 dummy in Model IV and V. This implies ECA, 2001 has a significant impact in reducing the energy intensity of Steel Industry.

The dummy variable, PAT capturing the impact of Perform, Achieve and Trade Mechanism, Phase-I (2012-2015) doesn't seem to have any significant

impact on reducing energy intensity of Steel industry as reported by results of Model II.

As reported by Model V, _Ipat_eca_2 dummy is significant implying PAT and ECA both simultaneously prevalent from 2012 to 2015 seems to have impact on energy intensity of Steel industry thereby reducing energy consumption.

4. Conclusion

Out of the eight sectors covered under PAT cycle-I (2012-15), one of the most energy-intensive sectors is the iron and steel sector contributing to 15 percent of total energy consumption out of these.

Though the energy consumption in these two sector is on declining trend but still it forms 20-40 percent of the total production cost of steel (Worldsteel Association, 2017).

Also, iron and steel sector is on a rising trend due to high global and domestic demand of crude steel in the market and attained third rank after China and Japan. Total 67 plants of iron and steel are assigned energy reduction targets. For Iron and steel sector the threshold limit of energy consumption per annum is 30000 TOE (BEE, 2017).

The decline in energy consumption till 2011 by this sector can also be attributed to Energy Conservation Act implemented in the year 2001 along with other factors. This is also confirmed by the empirical results in our results that ECA has a significant impact over reduction of energy intensity of the steel firms.

PAT doesn't seem to have much impact over energy intensity alone (Model II) but the years where both PAT and ECA are prevalent, i.e., from 2012 to 2015 there seems to be a significant impact of around 0.050 reduction in energy intensity (Model V). Though, by the end of first PAT cycle-I, energy savings equivalent of 2.10 Million tonne of oil equivalent annually has been achieved, which is around 41% higher than the saving targets from 67 Nos. of notified DCs. PAT may not have seem to impact much by our empirical results might be because PAT has defined Designated consumers on the basis of plant level data and due to non-availability of data we are bound to take firm level data for our analysis.

There is one more observation from the empirical results, that profit margin intensity was found to be negatively related to energy intensity implying more profitable firms invest more in energy efficiency.

REFERENCES

BP, (2016). Statistical Review of World Energy.

BHANDARI, DIVITA et al., (2017). The perform, achieve and trade scheme in India: An effectiveness analysis, Renewable and Sustainable Energy Reviews, Elsevier.

BUREAU, E. T., (2017). India's energy consumption to grow faster than major economies, The Economics Times, Retrieved from https://economictimes.indiatimes.com/industry/energy/oil-gas/indias-energy-consumption-to-grow-faster-than-major-economies/articleshow/56800587.cms.

CENTRE FOR MONITORING INDIAN ECONOMY, CMIE ProwessIQ Database, VERSION 1.81.

CII, (2013). Technology Compendium on Energy saving Opportunities Iron & Steel Sector, Shakti Sustainable Energy Foundation, Bureau of Energy Efficiency, BEE, August.

IEA, (2015). India Energy Outlook, International Energy Agency, World Energy Outlook Special Report, Paris, France

IEA, (2012). Energy Transition for Industry: India and the Global Context, International Energy Agency, Paris, France.

KRISHNAN, S. S. et al., (2013), A Study of Energy Efficiency in Indian Iron and Steel Industry, Shakti Sustainable Energy Foundation, Center for Study of Science, Technology & Policy, December.

MINISTRY OF STEEL, (2017). Energy And Environment Management In Iron & Steel Sector, Government of India, Retrieved from, http://steel.gov.in/technicalwing/energy-and-environment-management-iron-steel-sector.

MALAVIKA VYAWAHARE, (2017). India saw largest rise in GHG emissions in 2016 among major emitters, Hindustan Times, New Delhi, Retrieved from https://www.hindustantimes.com/india-news/india-among-highest-greenhouse-gas-emitters-in-2016-big-coal-consumer/story-juJex1dknBvLxmQ275YN0K.html7.

MUKHERJEE, K., (2008). Energy use efficiency in the Indian manufacturing sector: An interstate analysis, Energy Policy, Vol. 36, pp. 662–672.

OAK, HENA, (2017). Factors Influencing Energy Intensity of Indian Cement Industry, International Journal of Environmental Science and Development, Vol. 8 (5), May.

KUMAR, ALOK, (2003). Energy Intensity: A Quantitative Exploration for Indian Manufacturing, Working Paper, Indira Gandhi Institute of Development Research, Mumbai.

SAHU, SANTOSH, NARAYANAN, K., (2009). Determinants of Energy Intensity: A Preliminary Investigation of Indian Manufacturing Industries", Paper presented in the 44th Conference of The Indian Econometrics Society, at Guwahati University, Assam, India & Available at, http://mpra.ub.uni-muenchen.de/16606/.

TENG, (2012). Indigenous R&D, technology imports and energy consumption intensity: Evidence from industrial sectors in China, Energy Procedia, Vol. 16, pp. 2019–2026.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 123–138, DOI 10.21307/stattrans-2019-018

VARIABLE SELECTION IN MULTIVARIATE FUNCTIONAL DATA CLASSIFICATION

Tomasz Górecki[1], **Mirosław Krzyśko**[2], **Waldemar Wołyński**[3]

ABSTRACT

A new variable selection method is considered in the setting of classification with multivariate functional data (Ramsay and Silverman (2005)). The variable selection is a dimensionality reduction method which leads to replace the whole vector process, with a low-dimensional vector still giving a comparable classification error. Various classifiers appropriate for functional data are used. The proposed variable selection method is based on functional distance covariance (dCov) given by Székely and Rizzo (2009, 2012) and the Hilbert-Schmidt Independent Criterion (HSIC) given by Gretton et al. (2005). This method is a modification of the procedure given by Kong et al. (2015). The proposed methodology is illustrated with a real data example.

Key words: multivariate functional data, variable selection, dCov, HSIC, classification.

1. Introduction

In recent years, much attention has been paid to methods for representing data as functions or curves. Such data are known in the literature as functional data (Ramsay and Silverman (2005), Horváth and Kokoszka (2012)). Applications of functional data can be found in various fields, including medicine, economics, meteorology and many others. In many applications there is a need to use statistical methods for objects characterized by multiple variables observed at many time points (doubly multivariate data). Such data are called multivariate functional data. In this paper we focus on the classification problem for multivariate functional data. In many cases, in the classification procedures, the number of predictors p is significantly greater than the sample size n. Thus, it is natural to assume that only a small number of predictors are relevant to response Y.

Various basic classification methods have also been adapted to functional data, such as linear discriminant analysis (Hastie et al. (1995)), logistic regression (Rossi

[1]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. E-mail: tomasz.gorecki@amu.edu.pl. ORCID ID: https://orcid.org/0000-0002-9969-5257.

[2]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. Interfaculty Institute of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland. E-mail: mkrzysko@amu.edu.pl. ORCID ID: https://orcid.org/0000-0001-0075-4432.

[3]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland.
E-mail: wolynski@amu.edu.pl. ORCID ID: https://orcid.org/0000-0002-0777-9163.

et al. (2002)), penalized optimal scoring (Ando (2009)), $k$nn (Ferraty and Vieu (2003)), SVM (Rossi and Villa (2006)), and neural networks (Rossi et al. (2005)). Moreover, the combining of classifiers has been extended to functional data (Ferraty and Vieu (2009)). Górecki et al. (2016) adapted multivariate regression models to the classification of multivariate functional data. Gretton et al. (2005) defined the measure of dependence between random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ called the Hilbert-Schmidt Independence Criterion (HSIC) and proved that this measure is equal to zero if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent to each other when using universal kernels, such as the Gaussian kernels. Based on the idea of HSIC between two random vectors, we introduced the HSIC between two random processes.

Székely et al. (2007), Székely and Rizzo (2009, 2012, 2013) defined the measures of dependence between random vectors: the distance covariance (dCov). These authors showed that for all random variables with finite first moments, dCov generalizes the idea of covariance in two ways. Firstly, this coefficient can be applied when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are of any dimensions and not only for the simple case where $p = q = 1$. Secondly, dCov is equal to zero if and only if there is independence between the random vectors. Indeed, the distance covariance measures a linear relationship and can be equal to 0 even when the variables are related. Based on the idea of the distance covariance between two random vectors, we introduced the functional distance covariance between two random processes. We select a set of important predictors with a large value of functional distance covariance or functional Hilbert-Schmidt Independent Criterion. Our selection procedure is a modification of the procedure given by Kong et al. (2015).

An entirely different approach to the variable selection in functional data classification is presented by Berrendero et al. (2016). It is clear that variable selection has, at least, an advantage when compared with other dimension reduction methods (functional principal component analysis (FPCA), see Górecki et al. (2014), Jacques and Preda (2014), functional partial least squares (FPLS) methodology, see Delaigle and Hall (2012), and other methods) based on general projections: the output of any variable selection method is always directly interpretable in terms of the original variables, provided that the required number d of the selected variables is not too large.

The rest of this paper is organized as follows. In Section 2 we present the classification procedures used through the paper. In Section 3 we present the problem of representing functional data by orthonormal basis functions. In Section 4, we define a functional distance covariance. In Section 5 we define a functional HSIC. In Section 6 we propose a variable selection procedure based on the functional distance covariance and on HSIC. In Section 7 we illustrate the proposed methodology through a real data example. We conclude in Section 8.

2. Classifiers

The classification problem involves determining a procedure by which a given object can be assigned to one of q populations based on observation of p features of that

object.

The object being classified can be described by a random pair $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top \in \mathbb{R}^p$ and $Y \in \{1, \ldots, q\}$. An automated classifier can be viewed as a method of estimating the posterior probability of membership in groups. For a given $\boldsymbol{X}$, a reasonable strategy is to assign $\boldsymbol{X}$ to that class with the highest posterior probability. This strategy is called the Bayes' rule classifier.

2.1. Linear and quadratic discriminant classifiers

Now we make the Bayes' rule classifier more specific by the assumption that all multivariate probability densities are multivariate normal having arbitrary mean vectors and a common covariance matrix. We shall call this model the linear discriminant classifier (LDC). Assuming that class-covariance matrices are different, we obtain quadratic discriminant classifier (QDC).

2.2. Naive Bayes classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with independence assumptions. When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a one-dimensional normal distribution or we estimate density by kernel method.

2.3. k-nearest neighbour classifier

Most often we do not have sufficient knowledge of the underlying distributions. One of the important nonparametric classifiers is a k-nearest neighbour classifier (kNN classifier). Objects are assigned to the class having the majority in the k nearest neighbours in the training set.

2.4. Multinomial logistic regression

It is a classification method that generalizes logistic regression to multiclass problem using one vs. all approach.

3. Functional data

We now assume that the object being classified is described by a p-dimensional random process $\boldsymbol{X} = (X_1, X_2, ..., X_p)^\top \in L_2^p(I)$, where $L_2(I)$ is the Hilbert space of square-integrable functions, and $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{0}$.

Moreover, assume that the kth component of the vector $\boldsymbol{X}$ can be represented by a finite number of orthonormal basis functions $\{\varphi_b\}$

$$X_k(t) = \sum_{b=0}^{B_k} \alpha_{kb}\varphi_b(t),\ t \in I,\ k = 1, \ldots, p,$$

where $\alpha_{k0}, \alpha_{k1}, \ldots, \alpha_{kB_k}$ are the unknown coefficients.

Let $\boldsymbol{\alpha} = (\alpha_{10}, \ldots, \alpha_{1B_1}, \ldots, \alpha_{p0}, \ldots, \alpha_{pB_p})^\top \in \mathbb{R}^{K+p}$, $K = B_1 + \cdots + B_p$ and

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}_1^\top(t) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}_2^\top(t) & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\varphi}_p^\top(t) \end{bmatrix},$$

where $\boldsymbol{\varphi}_k(t) = (\varphi_0(t), ..., \varphi_{B_k}(t))^\top$, $k = 1, ..., p$.

Using the above matrix notation, the process $\boldsymbol{X}$ can be represented as:

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\alpha}, \tag{1}$$

where $\mathrm{E}(\boldsymbol{\alpha}) = \mathbf{0}$. This means that the realizations of the process $\boldsymbol{X}$ are in finite dimensional subspace of $L_2^p(I)$. We will denote this subspace by $\mathscr{L}_2^p(I)$.

We can estimate the vector $\boldsymbol{\alpha}$ on the basis of n independent realizations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of the random process $\boldsymbol{X}$ (functional data). We will denote this estimator by $\hat{\boldsymbol{\alpha}}$.

Typically data are recorded at discrete moments in time. Let x_{kj} denote an observed value of the feature X_k, $k = 1, 2, \ldots, p$ at the jth time point t_j, where $j = 1, 2, ..., J$. Then our data consist of the pJ pairs (t_j, x_{kj}). These discrete data can be smoothed by continuous functions x_k and I is a compact set such that $t_j \in I$, for $j = 1, ..., J$.

Details of the process of transformation of discrete data to functional data can be found in Ramsay and Silverman (2005) or in Górecki et al. (2014).

4. Distance covariance (dCov)

For the jointly distributed random process $\boldsymbol{X} \in L_2^p(I)$ and the random vector $\boldsymbol{Y} \in \mathbb{R}^q$, let

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l}, \boldsymbol{m}) = \mathrm{E}\{\exp[i\langle \boldsymbol{l}, \boldsymbol{X}\rangle + i\langle \boldsymbol{m}, \boldsymbol{Y}\rangle_q]\}$$

be the joint characteristic function of $(\boldsymbol{X}, \boldsymbol{Y})$, where

$$\langle \boldsymbol{l}, \boldsymbol{X}\rangle = \int_I \boldsymbol{l}'(t)\boldsymbol{X}(t)dt$$

and

$$\langle \boldsymbol{m}, \boldsymbol{Y}\rangle = \boldsymbol{m}'\boldsymbol{Y}.$$

Moreover, we define the marginal characteristic functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows: $f_{\boldsymbol{X}}(\boldsymbol{l}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l}, \mathbf{0})$ and $f_{\boldsymbol{Y}}(\boldsymbol{m}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\mathbf{0}, \boldsymbol{m})$.

Here, for generality, we assume that $\boldsymbol{Y} \in \mathbb{R}^q$, although the label Y in the classification problem is a random variable, with values in $\{1, \ldots, q\}$. Label $\boldsymbol{Y}$ has to be transformed into the label vector $\boldsymbol{Y} = (Y_1, \ldots, Y_q)'$, where $Y_i = 1$ for $i = 1, \ldots, q$ if $\boldsymbol{X}$ belongs to class i, and 0 otherwise.

Now, let us assume that $\boldsymbol{X} \in \mathscr{L}_2^p(I)$. Then, the process $\boldsymbol{X}$ has the representation (1).

In this case, we may assume (Ramsay and Silverman (2005)) that the vector weight function $\boldsymbol{l}$ and the process $\boldsymbol{X}$ are in the same space, i.e. the function $\boldsymbol{l}$ can be written in the form

$$\boldsymbol{l}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\lambda}, \tag{2}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{K+p}$.

Hence

$$\langle \boldsymbol{l}, \boldsymbol{X} \rangle = \int_I \boldsymbol{l}'(t)\boldsymbol{X}(t)dt = \boldsymbol{\lambda}'[\int_I \boldsymbol{\Phi}'(t)\boldsymbol{\Phi}(t)dt]\boldsymbol{\alpha} = \boldsymbol{\lambda}'\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ are vectors occurring in the representations (1) and (2) of the process $\boldsymbol{X}$ and function $\boldsymbol{l}$, and

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{m}) = \mathrm{E}\{\exp[i\boldsymbol{\lambda}'\boldsymbol{\alpha} + i\boldsymbol{m}'\boldsymbol{Y}]\} = f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m}),$$

where $f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m})$ is the joint characteristic function of the pair of random vectors $(\boldsymbol{\alpha},\boldsymbol{Y})$.

On the basis of the idea of distance covariance between two random vectors (Székely et al. (2007)), we can introduce functional distance covariance between random process $\boldsymbol{X}$ and random vector $\boldsymbol{Y}$.

Definition 1. *A nonnegative number* $\mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y})$ *defined by*

$$\mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{dCov}(\boldsymbol{\alpha},\boldsymbol{Y}),$$

where

$$\mathrm{dCov}^2(\boldsymbol{\alpha},\boldsymbol{Y}) = \frac{1}{C_{K+p}C_q}\int_{\mathbb{R}^{K+p+q}} \frac{|f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\lambda})f_{\boldsymbol{Y}}(\boldsymbol{m})|^2}{\|\boldsymbol{\lambda}\|_{K+p}^{K+p+1}\|\boldsymbol{m}\|_q^{q+1}} d\boldsymbol{\lambda}\, d\boldsymbol{m},$$

and $|z|$ *denotes the modulus of* $z \in \mathbb{C}$, $\|\boldsymbol{\lambda}\|_{K+p}$, $\|\boldsymbol{m}\|_q$ *the standard Euclidean norms on the corresponding spaces* V *chosen to produce scale free and rotation invariant measure that does not go to zero for dependent random vectors, and*

$$C_r = \frac{\pi^{\frac{1}{2}(r+1)}}{\Gamma(\frac{1}{2}(r+1))}$$

is half the surface area of the unit sphere in $\mathbb{R}^{r+1}$*, is called a functional distance covariance between the random process* $\boldsymbol{X}$ *and the random vector* $\boldsymbol{Y}$*.*

We can estimate functional distance covariance using data set $\boldsymbol{S} = \{(\hat{\boldsymbol{\alpha}}_1,\boldsymbol{y}_1),\ldots,(\hat{\boldsymbol{\alpha}}_n,\boldsymbol{y}_n)\}$.

Let

$$\bar{\boldsymbol{\alpha}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\alpha}}_k, \quad \bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_k,$$
$$\tilde{\boldsymbol{\alpha}}_k = \hat{\boldsymbol{\alpha}}_k - \bar{\boldsymbol{\alpha}}, \quad \tilde{\boldsymbol{y}}_k = \boldsymbol{y}_k - \bar{\boldsymbol{y}},\ k = 1,\ldots,n$$

and

$$\boldsymbol{A} = (a_{kl}), \quad \boldsymbol{B} = (b_{kl}),$$
$$\tilde{\boldsymbol{A}} = (A_{kl}), \quad \tilde{\boldsymbol{B}} = (B_{kl}),$$

where

$$a_{kl} = \|\hat{\boldsymbol{\alpha}}_k - \hat{\boldsymbol{\alpha}}_l\|_{K+p}, \quad b_{kl} = \|\boldsymbol{y}_k - \boldsymbol{y}_l\|_q,$$
$$A_{kl} = \|\tilde{\boldsymbol{\alpha}}_k - \tilde{\boldsymbol{\alpha}}_l\|_{K+p}, \quad B_{kl} = \|\tilde{\boldsymbol{y}}_k - \tilde{\boldsymbol{y}}_l\|_q, \quad k,l = 1,\ldots,n.$$

Hence

$$\tilde{\boldsymbol{A}} = \boldsymbol{HAH}, \quad \tilde{\boldsymbol{B}} = \boldsymbol{HBH},$$

where

$$\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$$

is the centring matrix.

On the basis of the result of Székely et al. (2007), we have

$$\mathrm{dCov}(\boldsymbol{S}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}.$$

5. Hilbert-Schmidt Independent Criterion (HSIC)

Let $\boldsymbol{\phi}$ be a mapping from L_2^p to an inner product feature space $\mathscr{H}$, and $\boldsymbol{\psi}$ be a mapping from R^q to an inner product feature space $\mathscr{G}$.

Definition 2. *The cross-covariance operator $\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}}: \mathscr{G} \to \mathscr{H}$ is a linear operator defined as*

$$\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}} = \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}[\boldsymbol{\phi}(\boldsymbol{X}) \otimes \boldsymbol{\psi}(\boldsymbol{Y})] - \mu_{\boldsymbol{X}} \otimes \mu_{\boldsymbol{Y}},$$

for all $f \in \mathscr{H}$ and $g \in \mathscr{G}$, where the tensor product operator $f \otimes g: \mathscr{G} \to \mathscr{H}$, $f \in \mathscr{H}$, $g \in \mathscr{G}$, is defined as

$$(f \otimes g)h = f\langle g,h\rangle_{\mathscr{G}}, \text{ for all } h \in \mathscr{G}.$$

This is a generalization of the cross-covariance matrix between random vectors.

Moreover, by the definition of the Hilbert-Schmidt (HS) norm, we can compute the HS norm of $f \otimes g$ via

$$\|f \otimes g\|_{HS}^2 = \|f\|_{\mathscr{H}}^2 \|g\|_{\mathscr{G}}^2.$$

Gretton et al. (2005) defined the Hilbert-Schmidt Independence Criterion (HSIC) in the following way:

Definition 3. *Hilbert-Schmidt Independence Criterion (HSIC) is the squared Hilbert-Schmidt norm of the cross-covariance operator*

$$\mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y}) = \|\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}}\|_{HS}^2.$$

Now, let

$$k\colon \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$$

be a kernel function on $\mathbb{R}^p$.

This raises an interesting question: given a function of two variables $k(\boldsymbol{x},\boldsymbol{x}')$, does there exist a function $\boldsymbol{\phi}$ such that $k(\boldsymbol{x},\boldsymbol{x}') = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle_{\mathscr{H}}$? The answer is provided by Mercer's theorem (1909), which says, roughly, that if k is positive semi-definite then such a $\boldsymbol{\phi}$ exists.

Often, we will not know $\boldsymbol{\phi}$, but a kernel function k, which encodes the inner product in $\mathscr{H}$, instead.

Popular positive semi-definite kernel functions on $\mathbb{R}^p$ include the polynomial kernel of degree $d > 0$, $k(\boldsymbol{x},\boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^d$, the Gaussian kernel $k(\boldsymbol{x},\boldsymbol{x}') = \exp(-\lambda \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$, $\lambda > 0$, and the Laplace kernel $k(\boldsymbol{x},\boldsymbol{x}') = \exp(-\lambda \|\boldsymbol{x} - \boldsymbol{x}'\|)$, $\lambda > 0$. In this paper we use, the Gaussian kernel.

A feature mapping $\boldsymbol{\phi}$ is centred by subtracting from it its expectation, that is transforming $\boldsymbol{\phi}(\boldsymbol{x})$ to $\tilde{\boldsymbol{\phi}}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}) - \mathrm{E}_{\boldsymbol{X}}[\boldsymbol{\phi}(\boldsymbol{X})]$. Centring a positive semi-definite kernel function k consists in centring in the feature mapping $\boldsymbol{\phi}$ associated to k. Thus, the centred kernel $\tilde{k}$ associated to k is defined by

$$\begin{aligned}\tilde{k}(\boldsymbol{x},\boldsymbol{x}') &= \langle \boldsymbol{\phi}(\boldsymbol{x}) - \mathrm{E}_{\boldsymbol{X}}[\boldsymbol{\phi}(\boldsymbol{X})], \boldsymbol{\phi}(\boldsymbol{x}') - \mathrm{E}_{\boldsymbol{X}'}[\boldsymbol{\phi}(\boldsymbol{X}')] \rangle \\ &= k(\boldsymbol{x},\boldsymbol{x}') - \mathrm{E}_{\boldsymbol{X}}[k(\boldsymbol{X},\boldsymbol{x}')] - \mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{x},\boldsymbol{X}')] + \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{X},\boldsymbol{X}')],\end{aligned}$$

assuming the expectations exist. Here, the expectation is taken over independent copies $\boldsymbol{X}, \boldsymbol{X}'$. We see that, $\tilde{k}$ is also a positive semi-definite kernel. Note also that for a centred kernel $\tilde{k}$, $\mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[\tilde{k}(\boldsymbol{X},\boldsymbol{X}')] = 0$, that is, centring the feature mapping implies centring the kernel function.

Let $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ be a finite subset of $\mathbb{R}^p$. A feature mapping $\boldsymbol{\phi}$ is centred by subtracting from it its empirical expectation, i.e. leading to $\tilde{\boldsymbol{\phi}}(\boldsymbol{x}_i) = \boldsymbol{\phi}(\boldsymbol{x}_i) - \overline{\boldsymbol{\phi}}$, where $\overline{\boldsymbol{\phi}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{x}_i)$. The kernel matrix $\boldsymbol{K} = (K_{ij})$ associated to the kernel function k and the set $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ is centred by replacing it with $\widetilde{\boldsymbol{K}} = (\widetilde{K}_{ij})$ defined for all $i, j =$

$1,2,\ldots,n$ by

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{n}\sum_{i=1}^{n} K_{ij} - \frac{1}{n}\sum_{j=1}^{n} K_{ij} + \frac{1}{n^2}\sum_{i,j=1}^{n} K_{ij},$$

where $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, \ldots, n$.

The centred kernel matrix $\widetilde{\boldsymbol{K}}$ is a positive semi-definite matrix. Also, as with the kernel function $\frac{1}{n^2}\sum_{i,j}^{n} \tilde{K}_{ij} = 0$.

Let $\langle \cdot, \cdot \rangle_F$ denote the Frobenius product and $\| \cdot \|_F$ the Frobenius norm defined for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$ by

$$\begin{aligned} \langle \boldsymbol{A}, \boldsymbol{B} \rangle_F &= \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B}), \\ \|\boldsymbol{A}\|_F &= (\langle \boldsymbol{A}, \boldsymbol{A} \rangle_F)^{1/2}. \end{aligned}$$

Then, for any kernel matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$, the centred kernel matrix $\widetilde{\boldsymbol{K}}$ can be expressed as follows (Schölkopf et al.(1998)):

$$\widetilde{\boldsymbol{K}} = \boldsymbol{H}\boldsymbol{K}\boldsymbol{H},$$

where $\boldsymbol{H}$ ia a centering matrix.

Since $\boldsymbol{H}$ is the idempotent matrix ($\boldsymbol{H}^2 = \boldsymbol{H}$), then we get for any two kernel matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ based on the subset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of $\mathbb{R}^p$ and the subset $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of $\mathbb{R}^q$, respectively,

$$\langle \widetilde{\boldsymbol{K}}, \widetilde{\boldsymbol{L}} \rangle_F = \langle \boldsymbol{K}, \widetilde{\boldsymbol{L}} \rangle_F = \langle \widetilde{\boldsymbol{K}}, \boldsymbol{L} \rangle_F.$$

We may express HSIC in terms of kernel functions (Gretton et al. (2005)):

$$\begin{aligned} \operatorname{HSIC}(\boldsymbol{X}, \boldsymbol{Y}) = {} & \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}[k(\boldsymbol{X}, \boldsymbol{X}')l(\boldsymbol{Y}, \boldsymbol{Y}')] \\ & + \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{X}, \boldsymbol{X}')]\,\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[l(\boldsymbol{Y}, \boldsymbol{Y}')] \\ & - 2\,\mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}[\mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{X}, \boldsymbol{X}')]\,\mathrm{E}_{\boldsymbol{Y}'}[l(\boldsymbol{Y}, \boldsymbol{Y}')]]. \end{aligned}$$

Here, $\mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}$ denotes the expectation over independent pairs $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}', \boldsymbol{Y}')$.

Let

$$k^\star : \mathscr{L}_2^p(I) \times \mathscr{L}_2^p(I) \to \mathbb{R}$$

be a kernel function on $\mathscr{L}_2^p(I)$. For the multivariate functional data the Gaussian kernel has the form:

$$k^\star(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\lambda \|\boldsymbol{x} - \boldsymbol{x}'\|^2),\ \lambda > 0.$$

From the orthonormality of the basis functions, we have:

$$\begin{aligned}\|\boldsymbol{x}-\boldsymbol{x}'\|^2 &= \int_I (\boldsymbol{x}(t)-\boldsymbol{x}'(t))^\top(\boldsymbol{x}(t)-\boldsymbol{x}'(t))dt \\ &= \|\hat{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}'\|^2.\end{aligned}$$

Hence

$$k^\star(\boldsymbol{x},\boldsymbol{x}') = k(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\alpha}}'),$$

where $\hat{\boldsymbol{\alpha}}_1,\ldots,\hat{\boldsymbol{\alpha}}_n$ are vectors occurring in the representation (1).

Definition 4. *The empirical HSIC for functional data is defined as*

$$\mathrm{HSIC}(S^\star) = \frac{1}{n^2}\langle \boldsymbol{K}^\star,\boldsymbol{L}^\star\rangle_F,$$

where $S^\star = \{(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\}$, $\boldsymbol{K}^\star$ *and* $\boldsymbol{L}^\star$ *are kernel matrices based on the subsets* $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$, *and* $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n\}$ *of* $\mathscr{L}_2^p(I)$ *and* $\mathbb{R}^q$, *respectively.*

But $\boldsymbol{K}^\star = \boldsymbol{K}$, where $\boldsymbol{K}$ is the kernel matrix of size $n\times n$, which has its (i,j)th element K_{ij} given by $K_{ij} = k(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\alpha}}_j)$. $\boldsymbol{L}$ is the kernel matrix of size $n\times n$, which has its (i,j)th element L_{ij} given by $L_{ij} = l(\boldsymbol{y}_i,\boldsymbol{y}_j)$.

Hence

$$\mathrm{HSIC}(S^\star) = \mathrm{HSIC}(S),$$

where $S = \{(\hat{\boldsymbol{\alpha}}_1,\boldsymbol{y}_1),\ldots,(\hat{\boldsymbol{\alpha}}_n,\boldsymbol{y}_n)\}$.

6. Variable selection based on the fuctional dCov and the functional HSIC

In this Section we propose the selection procedure built on the functional dCov and the functional HSIC. Let $\boldsymbol{Y} = (Y_1,\ldots,Y_q)'$, be the response vector, and $\boldsymbol{X} = (X_1,\ldots,X_p)'$ be the predictor p-dimensional process. Assume that only a small number of predictors are relevant to $\boldsymbol{Y}$. We will define an irrelevant variable to be one whose value is statistically independent of label vector $\boldsymbol{Y}$ and of the other variables $X_1,\ldots,X_p$.

We select a set of important predictors with large functional $\mathrm{dCov}(\boldsymbol{S})$ or with large functional $\mathrm{HSIC}(\boldsymbol{S})$.

We utilize the functional dCov because it allows for arbitrary relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$, regardless of whether it is linear or nonlinear. We would like an assurance that irrelevant variables do not increase dCov. Kong et al. (2015) proved the following theorem.

Theorem 1. *Let* $\boldsymbol{Z} = (\boldsymbol{X}^\top, X_{p+1})^\top$, *where* X_{p+1} *is an irrelevant variable. Then*

$$\mathrm{dCov}(\boldsymbol{Z},\boldsymbol{Y}) \le \mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y}).$$

Gretton et al. (2005) proved that $\mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent of each other. This is the direct motivation why we may also choose HSIC to measure the dependence. For the Gaussian kernel the following result is true.

Theorem 2. *Let* $\boldsymbol{Z} = (\boldsymbol{X}^\top, X_{p+1})^\top$*, where* X_{p+1} *is an irrelevant variable. Then*

$$\mathrm{HSIC}(\boldsymbol{Z},\boldsymbol{Y}) \leq \mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y}).$$

Proof. Since the variable X_{p+1} is independent of the label vector $\boldsymbol{Y}$ and the other variables $X_1,\ldots,X_p$, functions of these variables are also independent.

Hence

$$\begin{aligned}
\mathrm{HSIC}(\boldsymbol{Z},\boldsymbol{Y}) &= \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}[k(\boldsymbol{Z},\boldsymbol{Z}')l(\boldsymbol{Y},\boldsymbol{Y}')] + \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{Z},\boldsymbol{Z}')]\,\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')] \\
&\quad - 2\,\mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}\{\mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{Z},\boldsymbol{Z}')]\,\mathrm{E}_{\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')]\} \\
&= \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}[k(\boldsymbol{X},\boldsymbol{X}')\exp(-\lambda(X_{p+1}-X'_{p+1})^2)l(\boldsymbol{Y},\boldsymbol{Y}')] \\
&\quad + \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{X},\boldsymbol{X}')]\exp(-\lambda(X_{p+1}-X'_{p+1})^2)\,\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')] \\
&\quad - 2\,\mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}\{\mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{X},\boldsymbol{X}')\exp(-\lambda(X_{p+1}-X'_{p+1})^2)]\,\mathrm{E}_{\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')]\} \\
&= \mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y})\exp(-\lambda(X_{p+1}-X'_{p+1})^2) \leq \mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y}),
\end{aligned}$$

because $\exp(-\lambda(X_{p+1}-X'_{p+1})^2) \leq 1$, for $\lambda > 0$. □

The functional dCov and functional HSIC also permit univariate and multivariate response. Thus, this procedure is completely model-free.

We implemented the above theorems as a stopping rule in the selections of responses. The procedure took the following steps:

1. Calculate marginal functional dCov or functional HSIC for X_k, $k = 1,\ldots,p$ with the response $\boldsymbol{Y}$.
2. Rank the variables in decreasing order of the selected measure. Denote the ordered predictors as $X_{(1)},X_{(2)},\ldots,X_{(p)}$. Start with $\boldsymbol{X}_S = \{X_{(1)}\}$.
3. For k from 2 to p, keep adding $X_{(k)}$ to $\boldsymbol{X}_S$ if $\mathrm{dCov}(\boldsymbol{X}_S,\boldsymbol{Y})$ or $\mathrm{HSIC}(\boldsymbol{X}_S,\boldsymbol{Y})$ does not decrease. Stop otherwise.

7. Example

The described method was employed here to select the variables (pillars) in the classification problem of 115 countries in the period 2008-2017. Table 1 describes the variables (pillars) used in the analysis.

Table 1. Variables (pillars) used in analysis, 2008-2017

No.	Variable (pillar)
1.	Institutions
2.	Infrastructure
3.	Macroeconomic environment
4.	Health and primary education
5.	Higher education and training
6.	Goods market efficiency
7.	Labour market efficiency
8.	Financial market development
9.	Technological readiness
10.	Market size
11.	Business sophistication
12.	Innovation

For this purpose, the use was made of data published by the World Economic Forum (WEF) in its annual reports (`http://www.weforum.org`). Those are comprehensive data, describing exhaustively various socio-economic conditions or spheres of individual states. WEF experts have divided discussed countries into five groups (Figure 1).

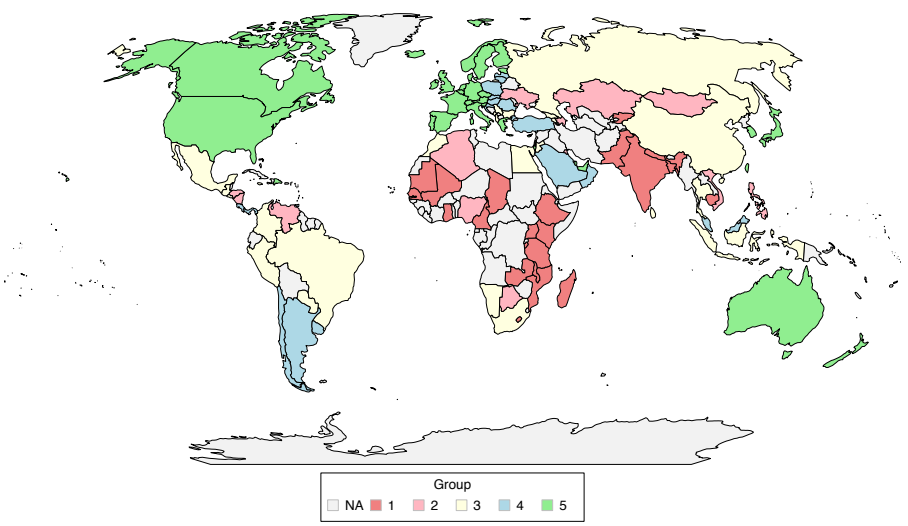

Figure 1: 115 countries used in the analysis

The data were transformed into functional data. Calculations were performed using the Fourier basis. In view of a small number of time periods, for each variable the maximum number of basis components was taken to be equal to five.

In the next step we applied the method of selecting variables described earlier (we stopped the procedure if the increase in the selected measure was less than 0.05). In such a way we obtained 5 variables (Figure 2 and Figure 3).

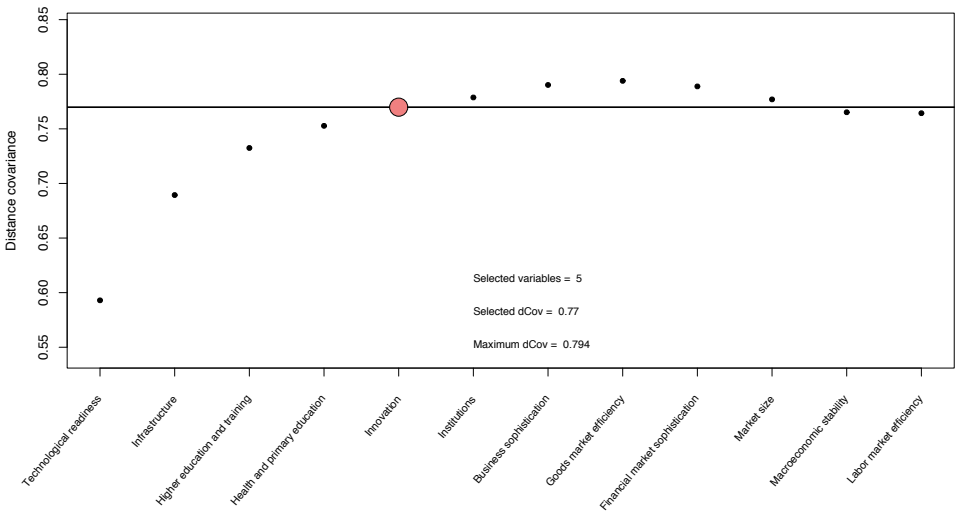

Figure 2: Variables selection for functional dCov

Next, we applied the described classifiers to reduced functional data and to full functional data. To estimate the error rate of the classifiers we used LOO CV (leave-one-out cross validation) method. The results are in Table 2.

Table 2. Classification accuracy (in %)

Classifier	Selected variables (5)	All variables (12)
LDC	71.30	66.09
kNN ($k = 1, \ldots, 8$)	77.39	71.30
Naive Bayes (normal)	69.57	65.22
Naive Bayes (kernel)	67.83	62.61
Logistic regression	60.87	56.52

We can observe that the error rate decreases if we reduce our data set. We can also notice that the order of classifiers stays unchanged (the best classifier for full data is kNN, and the same is the best for reduced data).

During the calculations we used `R` (R Core Team (2018)) software and `caret` (Kuhn (2018)), `energy` (Rizzo and Székely (2018)) and `fda` (Ramsay et al. (2018)) packages.

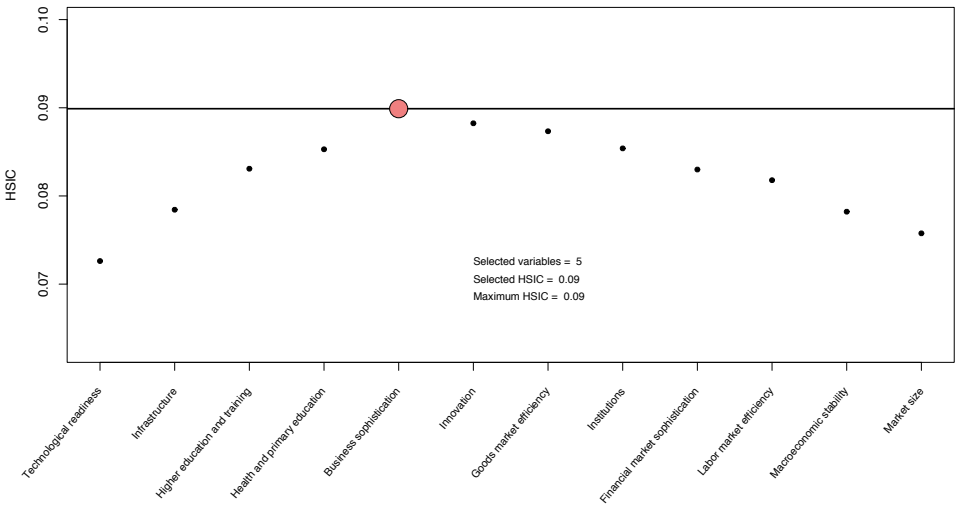

Figure 3: Variables selection for functional HSIC

8. Conclusions

The paper introduces variable selection for classification of multivariate functional data. The use of functional distance covariance or functional HSIC as a tool to reduce dimensionality of data set suggests that the technique provides useful results for classification of multivariate functional data. For analysed data set only five from twelve variables were included in the final model. We realize that the classification accuracy could drop slightly. However, we expect that this drop should be reasonable and in return we could gain a considerable amount of computation time.

In practice, it is important not to depend entirely on variable selection criteria because none of them works well under all conditions. So, our approach could be seen as a competitive to other variable selection methods and the full model without variables reduction. Finally, the researcher needs to evaluate the models using various diagnostic procedures.

REFERENCES

ANDO, T., (2009). Penalized optimal scoring for the classification of multi-dimensional functional data, Statistical Methodology, 6, pp. 565–576.

BERRENDERO, J. R., CUEVAS, A., TORRECILLA, J. L., (2016). Variable selection in functional data classification: a maxima-hunting proposal, Statistica Sinica, 26 (2), pp. 619–638.

DELAIGLE, A., HAAL, P., (2012). Methodology and theory for partial least squares applied to functional data. Annals of Statistics, 40, pp. 322–352.

FERRATY, F., VIEU, P., (2003). Curve discrimination. A nonparametric functional approach. Computational Statistics & Data Analysis, 44, pp. 161–173.

FERRATY, F., VIEU, P., (2009). Additive prediction and boosting for functional data. Computational Statistics & Data Analysis, 53 (4), pp. 1400–1413.

GÓRECKI, T., KRZYŚKO, M., WASZAK, Ł., WOŁYŃSKI, W., (2014). Methods of reducing dimension for functional data, Statistics in Transition new series, 15, pp. 231–242.

GÓRECKI, T., KRZYŚKO, M., WOŁYŃSKI, W., (2016). Multivariate functional regression analysis with application to classification problems, In: Analysis of Large and Complex Data, Studies in Classification, Data Analysis, and Knowledge Organization, Eds.: Wilhelm Adalbert F. X., Kestler Hans A., Springer International Publishing, pp. 173–183.

GRETTON, A., BOUSQUET, O., SMOLA, A., SCHÖLKOPF, B., (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In: Algorithmic Learning Theory (S., Jain, H. U., Simon and E., Tomita, eds.), Lecture Notes in Computer Science, 3734, pp. 63–77, Springer, Berlin.

HASTIE, T. J., TIBSHIRANI, R. J., BUJA, A., (1995). Penalized discriminant analysis, Annals of Statistics, 23, pp. 73–102.

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

JACQUES, J., PREDA, C., (2014). Model-based clustering for multivariate functional data, Computational Statistics & Data Analysis, 71, pp. 92–106.

KONG, J., WANG, S., WAHBA G., (2015). Using distance covariance for improved variable selection with application to learning genetic risk models, Statistics in Medicine, 34, pp. 1708–1720.

KUHN, M., Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt, (2018), caret: Classification and Regression Training. R package version 6.0-80, https://CRAN.R-project.org/package=caret.

R Core Team (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

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

RAMSAY, J. O., WICKHAM, H. GRAVES, S., HOOKER, G., (2018). fda: Functional Data Analysis, R package version 2.4.8, https://CRAN.R-project.org/package=fda.

RIZZO, M. L., SZÉKELY, G. J., (2018). energy: E-Statistics: Multivariate Inference via the Energy of Data, R package version 1.7-5, https://CRAN.R-project.org/package=energy.

ROSSI, F., DELANNAYC, N., CONAN-GUEZA, B., VERLEYSENC, M., (2005). Representation of functional data in neural networks, Neurocomputing, 64, pp. 183–210.

ROSSI, F., VILLA, N., (2006). Support vector machines for functional data classification, Neural Computing, 69, pp. 730–742.

ROSSI, N., WANG, X., RAMSAY, J.O., (2002). Nonparametric item response function estimates with EM algorithm, Journal of Educational and Behavioral Statistics, 27, pp. 291–317.

SCHÖLKOPF, B., SMOLA, A. J., MÜLLER, K. R., (1998). Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation, 10, pp. 1299–1319.

SZÉKELY, G. J., RIZZO, M. L., BAKIROV, N. K., (2007). Measuring and testing dependence by correlation of distances, The Annals of Statistics, 35 (6), pp. 2769–2794.

SZÉKELY, G. J., RIZZO, M. L., (2009). Brownian distance covariance, Annals of Applied Statistics, 3 (4), pp. 1236–1265.

SZÉKELY, G. J., RIZZO, M. L., (2012). On the uniqueness of distance covariance, Statistical Probability Letters, 82 (12), pp. 2278–2282.

SZÉKELY, G. J., RIZZO, M. L., (2013). The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis, 117, pp. 193–213.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 139–153, DOI 10.21307/stattrans-2019-019

TESTING HYPOTHESES ABOUT STRUCTURE OF PARAMETERS IN MODELS WITH BLOCK COMPOUND SYMMETRIC COVARIANCE STRUCTURE

Roman Zmyślony[1], **Arkadiusz Kozioł**[2]

ABSTRACT

In this article we deal with testing the hypotheses of the so-called structured mean vector and the structure of a covariance matrix. For testing the above mentioned hypotheses Jordan algebra properties are used and tests based on best quadratic unbiased estimators (BQUE) are constructed. For convenience coordinate-free approach (see Kruskal (1968) and Drygas (1970)) is used as a tool for characterization of best unbiased estimators and testing hypotheses. To obtain the test for mean vector, linear function of mean vector with the standard inner product in null hypothesis is changed into equivalent hypothesis about some quadratic function of mean parameters (it is shown that both hypotheses are equivalent and testable). In both tests the idea of the positive and negative part of quadratic estimators is applied to get the test, statistics which have F distribution under the null hypothesis. Finally, power functions of the obtained tests are compared with other known tests like LRT or Roy test. For some set for parameters in the model the presented tests have greater power than the above mentioned tests. In the article we present new results of coordinate-free approach and an overview of existing results for estimation and testing hypotheses about BCS models.

Key words: coordinate-free approach, Jordan algebra, multivariate model, block compound symmetric covariance structure, best unbiased estimators, testing structure of mean vector, testing independence of block variables.

1. Coordinate-free approach and Jordan algebra

1.1. Expectation and covariance operator in finite dimensional space with inner product

Let $\mathscr{K}(\cdot,\cdot)$ be a finite dimensional space with an inner product $(\boldsymbol{a},\boldsymbol{b})$.

Definition 1. *We say that the vector $\boldsymbol{\eta}\in\mathscr{K}$ is the expectation of a random vector $\boldsymbol{y}\in\mathscr{K}$ if there exists $\boldsymbol{\eta}$ such that for all $\boldsymbol{a}\in\mathscr{K}$ the expectation*

$$\mathrm{E}(\boldsymbol{a},\boldsymbol{y})=(\boldsymbol{a},\boldsymbol{\eta}). \tag{1}$$

[1]Faculty of Mathematics, Computer Science and Econometrics, University of Zielona Góra, Szafrana 4a, 65-516 Zielona Góra, Poland. E-mail: r.zmyslony@wmie.uz.zgora.pl. ORCID ID: https://orcid.org/0000-0001-8029-0578.

[2]Faculty of Mathematics, Computer Science and Econometrics, University of Zielona Góra, Szafrana 4a, 65-516 Zielona Góra, Poland. E-mail: a.koziol@wmie.uz.zgora.pl. ORCID ID: https://orcid.org/0000-0002-5347-4109.

Lemma 1. *Expectation $\boldsymbol{\eta}$ is uniquely defined and does not depend on the choice of inner product.*

Proof. Suppose that $\boldsymbol{\eta}_1$ different from $\boldsymbol{\eta}_2$ are two expectation vectors, then for all $\boldsymbol{a} \in \mathscr{K}$ we have that $\mathrm{E}(\boldsymbol{a}, \boldsymbol{y}) = (\boldsymbol{\eta}_1, \boldsymbol{a}) = (\boldsymbol{\eta}_2, \boldsymbol{a})$. This is equivalent to $(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2, \boldsymbol{a}) = 0$ for all $\boldsymbol{a} \in \mathscr{K}$ and this is equivalent to $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$.
To prove the second part of this lemma let $[\cdot,\cdot]$ be an arbitrary inner product. Then, from the characterization of all inner products it is implied there exists a self-adjoint positive definite operator $\boldsymbol{A} = \boldsymbol{A}^*$ such that for all $\boldsymbol{a}, \boldsymbol{b} \in \mathscr{K}$ we have $[\boldsymbol{a}, \boldsymbol{b}] = (\boldsymbol{a}, \boldsymbol{A}\boldsymbol{b}) = (\boldsymbol{A}\boldsymbol{a}, \boldsymbol{b})$. From the definition of expectation we have that for all $\boldsymbol{a} \in \mathscr{K}$

$$\mathrm{E}[\boldsymbol{a}, \boldsymbol{y}] = [\boldsymbol{a}, \boldsymbol{\eta}_{[\cdot,\cdot]}]. \tag{2}$$

On the other hand, for all $\boldsymbol{a} \in \mathscr{K}$

$$\mathrm{E}[\boldsymbol{a}, \boldsymbol{y}] = \mathrm{E}(\boldsymbol{a}, \boldsymbol{A}\boldsymbol{y}) = \mathrm{E}(\boldsymbol{A}\boldsymbol{a}, \boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{a}, \boldsymbol{\eta}_{(\cdot,\cdot)}) = (\boldsymbol{a}, \boldsymbol{A}\boldsymbol{\eta}_{(\cdot,\cdot)}) = [\boldsymbol{a}, \boldsymbol{\eta}_{(\cdot,\cdot)}]. \tag{3}$$

From (2) and (3) we have that $\boldsymbol{\eta}_{[\cdot,\cdot]} = \boldsymbol{\eta}_{(\cdot,\cdot)}$. □

Definition 2. *Operator $\boldsymbol{\Sigma}_{(\cdot,\cdot)}$ is a covariance operator if for all $\boldsymbol{a}, \boldsymbol{b} \in \mathscr{K}$*

$$\mathrm{cov}((\boldsymbol{a}, \boldsymbol{y}), (\boldsymbol{b}, \boldsymbol{y})) = (\boldsymbol{a}, \boldsymbol{\Sigma}_{(\cdot,\cdot)}\boldsymbol{b}). \tag{4}$$

The following lemma shows that the covariance operator depends on the choice of inner product.

Lemma 2. *Operator $\boldsymbol{\Sigma}_{(\cdot,\cdot)}$ is uniquely defined and depends on inner product $(\cdot,\cdot)$, i.e. under $[\cdot,\cdot] = (\cdot, \boldsymbol{A}\cdot)$ operator $\boldsymbol{\Sigma}_{[\cdot,\cdot]} = \boldsymbol{\Sigma}_{(\cdot,\cdot)}\boldsymbol{A}$.*

Proof. The proof of uniqueness of the covariance operator is similar to the proof of uniqueness of expectation. To prove the second part of the lemma note that from the definition we have

$$\mathrm{cov}([\boldsymbol{a}, \boldsymbol{y}], [\boldsymbol{b}, \boldsymbol{y}]) = [\boldsymbol{a}, \boldsymbol{\Sigma}_{[\cdot,\cdot]}\boldsymbol{b}]. \tag{5}$$

On the other hand,

$$\begin{aligned} \mathrm{cov}([\boldsymbol{a}, \boldsymbol{y}], [\boldsymbol{b}, \boldsymbol{y}]) &= \mathrm{cov}((\boldsymbol{A}\boldsymbol{a}, \boldsymbol{y}), (\boldsymbol{A}\boldsymbol{b}, \boldsymbol{y})) = (\boldsymbol{A}\boldsymbol{a}, \boldsymbol{\Sigma}_{(\cdot,\cdot)}\boldsymbol{A}\boldsymbol{b}) = (\boldsymbol{a}, \boldsymbol{A}\boldsymbol{\Sigma}_{(\cdot,\cdot)}\boldsymbol{A}\boldsymbol{b}) && (6) \\ &= [\boldsymbol{a}, \boldsymbol{\Sigma}_{(\cdot,\cdot)}\boldsymbol{A}\boldsymbol{b}] && (7) \end{aligned}$$

From (5) and (7) it follows that $\boldsymbol{\Sigma}_{[\cdot,\cdot]} = \boldsymbol{\Sigma}_{(\cdot,\cdot)}\boldsymbol{A}$. □

Remark 1. *Through the paper we deal with $\mathbb{R}^n$ and the standard inner product. In the space of $m \times n$ matrices, which is denoted by $\mathscr{M}^{m,n}$, the inner product is defined as $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}')$. The space of $n \times n$ symmetric matrices will be denoted by $\mathscr{S}^n$. Because of symmetry the inner product in $\mathscr{S}^n$ is $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B})$. Moreover, throughout the paper $\boldsymbol{A}'$ will stand for transpose of matrix $\boldsymbol{A}$.*

1.2. Special linear operator on space of $\mathscr{M}^{m,n}$

Definition 3. *Let $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$ be matrices with such dimensions that multiplication $\boldsymbol{ACB}$ is possible. Then:*

$$(\boldsymbol{A}\odot\boldsymbol{B})\boldsymbol{C}=\boldsymbol{ACB}.$$

In the following remark we will show the relation between the Kronecker product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ ($\boldsymbol{A}\otimes\boldsymbol{B}$), which have orders $k\times l$ and $p\times q$, respectively, and the special operator $\odot$. In this paper, the Kronecker product is defined as block matrix $\boldsymbol{A}\otimes\boldsymbol{B}=a_{ij}\boldsymbol{B}$ for $i=1,\ldots,k$ and $j=1,\ldots,l$.

The operator vec is a linear transformation which converts a matrix into a column vector by stacking the columns of the matrix under another. The inverse operator to vec is vec_p^{-1} which converts a column vector into a matrix with p rows, such that $\mathrm{vec}_p^{-1}(\mathrm{vec}(\boldsymbol{X}))=\boldsymbol{X}$ for all matrices $\boldsymbol{X}$ of order $p\times k$ and $\mathrm{vec}(\mathrm{vec}_p^{-1}(\boldsymbol{x}))=\boldsymbol{x}$ for all vectors $\boldsymbol{x}$ with dimension $pk\times 1$.

Remark 2. *Let $\boldsymbol{Y}$ be a matrix order $q\times l$. The operator $\odot$ has a following properties:*

- $(\boldsymbol{A}\otimes\boldsymbol{B})\mathrm{vec}(\boldsymbol{Y})=\mathrm{vec}\left((\boldsymbol{B}\odot\boldsymbol{A}')\boldsymbol{Y}\right)$;
- $\mathrm{vec}_p^{-1}\left((\boldsymbol{A}\otimes\boldsymbol{B})\mathrm{vec}(\boldsymbol{Y})\right)=(\boldsymbol{B}\odot\boldsymbol{A}')\boldsymbol{Y}$;
- $(\boldsymbol{A}\odot\boldsymbol{B})(\boldsymbol{C}\odot\boldsymbol{D})=\boldsymbol{AC}\odot\boldsymbol{DB}$.

1.3. Jordan algebra and its properties

An associative algebra can be transformed into a Jordan algebra by the Jordan product $A\circ B=\frac{AB+BA}{2}$ (see Schafer (1966)). Through the paper we deal with Jordan algebras of matrices „formally real" in the sense that if $\boldsymbol{A}^2+\boldsymbol{B}^2+\ldots=\boldsymbol{0}$ then $\boldsymbol{A}=\boldsymbol{B}=\ldots=\boldsymbol{0}$ (see Jordan, Neumann and Wigner (1934)).

A full characterization of irreducible Jordan algebras of matrices is given by Jordan, Neumann and Wigner (1934) (for more details see also Massam (1994), Massam and Neher (1997), Letac and Massam (1998), Massam and Neher (1998), Faraut and Korányi (1994)):

- The algebra $\mathscr{S}^n$ of all $n\times n$ ($n\geq 1$) symmetric matrices with trace inner product and operation $A\circ B$;
- The algebra $\mathscr{L}^n$ (Lorentz spin algebra);
- The algebra $\mathscr{H}_n$ of all $n\times n$ complex Hermitian matrices with trace inner product and operation $A\circ B$;
- The algebra $\mathscr{Q}_n$ of all $n\times n$ quaternion Hermitian matrices with trace inner product operation $A\circ B$;
- The algebra $\mathscr{O}_3$ of all 3×3 octonion Hermitian matrices with trace inner product and operation $A\circ B$.

Remark 3. *Note that all Jordan algebras can be represented as Cartesian product of all above Jordan algebras. For statistics, the most important are Cartesian product of* $\mathbb{R}$ *(as a special case of the first one with* $n = 1$*, where multiplication is commutative* $(ab = ba)$*) and* $\mathcal{S}^n$ *(for* $n \geq 2$*). They were named as quadratic subspaces by Seely (1971). If matrices in Jordan algebra commute, i.e.* $\boldsymbol{AB} = \boldsymbol{BA}$*, then this algebra is isomorphic to Cartesian product of* $\mathbb{R}$*.*

Some of properties which we will use through the paper are given in the lemma below.

Lemma 3. *Let* $\boldsymbol{\vartheta}$ *be a quadratic subspace of* $\mathcal{S}^n$*. Then:*

1. $\boldsymbol{A} \in \boldsymbol{\vartheta} \Rightarrow \boldsymbol{A}^k \in \boldsymbol{\vartheta}$*;*
2. $\boldsymbol{A}, \boldsymbol{B} \in \boldsymbol{\vartheta} \Rightarrow \boldsymbol{ABA} \in \boldsymbol{\vartheta}$*;*
3. $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \boldsymbol{\vartheta} \Rightarrow \boldsymbol{ABC} + \boldsymbol{CBA} \in \boldsymbol{\vartheta}$*;*
4. $\boldsymbol{P}^2 = \boldsymbol{P}$*,* $\boldsymbol{P} = \boldsymbol{P}'$ *and* $\forall \boldsymbol{V} \in \boldsymbol{\vartheta}\ \boldsymbol{PV} = \boldsymbol{VP} \Rightarrow \boldsymbol{M\vartheta M'} = \boldsymbol{M\vartheta}$ *is a quadratic subspace, where matrix* $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}$*, while* $\boldsymbol{I}$ *stands for identity matrix;*
5. *If* $\boldsymbol{Q}$ *is an orthogonal matrix then* $\boldsymbol{Q\vartheta Q'}$ *is also a quadratic subspace.*

For the proof see Seely (1971) and also Zmyślony (1979).

2. Estimation and testing hypotheses in mixed models for univariate case

In this section we deal with estimation and testing hypotheses using coordinate-free approach and properties of Jordan algebra.

2.1. Estimation of parameters in mixed models

The well-known normal mixed model can be expressed as follows

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta}, \boldsymbol{V}(\boldsymbol{\sigma})), \tag{8}$$

where $\boldsymbol{\sigma} = \left(\sigma_1^2, \ldots, \sigma_m^2\right)'$ and $\boldsymbol{V}(\boldsymbol{\sigma}) = \sum_{i=1}^m \sigma_i^2 \boldsymbol{V}_i$. We shall note that $\mathcal{K} = \mathbb{R}^n$ with the standard inner product, $\mathcal{X} = \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\}$ and $\boldsymbol{\vartheta} = \operatorname{sp}\left\{\sum_{i=1}^m \sigma_i^2 \boldsymbol{V}_i : \sigma_i \geq 0, \boldsymbol{V}_i \text{ are known}\right\}$.

Remark 4. *We assume that there exists* $\boldsymbol{\sigma}_0$ *such that* $\boldsymbol{V}\left(\boldsymbol{\sigma}_0\right) = \boldsymbol{I}$*.*

Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\sigma}\}$ and $g(\boldsymbol{\theta})$ be a real-valued function. We consider the following classes of linear and quadratic estimators, respectively
$\mathcal{A} = \{(\boldsymbol{a}, \boldsymbol{y}) : \boldsymbol{a} \in \mathbb{R}^n, \mathrm{E}(\boldsymbol{a}, \boldsymbol{y}) = g(\boldsymbol{\theta})\}$,
$\mathcal{B} = \left\{\langle \boldsymbol{B}, \boldsymbol{yy'} \rangle : \boldsymbol{B} \in \mathcal{S}^n, \mathrm{E}\langle \boldsymbol{B}, \boldsymbol{yy'} \rangle = g(\boldsymbol{\theta})\right\}$.

Remark 5. *In the class $\mathscr{B}$ with $\mathscr{K} = \mathcal{S}^n$ and the inner product $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{AB})$ we get that the expectation of $\boldsymbol{yy}'$ is:*

$$\mathrm{E}\left(\boldsymbol{yy}'\right) = \boldsymbol{V}(\boldsymbol{\sigma}) + \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{X} = \boldsymbol{V}(\boldsymbol{\sigma}) + \mathrm{E}(\boldsymbol{y})\mathrm{E}(\boldsymbol{y})'$$

and covariance of $\boldsymbol{yy}'$ is:

$$\operatorname{cov}\left(\boldsymbol{yy}'\right) = 2\left[\mathrm{E}\left(\boldsymbol{yy}'\right) \otimes \mathrm{E}\left(\boldsymbol{yy}'\right) - \mathrm{E}(\boldsymbol{y})\mathrm{E}(\boldsymbol{y})' \otimes \mathrm{E}(\boldsymbol{y})\mathrm{E}(\boldsymbol{y})'\right].$$

We recall the definition of estimable function of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$.

Definition 4. *A function $\left[\boldsymbol{c}, \boldsymbol{\beta}\right]$ is said to be estimable if there exists a linear unbiased estimator for this function i.e. $\mathrm{E}(\boldsymbol{a}, \boldsymbol{y}) = \left[\boldsymbol{c}, \boldsymbol{\beta}\right]$.*

In the following two theorems conditions for the existence of estimators with optimal properties in a mixed linear model are given.

Theorem 1. *For any estimable function $\left[\boldsymbol{c}, \boldsymbol{\beta}\right]$ there exists its best linear unbiased estimator if and only if for $i = 1, \ldots, m$ holds $\boldsymbol{PV}_i = \boldsymbol{V}_i\boldsymbol{P}$, where $\boldsymbol{P} = \boldsymbol{XX}^+$, while $\boldsymbol{X}^+$ is Moore-Penrose inverse of matrix $\boldsymbol{X}$.*

Theorem 2. *For any estimable function $[\boldsymbol{c}, \boldsymbol{\sigma}]$ there exists its best quadratic unbiased estimator if and only if $\boldsymbol{M}\vartheta\boldsymbol{M}$ is quadratic subspace, where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{XX}^+$.*

Theorem 3. *For any quadratic estimable function there exists best quadratic unbiased estimators (BQUE) if and only if $\operatorname{sp}\left\{\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_m\right\}$ is quadratic subspace.*

For proofs of these theorems see Zmyślony (1978, 1980). From Seely (1972, 1977) and Zmyślony (1980), and since the estimators are functions of complete sufficient statistics, the following remarks follows.

Remark 6. *Best linear unbiased estimators and best quadratic unbiased estimators are best unbiased estimators.*

Suppose that

$$\underset{n\times 1}{\boldsymbol{y}} \sim \mathcal{N}(\mu\mathbf{1}, \boldsymbol{V}(\boldsymbol{\sigma})),$$

where $\mu \in \mathbb{R}$ and $\boldsymbol{V}(\boldsymbol{\sigma}) = \sigma_1^2\boldsymbol{V}_1 + \sigma_2^2\boldsymbol{V}_2 + \sigma_3^2\boldsymbol{V}_3$, while

$$\boldsymbol{V}_1 = \begin{bmatrix} \underset{n_1\times n_1}{\mathbf{11}'} & \mathbf{0} \\ \mathbf{0} & \underset{n_2\times n_2}{\mathbf{0}} \end{bmatrix}, \boldsymbol{V}_2 = \begin{bmatrix} \underset{n_1\times n_1}{\mathbf{0}} & \mathbf{0} \\ \mathbf{0} & \underset{n_2\times n_2}{\mathbf{11}'} \end{bmatrix}, \boldsymbol{V}_3 = \begin{bmatrix} \underset{n_1\times n_1}{\boldsymbol{I}} & \mathbf{0} \\ \mathbf{0} & \underset{n_2\times n_2}{\boldsymbol{I}} \end{bmatrix}.$$

Since the expectation of $\boldsymbol{yy}'$ is

$$\mathrm{E}\left(\boldsymbol{yy}'\right) = \mu^2\mathbf{11}' + \boldsymbol{V}(\boldsymbol{\sigma}) = \mu^2\mathbf{11}' + \sigma_1^2\boldsymbol{V}_1 + \sigma_2^2\boldsymbol{V}_2 + \sigma_3^2\boldsymbol{V}_3,$$

three following conditions for this model are satisfied:

1. $\vartheta = \mathrm{sp}\left\{\mathbf{1}\mathbf{1}', \boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{V}_3\right\}$ is a quadratic subspace,
2. $\frac{1}{n}\mathbf{1}\mathbf{1}'$ does not commute with $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$,
3. according to first characterization of Jordan algebra it means that ϑ can be represented as Cartesian product of 2×2 symmetric matrices and $\sigma_3^2\boldsymbol{I}$.

Remark 7. *According to Theorem 1 note that* $\boldsymbol{P} = \frac{1}{n}\mathbf{1}\mathbf{1}'$ *does not commute with* $\boldsymbol{V}(\boldsymbol{\sigma})$ *and thus the BLUE of* μ *does not exist. However, from Theorem 3, there exists the BQUE for* μ^2*.*

2.2. Tests for variance components based on unbiased estimators

For the normal model of the form given in (8) we consider the following hypotheses

$$H_0 : \sigma_i^2 = 0 \quad vs \quad H_1 : \sigma_i^2 > 0.$$

Let $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ be an unbiased estimator of σ_i^2. Moreover, let $\boldsymbol{A}_+$, $\boldsymbol{A}_-$ stand for positive and negative part of matrix $\boldsymbol{A}$, respectively.

Remark 8. *For* $i < k$ *the estimator* $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ *is "not defined", that is* $\boldsymbol{A} = \boldsymbol{A}_+ - \boldsymbol{A}_-$*, where* $\boldsymbol{A}_+, \boldsymbol{A}_- \geq 0$*, i.e.* $\boldsymbol{A}_+, \boldsymbol{A}_-$ *are nonnegative definite matrices different than* $\mathbf{0}$*. Note that*

- *if* H_0 *is true, then* $\mathrm{E}\left(\boldsymbol{y}'\boldsymbol{A}_+\boldsymbol{y}\right) = \mathrm{E}\left(\boldsymbol{y}'\boldsymbol{A}_-\boldsymbol{y}\right)$,
- *if* H_1 *is true, then* $\mathrm{E}\left(\boldsymbol{y}'\boldsymbol{A}_+\boldsymbol{y}\right) > \mathrm{E}\left(\boldsymbol{y}'\boldsymbol{A}_-\boldsymbol{y}\right)$.

Corollary 1. *The test should reject hypothesis*

$$H_0 : \sigma_i^2 = 0$$

if statistic

$$F = \frac{\boldsymbol{y}'\boldsymbol{A}_+\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{A}_-\boldsymbol{y}}$$

is sufficiently large.

Let us consider three conditions for commutative Jordan algebra, i.e. for all elements $\boldsymbol{A}$ and $\boldsymbol{B}$ of such algebra $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$:

1. $\mathrm{sp}\left\{\boldsymbol{M}\boldsymbol{V}_1\boldsymbol{M}, \ldots, \boldsymbol{M}\boldsymbol{V}_k\boldsymbol{M}\right\}$ is a commutative Jordan algebra,
2. $\mathrm{sp}\left\{\left\{\boldsymbol{M}\boldsymbol{V}_1\boldsymbol{M}, \ldots, \boldsymbol{M}\boldsymbol{V}_k\boldsymbol{M}\right\} \setminus \left\{\boldsymbol{M}\boldsymbol{V}_i\boldsymbol{M}\right\}\right\}$ is a commutative Jordan algebra,
3. $F = \frac{\boldsymbol{y}'\boldsymbol{A}_+\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{A}_-\boldsymbol{y}}$ has F-Snedecor distribution under $H_0 : \sigma_i^2 = 0$.

Theorem 4. *The first and second from the above conditions imply the third condition.*

For proof see Michalski and Zmyślony (1996).

Theorem 5. *The first and third from the above conditions imply the second condition.*

Theorem 6. *Let us assume that a subspace*

$$\mathrm{sp}\left\{\boldsymbol{M}\boldsymbol{V}_1\boldsymbol{M},\ldots,\boldsymbol{M}\boldsymbol{V}_k\boldsymbol{M}\right\}$$

is a commutative Jordan algebra, while

$$\mathrm{sp}\left\{\left\{\boldsymbol{M}\boldsymbol{V}_1\boldsymbol{M},\ldots,\boldsymbol{M}\boldsymbol{V}_k\boldsymbol{M}\right\}\setminus\left\{\boldsymbol{M}\boldsymbol{V}_i\boldsymbol{M}\right\}\right\}$$

is not a commutative Jordan algebra. Then, statistic

$$F=\frac{\boldsymbol{y}'\boldsymbol{A}_+\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{A}_-\boldsymbol{y}}$$

has a generalized F-Snedecor distribution under $H_0:\sigma_i^2=0$, where $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ is BQUE of parameter σ_i^2 (see Fonseca et al. (2002)).

3. Block compound symmetric covariance structure in doubly multivariate data

3.1. Covariance structure

The $(mu\times mu)$–dimensional BCS covariance structure for m-variate observations over u factor levels is defined as:

$$\begin{aligned}\boldsymbol{\Gamma} &= \begin{bmatrix}\boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1 & \ldots & \boldsymbol{\Gamma}_1\\ \vdots & \ddots & & \vdots\\ \vdots & & \ddots & \vdots\\ \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_1 & \ldots & \boldsymbol{\Gamma}_0\end{bmatrix}\\ &= \left(\boldsymbol{\Gamma}_0-\boldsymbol{\Gamma}_1\right)\odot\boldsymbol{I}_u+\boldsymbol{\Gamma}_1\odot\boldsymbol{J}_u\\ &= \boldsymbol{\Gamma}_0\odot\boldsymbol{I}_u+\boldsymbol{\Gamma}_1\odot\left(\boldsymbol{J}_u-\boldsymbol{I}_u\right)\end{aligned}$$

with $\boldsymbol{J}_u=\mathbf{1}_u\mathbf{1}_u'$. The above BCS structure can be also written as a sum of two orthogonal matrices (i.e. the product of orthogonal matrices is equal to matrix $\mathbf{0}$):

$$\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}_0-\boldsymbol{\Gamma}_1)\odot\left(\boldsymbol{I}_u-\frac{1}{u}\boldsymbol{J}_u\right)+\left(\boldsymbol{\Gamma}_0+(u-1)\boldsymbol{\Gamma}_1\right)\odot\frac{1}{u}\boldsymbol{J}_u.$$

The following assumptions for matrices $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}_1$ in BCS structure

1. $\boldsymbol{\Gamma}_0$ is a positive definite symmetric $m\times m$ matrix,
2. $\boldsymbol{\Gamma}_1$ is a symmetric $m\times m$ matrix,
3. $\boldsymbol{\Gamma}_0+(u-1)\boldsymbol{\Gamma}_1$ is a positive definite matrix,

4. $\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_1$ is a positive definite matrix.

imply that the $um \times um$ matrix $\boldsymbol{\Gamma}$ is positive definite (for the proof see Lemma 2.1 in Roy and Leiva (2011)). This result follows also from the property of rank for strong orthogonality of matrices.

3.2. Normal model with BCS covariance structure

The normal BCS model can be written in the following way:

$$\underset{um\times n}{\boldsymbol{Y}} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\right] \sim \mathcal{N}\left((\boldsymbol{I}_{um} \odot \mathbf{1}_n')\boldsymbol{\mu}, \boldsymbol{\Gamma} \odot \boldsymbol{I}_n\right) \tag{9}$$

with $\boldsymbol{\Gamma}$ defined in 3.1. In this model we assume that the mean vector changes over sites or over time points so $\boldsymbol{\mu}$ has um components. Matrix $\boldsymbol{Y}$ contains n independent normally distributed random column vectors, which are identically distributed with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Gamma}$.
Let us consider orthogonal transformation $\boldsymbol{I}_{um} \odot \boldsymbol{Q}$ on $\boldsymbol{Y}_{um\times n}$, where $\boldsymbol{Q}$ is an orthogonal matrix of order n.

Proposition 1. *If* $\operatorname{cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{\Sigma} \odot \boldsymbol{I}$ *with any covariance matrix* $\boldsymbol{\Sigma}$ *then covariance is invariant with respect to transformation* $\boldsymbol{I} \odot \boldsymbol{Q}$ *on* $\boldsymbol{Y}$*.*

In the next proposition we show that orthogonal transformation saves commutativity of projectors with covariance matrices as well as the property of quadratic subspace.

Proposition 2. *Let* $\boldsymbol{\vartheta}_{\boldsymbol{\Sigma}_{\boldsymbol{Y}}}$ *be the space generated by covariance matrices* $\boldsymbol{\Sigma}$ *and let* $\boldsymbol{P}_{\mathrm{E}(\boldsymbol{Y})}$ *denote orthogonal projector onto the subspace of mean matrix of a random matrix* $\boldsymbol{Y}$*. Moreover, let* $\boldsymbol{U} = \boldsymbol{Q}(\boldsymbol{Y})$*, where* $\boldsymbol{Q}$ *is an arbitrary orthogonal operator. Then:*

(i) *If* $\boldsymbol{P}_{\mathrm{E}(\boldsymbol{Y})}\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{P}_{\mathrm{E}(\boldsymbol{Y})}$ *then* $\boldsymbol{P}_{\mathrm{E}(\boldsymbol{U})}\boldsymbol{\Sigma}_{\boldsymbol{U}} = \boldsymbol{\Sigma}_{\boldsymbol{U}}\boldsymbol{P}_{\mathrm{E}(\boldsymbol{U})}$. (10)

(ii) *If* $\boldsymbol{\vartheta}_{\boldsymbol{\Sigma}_{\boldsymbol{Y}}}$ *is a quadratic subspace then* $\boldsymbol{\vartheta}_{\boldsymbol{\Sigma}_{\boldsymbol{U}}}$ *is a quadratic subspace.* (11)

For the special case of $\boldsymbol{Q} = \boldsymbol{Q}_1 \odot \boldsymbol{Q}_2$ we get the following:

Lemma 4. *Since the space* $\boldsymbol{\vartheta}_{\operatorname{cov}(\boldsymbol{Y})}$ *generated by covariance matrices* $\boldsymbol{\Gamma} \odot \boldsymbol{I}$ *is a quadratic subspace and orthogonal projector* $\boldsymbol{P}_{\mathrm{E}(\boldsymbol{Y})} = \boldsymbol{I}_{um} \odot \frac{1}{n}\boldsymbol{J}_n$ *commutes with covariance matrices, we have:*

$\boldsymbol{P}_{\mathrm{E}(\boldsymbol{U})}$ *commutes with* $\operatorname{cov}(\boldsymbol{U})$ *and* $\boldsymbol{\vartheta}_{\operatorname{cov}(\boldsymbol{U})}$ *is a quadratic subspace.*

For the proof that for the model *(9)* $\boldsymbol{\vartheta}_{\operatorname{cov}(\boldsymbol{Y})}$ is a quadratic subspace and assumption that commutativity of $\boldsymbol{P}_{\mathrm{E}(\boldsymbol{Y})}$ holds see Roy et al. (2016).

3.3. Testing hypotheses about structure of expectation

In this section we consider testing hypotheses about the parameters of the mean vector. These results can be also found in Zmyślony et al. (2018). For this reason

we use the two following orthogonal transformations:

1. $\boldsymbol{U} = \left[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\right] = (\boldsymbol{I}_{um} \odot \boldsymbol{Q}_2)\boldsymbol{Y}$, where $\boldsymbol{Q}_2 = \left[\frac{1}{\sqrt{n}}\mathbf{1}_n \vdots \boldsymbol{K}_{\mathbf{1}_n}\right]$ is Helmert matrix, such that $\boldsymbol{K}'_{\mathbf{1}_n}\boldsymbol{K}_{\mathbf{1}_n} = \boldsymbol{I}_{n-1}$ and $\boldsymbol{K}'_{\mathbf{1}_n}\mathbf{1}_n = \mathbf{0}$,

2. $\boldsymbol{W}_i = (\boldsymbol{I} \odot \boldsymbol{Q}_1)\boldsymbol{U}_i$, where $\boldsymbol{U}_i = \mathrm{vec}^{-1}\left(\boldsymbol{u}_i\right)$ is a matrix of size $m \times u$ and $\boldsymbol{Q}_1 = \left[\frac{1}{\sqrt{u}}\mathbf{1}_u \vdots \boldsymbol{K}_{\mathbf{1}_u}\right]$

which are useful for constructing the test statistic.
Now, we formulate the null hypothesis for structure of mean

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_u.$$

This hypothesis can be written equivalently as

$$H_0 : \boldsymbol{\mu}_2^{(c)} = \boldsymbol{\mu}_3^{(c)} = \ldots = \boldsymbol{\mu}_u^{(c)} = 0,$$

where $\boldsymbol{\mu}_j^{(c)} = \sqrt{nu}\sum_{l=1}^{u}\boldsymbol{k}_{l,j-1}\boldsymbol{\mu}_l$, while $\boldsymbol{k}_{l,j-1}$ is $l, j-1$-th element of $\boldsymbol{K}_{\mathbf{1}_u}$.
Following the idea of Michalski and Zmyślony (1999) this hypothesis is equivalent to

$$H_0 : \sum_{j=2}^{u}\boldsymbol{\mu}_j^{(c)}\boldsymbol{\mu}_j^{(c)\prime} = 0.$$

One can prove that quadratic estimator of $\sum_{j=2}^{u}\boldsymbol{\mu}_j^{(c)}\boldsymbol{\mu}_j^{(c)\prime}$ is a function of complete sufficient statistics (see Roy et al. (2016)) and has the following form:

$$\widehat{\sum_{j=2}^{u}\boldsymbol{\mu}_j^{(c)}\boldsymbol{\mu}_j^{(c)\prime}} = \sum_{j=2}^{u}\widehat{\boldsymbol{\mu}}_j^{(c)}\widehat{\boldsymbol{\mu}}_j^{(c)\prime} - (u-1)\widehat{\boldsymbol{\Gamma}}_0 - \widehat{\boldsymbol{\Gamma}}_1, \tag{12}$$

where $\widehat{\boldsymbol{\Gamma}}_0$ and $\widehat{\boldsymbol{\Gamma}}_1$ are best unbiased estimators (BUE) for $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}_1$, respectively. For details see Roy et al. (2016).
Note that

$$\sum_{j=2}^{u}\widehat{\boldsymbol{\mu}}_j^{(c)}\widehat{\boldsymbol{\mu}}_j^{(c)\prime} \overset{\mathrm{df}}{=} (u-1)\widehat{\boldsymbol{\Delta}}_2$$

is the positive part and

$$(u-1)\widehat{\boldsymbol{\Gamma}}_0 - \widehat{\boldsymbol{\Gamma}}_1 \overset{\mathrm{df}}{=} (u-1)\widehat{\boldsymbol{\Delta}}_1$$

is the negative part of estimator in (12).

Under the null hypothesis the positive part has Wishart distribution and the negative part multiplied by $(n-1)$ is Wishart distributed with the same covariance matrix

$\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_1$

$$(n-1)(u-1)\widehat{\boldsymbol{\Delta}}_1 \sim \mathcal{W}_m\left(\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_1, (n-1)(u-1)\right),$$
$$(u-1)\widehat{\boldsymbol{\Delta}}_2 \sim \mathcal{W}_m\left(\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_1, u-1\right),$$

where $\widehat{\boldsymbol{\Delta}}_1$ and $\widehat{\boldsymbol{\Delta}}_2$ are independent.

Lemma 5. *If $\boldsymbol{W}_1 \sim \mathcal{W}_m(\boldsymbol{\Sigma}, n_1)$ and $\boldsymbol{W}_2 \sim \mathcal{W}_m(\boldsymbol{\Sigma}, n_2)$ are independent, then for every fixed vector $\boldsymbol{x} \neq 0 \in \mathbb{R}^m$:*

$$F = \frac{n_2 \boldsymbol{x}' \boldsymbol{W}_1 \boldsymbol{x}}{n_1 \boldsymbol{x}' \boldsymbol{W}_2 \boldsymbol{x}} \sim F_{n_1, n_2}.$$

Now, we give the theorem from Zmyślony et al. (2018).

Theorem 7. *Under the null hypothesis, the statistic*

$$F = \frac{\boldsymbol{x}' \sum_{j=2}^{u} \widehat{\boldsymbol{\mu}}_j^{(c)} \widehat{\boldsymbol{\mu}}_j^{(c)'} \boldsymbol{x}}{(u-1)\boldsymbol{x}' \left(\widehat{\boldsymbol{\Gamma}}_0 - \widehat{\boldsymbol{\Gamma}}_1\right) \boldsymbol{x}} = \frac{\boldsymbol{x}' \widehat{\boldsymbol{\Delta}}_2 \boldsymbol{x}}{\boldsymbol{x}' \widehat{\boldsymbol{\Delta}}_1 \boldsymbol{x}} \tag{13}$$

has F distribution with $(u-1)$ and $(n-1)(u-1)$ degrees of freedom for any fixed $\boldsymbol{x}$.

3.4. Testing hypotheses about $\boldsymbol{\Gamma}_1$

In this section we consider the following hypotheses about parameters in matrix $\boldsymbol{\Gamma}_1$ under assumption that all elements of $\boldsymbol{\Gamma}_1$ are nonnegative or nonpositive:

$$H_0 : \boldsymbol{\Gamma}_1 = \mathbf{0} \;\; vs. \;\; H_1 : \boldsymbol{\Gamma}_1 \neq \mathbf{0}.$$

The presented results can be also found in Fonseca et al. (2018). From Roy et al. (2015) we get that matrices:

$$(n-1)(u-1)\widehat{\boldsymbol{\Delta}}_1 = (n-1)(u-1)(\widehat{\boldsymbol{\Gamma}}_0 - \widehat{\boldsymbol{\Gamma}}_1) \sim \mathcal{W}_m(\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_1, (n-1)(u-1)),$$
$$(n-1)\widehat{\boldsymbol{\Delta}}_2 = (n-1)(\widehat{\boldsymbol{\Gamma}}_0 + (u-1)\widehat{\boldsymbol{\Gamma}}_1) \sim \mathcal{W}_m(\boldsymbol{\Gamma}_0 + (u-1)\boldsymbol{\Gamma}_1, (n-1))$$

are independent. It is easy to show that:

$$\widehat{\boldsymbol{\Gamma}}_1 = \frac{\widehat{\boldsymbol{\Delta}}_2 - \widehat{\boldsymbol{\Delta}}_1}{u}.$$

Under the framework given in Michalski and Zmyślony (1996) a positive part of $\widehat{\boldsymbol{\Gamma}}_1$ is given by:

$$\widehat{\boldsymbol{\Gamma}}_{1+} = \frac{\widehat{\boldsymbol{\Delta}}_2}{u}$$

and a negative part is given by:

$$\widehat{\boldsymbol{\Gamma}}_{1-} = \frac{\widehat{\boldsymbol{\Delta}}_1}{u}.$$

Note that estimator of $\mathbf{\Gamma}_1$ is given by:

$$\widehat{\mathbf{\Gamma}}_1 = \widehat{\mathbf{\Gamma}}_{1+} - \widehat{\mathbf{\Gamma}}_{1-} = \frac{\widehat{\mathbf{\Delta}}_2 - \widehat{\mathbf{\Delta}}_1}{u}.$$

For the proof of the following theorem see Fonseca et al. (2018).

Theorem 8. *Under the hypothesis* $H_0 : \mathbf{\Gamma}_1 = 0$ *the test statistic:*

$$F = \frac{\mathbf{1}'\widehat{\mathbf{\Gamma}}_{1+}\mathbf{1}}{\mathbf{1}'\widehat{\mathbf{\Gamma}}_{1-}\mathbf{1}} \tag{14}$$

has F *distribution with* $(n-1)$ *and* $(n-1)(u-1)$ *degrees of freedom.*

3.5. Testing hypotheses about single parameters

$$H_0 : \sigma_{ij}^{(1)} = 0 \;\; vs. \;\; H_1 : \sigma_{ij}^{(1)} \neq 0$$

In order to conduct F test for testing hypotheses about single parameter, i.e. $H_0 : \sigma_{ii}^{(1)} = 0$ for given $i = 1, \ldots, m$, vectors $\mathbf{1}$ in (14) should be replaced by

$$\boldsymbol{e}_i = (0, \ldots, 0, \underbrace{1}_{i\text{th position}}, 0, \ldots, 0)'.$$

If $\sigma_{ii}^{(1)}$ and $\sigma_{jj}^{(1)}$ are equal to zeros then for parameters $\sigma_{ij}^{(1)}$, $i < j$, $i = 1, \ldots, m$, instead of vectors $\mathbf{1}$ in (14) one should insert

$$\boldsymbol{e}_i - \boldsymbol{e}_j = (0, \ldots, 0, \underbrace{1}_{i\text{th position}}, 0, \ldots, \underbrace{-1}_{j\text{th position}}, 0, \ldots, 0)'.$$

Remark 9. *Testing single contrast of parameters can be done in a similar way using vector* e_i *defined above instead of* $\mathbf{1}_u$.

4. Data application

In this section we use a data set from Johnson and Wichern (2007) for estimation parameters and testing hypotheses, presented in previous section, about the structure of expectation and covariance parameters in model (9). These data contain measures of mineral content of three bones for 25 women: radius, humerus and ulna. Each measurement was recorded on the dominant and non-dominant side.

Using the formula (4.13) and Theorem 1 from Roy et al. (2016) we get that BLUE for $\boldsymbol{\mu}$ is

$$\widehat{\boldsymbol{\mu}} = \begin{bmatrix} 0.84380 & 1.79268 & 0.70440 & 0.81832 & 1.73484 & 0.69384 \end{bmatrix},$$

where, in accordance with the order of variables, the first three values are the means for measurements of mineral content in dominant side of radius, humerus and ulna, respectively, while the last three values are the means for measurements of mineral content in non-dominant side for these bones.
From the same paper, using formulas (3.4) and (3.5) and Theorem 1 we get that BQUE for $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$ are

$$\widehat{\mathbf{\Gamma}}_0 = \begin{bmatrix} 0.01221 & 0.02172 & 0.00901 \\ 0.02172 & 0.07492 & 0.01682 \\ 0.00901 & 0.01682 & 0.01108 \end{bmatrix} \text{ and } \widehat{\mathbf{\Gamma}}_1 = \begin{bmatrix} 0.01038 & 0.01931 & 0.00824 \\ 0.01931 & 0.06678 & 0.01529 \\ 0.00824 & 0.01529 & 0.00807 \end{bmatrix},$$

respectively.

For testing hypotheses about the structure of expectation in test statistic (13) we take the vector $\boldsymbol{x} = \mathbf{1}_m$, so we consider the sum of elements of the positive and negative part of the estimator $\sum_{j=2}^{u} \widehat{\boldsymbol{\mu}_j^{(c)}} \boldsymbol{\mu}_j^{(c)\prime}$. Our test was compared with two well-known tests: likelihood ratio test (LRT) and Roy's test. Formulas of these tests statistics were given in Zmyślony et al. (2018) in Section 4. Calculated p-values for considered data example for all three tests are given in the table below.

Table 1: P-values in testing hypotheses about the structure of expectation and elements of $\mathbf{\Gamma}_1$ with the use of three different tests

Name of test	Test for $\boldsymbol{\mu}$	Test for $\mathbf{\Gamma}_1$
F test	0.0363	$1.06073 \cdot 10^{-9}$
LRT	0.1725	$1.807443 \cdot 10^{-13}$
Roy's test	0.1725	—

The same p-values for LRT and Roy's test in Table 1 follow from the fact that in case $u = 2$ both tests are equivalent. On the standard 5% level of significance we conclude on the p-value for F test that means are significantly different between two sides. For more details about the comparison of these three tests see Zmyślony et al. (2018).

Test F for testing hypotheses about elements of $\mathbf{\Gamma}_1$, whose statistic was given in (14), was compared with LRT, whose statistic was given in formula (3.3) in Fonseca et al. (2018). For both tests, on 5% level of significance, we can conclude that at least one element of $\mathbf{\Gamma}_1$ is different than 0.

5. Conclusion

This paper contains a review of results concerning estimation and testing hypotheses for univariate and multivariate linear models. The presented results are based on the properties of Jordan algebra. Moreover, the coordinate-free approach simplifies inference in linear models for both the univariate and multivariate case. It was presented how the methods of estimation and testing for a univariate model can be

extended to the multivariate case. Estimators of the parameters of the presented BCS covariance structure model and the data presenting measures of mineral content of bones can be found in Roy et al. (2016). The power of the proposed tests for expectation and covariance parameters, in the multivariate case, is compared with well-known tests such as LRT and Roy's test in Fonseca et al. (2018) and Zmyślony et al. (2018). As a result of the simulation study we can say that in some cases (for some alternatives) the tests proposed in this paper have greater power than LRT and Roy's test. The same data example, as for the estimation purpose, was used for testing hypotheses for covariance structure in Fonseca et al. (2018) and for testing hypotheses about the mean structure in Zmyślony et al. (2018).

REFERENCES

DRYGAS, H., (1970). The Coordinate-Free Approach to Gauss-Markov Estimation, Berlin, Heidelberg: Springer.

FARAUT, J., KORÁNYI, A., (1994). Analysis on symmetric cones, Oxford University Press, Oxford.

FONSECA, M., MEXIA, J. T., ZMYŚLONY, R., (2002). Exact distribution for the generalized F tests. Discussiones Mathematicae Probability and Statistics, 22, pp. 37–51.

FONSECA, M., KOZIOŁ, A., ZMYŚLONY, R., (2018). Testing hypotheses of covariance structure in multivariate data. Electronic Journal of Linear Algebra: International Conference on Matrix Analysis and its Applications, MAT TRIAD 2017, 33, pp. 53–62.

JOHNSON, R. A., WICHERN, D. W., (2007). Applied Multivariate Statistical Analysis, New Jersey, Pearson Prentice Hall, sixth ed., Englewood Cliffs.

JORDAN, P., NEUMANN, von, J., WIGNER, E., (1934). On an algebraic generalization of the quantum mechanical formalism. The Annals of Mathematics, 35 (1), pp. 29–64.

KOZIOŁ, A., ROY, A., FONSECA, M., ZMYŚLONY, R., LEIVA, R., (2018). Free-coordinate estimation for doubly multivariate data. Linear Algebra Appl., 547C, pp. 217–239.

KRUSKAL, W., (1968). When are Gauss-Markov and Least Squares Estimators Identical? A Coordinate-Free Approach. The Annals of Mathematical Statistics, 39 (1), pp. 70–75.

LETAC, G., MASSAM, H., (1998). Quadratic and inverse regressions for Wishart distributions, The Annals of Statistics, 26 (2), pp. 573–595.

MASSAM, H., (1994). An exact decomposition theorem and a unified view of some related distributions for a class of exponential transformations on symmetric cones, The Annals of Statistics, 22 (1), pp. 369–394.

MASSAM, H., NEHER, E., (1997). On transformations and determinants of Wishart variables on symmetric cones, Journal of Theoretical Probability, 10, pp. 867–902.

MASSAM, H., NEHER, E., (1998). Estimation and testing for lattice conditional independence models on Euclidean Jordan algebras, The Annals of Statistics, 26 (3), pp. 1051–1082.

MICHALSKI, A., ZMYŚLONY, R., (1996). Testing hypothesis for variance components in mixed linear models. Statistics, 27, pp. 297–310.

MICHALSKI, A., ZMYŚLONY, R., (1999). Testing hypotheses for linear functions of parameters in mixed linear models. Tatra Mountains Mathematical Publications, 17, pp. 103–110.

ROY, A., LEIVA, R., ŽEŽULA, I., KLEIN, D., (2015). Testing the equality of mean vectors for paired doubly multivariate observations in blocked compound symmetric covariance matrix setup. Journal of Multivariate Analysis, 137, pp. 50–60.

ROY, A., ZMYŚLONY, R., FONSECA, M., LEIVA, R., (2016). Optimal estimation for doubly multivariate data in blocked compound symmetric covariance structure, Journal of Multivariate Analysis, 144, pp. 81–90.

SCHAFER, R., (1966). An Introduction to Nonassociative Algebras, Academic Press New York and London.

SEELY, J. F., (1971). Quadratic subspaces and completeness. The Annals of Mathematical Statistics, 42 (2), pp. 710–721.

SEELY, J. F., (1972). Completeness for a family of multivariate normal distributions. The Annals of Mathematical Statistics, 43, pp. 1644–1647.

SEELY, J. F., (1977). Minimal sufficient statistics and completeness for multivariate normal families. Sankhya (Statistics). The Indian Journal of Statistics, 39(2), pp. 170–185.

ZMYŚLONY, R., (1978). A characterization of best linear unbiased estimators in the general linear model, Lecture Notes in Statistics, 2, pp. 365–373.

ZMYŚLONY, R. (1980). Completeness for a family of normal distributions, Mathematical Statistics, Banach Center Publications, 6, pp. 355–357.

ZMYŚLONY, R., ŽEŽULA, I., KOZIOŁ, A., (2018). Application of Jordan Algebra for testing hypotheses about structure of mean vector in model with block compound symmetric covariance structure. Electronic Journal of Linear Algebra: International Conference on Matrix Analysis and its Applications, MAT TRIAD 2017, 33, pp. 41–52.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 155–171, DOI 10.21307/stattrans-2019-020

EXTREME GRADIENT BOOSTING METHOD IN THE PREDICTION OF COMPANY BANKRUPTCY

Barbara Pawełek[1]

ABSTRACT

Machine learning methods are increasingly being used to predict company bankruptcy. Comparative studies carried out on selected methods to determine their suitability for predicting company bankruptcy have demonstrated high levels of prediction accuracy for the extreme gradient boosting method in this area. This method is resistant to outliers and relieves the researcher from the burden of having to provide missing data. The aim of this study is to assess how the elimination of outliers from data sets affects the accuracy of the extreme gradient boosting method in predicting company bankruptcy. The added value of this study is demonstrated by the application of the extreme gradient boosting method in bankruptcy prediction based on data free from the outliers reported for companies which continue to operate as a going concern. The research was conducted using 64 financial ratios for the companies operating in the industrial processing sector in Poland. The research results indicate that it is possible to increase the detection rate for bankrupt companies by eliminating the outliers reported for companies which continue to operate as a going concern from data sets.

Key words: XGBoost, company bankruptcy, machine learning, outlier.

1. Introduction

An important issue in economic and financial decision-making is to predict business failure (bankruptcy prediction, credit scoring) (Nanni and Lumini, 2009). A number of data classification methods are used in company bankruptcy prediction and in credit scoring (Baesens et al., 2003; Lessmann et al., 2015). In the paper by Baesens et al. (2003) the authors evaluate and compare different types of classifiers, for example logistic regression, discriminant analysis, *k*-nearest neighbour, neural networks, decision trees, support vector machines, least-squares support vector machines. Their results suggest that the neural network, least-squares support vector machines, logistic regression and linear discriminant analysis yield a very good performance. The authors of the paper by Lessmann et al. (2015) update the study of Baesens et al. (2003) and compare 41 different classification algorithms such as individual classifiers (Bayesian network, CART, extreme learning machine, kernalized ELM, *k*-nearest neighbour,

[1] Department of Statistics, Cracow University of Economics, Kraków, Poland.
E-mail: barbara.pawelek@uek.krakow.pl. ORCID ID: https://orcid.org/0000-0002-9589-6043.

J4.8, linear discriminant analysis, linear support vector machine, logistic regression, multilayer perceptron artificial neural network, naive Bayes, quadratic discriminant analysis, radial basis function neural network, regularized logistic regression, SVM with radial basis kernel function, voted perceptron), homogenous ensemble classifiers (alternating decision tree, bagged decision trees, bagged MLP, boosted decision trees, logistic model tree, random forest, rotation forest, stochastic gradient boosting), heterogeneous ensemble classifiers (simple average ensemble, weighted average ensemble, stacking, complementary measure, ensemble pruning via reinforcement learning, GASEN, hill-climbing ensemble selection, HCES with bootstrap sampling, matching pursuit optimization ensemble, top-*T* ensemble, clustering using compound error, *k*-means clustering, kappa pruning, margin distance minimization, uncertainty weighted accuracy, probabilistic model for classifier competence, *k*-nearest oracle). Their results suggest that heterogeneous ensembles classifiers perform well.

The main criterion for assessing the suitability of a bankruptcy prediction model (and a credit scoring model) is its prediction ability. In general, performance measures split into three types: the measures that assess the discriminatory ability of the model; the measures that assess the accuracy of the model's probability predictions; the measures that assess the correctness of the model's categorical predictions (Lessmann et al., 2015).

Researchers are seeking to identify sources of the errors committed when predicting company bankruptcy. One of the reasons for misclassification of objects is the heterogeneity of a research data set. Bankruptcy prediction models are developed on the basis of the financial ratios included in financial statements. An analysis of the financial details of the companies which went bankrupt and those which continue to operate as a going concern leads to the conclusion that some of the companies in Poland with unfavourable financial ratios do not go bankrupt (Pawełek et al., 2017). In light of the above considerations, the homogeneity (in terms of financial condition assessment) of the set of companies which continue to operate as a going concern is called into doubt.

Machine learning methods are increasingly used in company bankruptcy prediction (e.g. Brown and Mues, 2012; García et al., 2019; Pawełek, 2017). Comparative studies carried out on selected methods to determine their suitability for predicting company bankruptcy have demonstrated high levels of prediction accuracy for the extreme gradient boosting method (Xia et al., 2017; Zięba et al., 2016). The study by Zięba et al. (2016) adopted a bankruptcy prediction model for the companies operating in the industrial processing sector in Poland over time horizons of one, two, three, four and five years, using a number of machine learning methods (e.g. linear discriminant analysis, multilayer perceptron with a hidden layer, decision rules inducer, decision tree model, logistic regression, boosting algorithm AdaBoost, cost-sensitive boosting algorithm AdaCost, support vector machines, random forest, boosted trees trained with extreme gradient boosting). The databases subject to analysis were not free from outliers and the missing data were not imputed.

This research was undertaken to investigate the combined findings concerning the heterogeneous nature of the set of companies which continue to operate as a going concern in terms of their financial condition and the high

accuracy of the extreme gradient boosting method in predicting company bankruptcy.

The aim of this study is to present the results of our empirical research on the impact that the elimination of outliers from data may have on the accuracy of the extreme gradient boosting method in predicting company bankruptcy. The added value of the study is demonstrated by the proposed application of the extreme gradient boosting method in bankruptcy prediction based on data free from the outliers reported for companies which continue to operate as a going concern.

The study is divided as follows: Section 2 provides a description of the relevant databases and the extreme gradient boosting method; Section 3 outlines the research procedure used; the results of the empirical research are presented and discussed in Section 4; and the main findings of the study are summarised in Section 5.

Table 1. Financial ratios

Ratio	Description	Ratio	Description
W_1	net profit / total assets	W_{33}	operating expenses / short-term liabilities
W_2	total liabilities / total assets	W_{34}	operating expenses / total liabilities
W_3	working capital / total assets	W_{35}	profit on sales / total assets
W_4	current assets / short-term liabilities	W_{36}	total sales / total assets
W_5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	W_{37}	(current assets - inventories) / long-term liabilities
W_6	retained earnings / total assets	W_{38}	constant capital / total assets
W_7	EBIT / total assets	W_{39}	profit on sales / sales
W_8	book value of equity / total liabilities	W_{40}	(current assets - inventory - receivables) / short-term liabilities
W_9	sales / total assets	W_{41}	total liabilities / ((profit on operating activities + depreciation) * (12/365))
W_{10}	equity / total assets	W_{42}	profit on operating activities / sales
W_{11}	(gross profit + extraordinary items + financial expenses) / total assets	W_{43}	rotation receivables + inventory turnover in days
W_{12}	gross profit / short-term liabilities	W_{44}	(receivables * 365) / sales
W_{13}	(gross profit + depreciation) / sales	W_{45}	net profit / inventory
W_{14}	(gross profit + interest) / total assets	W_{46}	(current assets - inventory) / short-term liabilities
W_{15}	(total liabilities * 365) / (gross profit + depreciation)	W_{47}	(inventory * 365) / cost of products sold
W_{16}	(gross profit + depreciation) / total liabilities	W_{48}	EBITDA (profit on operating activities - depreciation) / total assets

Table 1. Financial ratios (cont.)

Ratio	Description	Ratio	Description
W_{17}	total assets / total liabilities	W_{49}	EBITDA (profit on operating activities - depreciation) / sales
W_{18}	gross profit / total assets	W_{50}	current assets / total liabilities
W_{19}	gross profit / sales	W_{51}	short-term liabilities / total assets
W_{20}	(inventory * 365) / sales	W_{52}	(short-term liabilities * 365) / cost of products sold)
W_{21}	sales (n) / sales (n-1)	W_{53}	equity / fixed assets
W_{22}	profit on operating activities / total assets	W_{54}	constant capital / fixed assets
W_{23}	net profit / sales	W_{55}	working capital
W_{24}	gross profit (in 3 years) / total assets	W_{56}	(sales - cost of products sold) / sales
W_{25}	(equity - share capital) / total assets	W_{57}	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
W_{26}	(net profit + depreciation) / total liabilities	W_{58}	total costs / total sales
W_{27}	profit on operating activities / financial expenses	W_{59}	long-term liabilities / equity
W_{28}	working capital / fixed assets	W_{60}	sales / inventory
W_{29}	logarithm of total assets	W_{61}	sales / receivables
W_{30}	(total liabilities - cash) / sales	W_{62}	(short-term liabilities *365) / sales
W_{31}	(gross profit + interest) / sales	W_{63}	sales / short-term liabilities
W_{32}	(current liabilities * 365) / cost of products sold	W_{64}	sales / fixed assets

Source: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data.

2. Data and method

The data used in this study are derived from the Emerging Markets Information Service (https://www.emis.com/pl). The empirical research was carried out using the five databases available at, https://archive.ics.uci.edu/ml/datasets/Polish+companies+ bankruptcy+data.

The research is primarily concerned with the companies operating in the industrial processing sector in Poland. A total of 64 financial ratios were used (Table 1). The research covers the period from 2000 to 2013.

The time horizons of one, two, three, four and five years were used to predict company bankruptcy. A prediction horizon of one or two years is typically adopted in the literature on bankruptcy prediction. However, in some cases the three-year time horizon is used. A time horizon of four or five years is also possible but is

rarely applied; this is due to the dynamic nature of the immediate or more distant business environment. An example of such research is the work by Zięba et al. (2016), which adopted a time horizon of one to five years. Due to the fact that the above work has inspired us to undertake this study, we have decided to analyse all five databases (i.e. five prediction horizons).

Machine learning methods are used in a number of research areas (e.g. Friedman et al., 2000). Two major factors determining the suitability of a machine learning method for predicting various developments are: the application of statistical methods for detecting and modelling the existing links between complex phenomena and the use of calculation algorithms designed for large data sets. One example of the machine learning method is gradient tree boosting (GTB) (Friedman, 2001). The gradient tree boosting method is also known as gradient boosting machine (GBM) or gradient boosted regression tree (GBRT). For the purposes of this research, we have adopted the extreme gradient boosting (XGBoost) method (Chen and Guestrin, 2016). XGBoost is a GTB algorithm, which is particularly useful in analysing large data sets. The innovative nature of the XGBoost method - in comparison with other tree boosting algorithms - lies in its use of a novel sparsity-aware algorithm for parallel tree learning.

The necessary calculations for the research work were made using the R software and the 'xgboost' package (Chen and Guestrin, 2016).

3. Research method description

The first step of our empirical research was to split the input data set into a training set and a test set. To ensure generalisability of the research findings, the splitting operation was repeated 30 times. The next step was to remove the outliers reported for companies which continue to operate as a going concern from the training set. The outliers were removed using quantiles (Pawełek et al., 2017; Wu et al., 2010). Then, the extreme gradient boosting method was applied, followed by calculations concerning two model selection criteria to compare the results obtained. The *error* criterion is used to verify the share of wrongly classified objects in the total number of objects, while the *AUC* criterion stands for *Area under the ROC Curve*. Once developed, the models were assessed in terms of their prediction accuracy (*Accuracy* – the share of correctly classified companies in the total number of objects, *Sensitivity* – the share of correctly classified bankrupts in the total number of bankrupts, *Specificity* – the share of correctly classified non-bankrupts in the total number of non-bankrupts). The final step was to verify the hypothesis concerning the location parameters for the populations from which the prediction accuracy measure concerned had been derived, as calculated for models developed based on outlier-free training sets.

Phases of research:

1) Random division of the set X^h, where $h = 1, 2, 3, 4, 5$, into the training set U^h and the test set T^h ($X^h = U^h \cup T^h$, where $\overline{\overline{U^h}} = \frac{2}{3}\overline{\overline{X^h}}$ and $\overline{\overline{T^h}} = \frac{1}{3}\overline{\overline{X^h}}$), while preserving the current structure to take account of bankrupt (B) and non-bankrupt (NB) companies. The splitting operation was repeated 30 times.

2) Providing four variants for each set U^h $(h = 1, 2, 3, 4, 5)$ using the quantiles Q_q and Q_{1-q}, where:

   a) $q = 0.00$ (i.e. outliers are not removed from the set),

   b) $q = 0.01$,

   c) $q = 0.05$,

   d) $q = 0.10$.

   If:

   $U^h = U^{h,B} \cup U^{h,NB}$ $(h = 1, 2, 3, 4, 5)$,

   where:

   $U^{h,B} = \{u_i^{h,B}: u_i^{h,B} \in U^h$ i $u_i^{h,B}$ is bankrupt$\}$,

   $U^{h,NB} = \{u_i^{h,NB}: u_i^{h,NB} \in U^h$ i $u_i^{h,NB}$ is not bankrupt$\}$.

   The following conversion formula is used:

$$\forall j = 1, 2, \dots, 64: u_{ij}^{h,NB} = \begin{cases} u_{ij}^{h,NB} \text{ if } Q_q^j \le u_{ij}^{h,NB} \le Q_{1-q}^j \\ Q_q^j \quad \text{if} \quad u_{ij}^{h,NB} < Q_q^j, \\ Q_{1-q}^j \text{ if} \quad Q_{1-q}^j < u_{ij}^{h,NB} \end{cases}$$

   the result of which is:

   $U_{0.00}^h$, $U_{0.01}^h$, $U_{0.05}^h$, $U_{0.10}^h$ $(h = 1, 2, 3, 4, 5)$.

3) Developing two models per training set U_q^h $(q = 0.00, 0.01, 0.05, 0.10; h = 1, 2, 3, 4, 5)$:

   – establishing the number of iterations (from 1 to 100) against the two criteria: *error* and *AUC*, which are used to measure the classification accuracy of the model based on the training set by means of cross-validation $(v = 3)$; the result is $M_q^{h,error}$ and $M_q^{h,auc}$ $(q = 0.00, 0.01, 0.05, 0.10;\ h = 1, 2, 3, 4, 5)$; Figures 1 and 2 show trends in *error* and *AUC* criteria values for the sample training set $U_{0.00}^1$, while Figures 3-5 show trends in *Accuracy*, *Sensitivity* and *Specificity* values for the sample test set $T_{0.00}^1$ (case of the *error* criterion);

- assessing the prediction accuracy of the developed models on the basis of the test set T^h: $M_q^{h,error}(T^h)$ and $M_q^{h,auc}(T^h)$ $(q = 0.00, 0.01, 0.05, 0.10; h = 1, 2, 3, 4, 5)$, using the following measures: *Accuracy*, *Sensitivity* and *Specificity*.

4) Verifying the hypothesis that the prediction accuracy measure concerned (*Accuracy*, *Sensitivity* or *Specificity*) – as calculated for models developed on the basis of outlier-free training sets $M_q^{h,error}(T^h)$ and $M_q^{h,auc}(T^h)$ $(q = 0.00, 0.01, 0.05, 0.10;\ h = 1, 2, 3, 4, 5)$ – originated from the populations with the same location parameters.

   We used the following tests:

   - the Kruskal-Wallis test,
   - Dunn's post hoc test (the version including the Bonferroni correction for multiple testing).

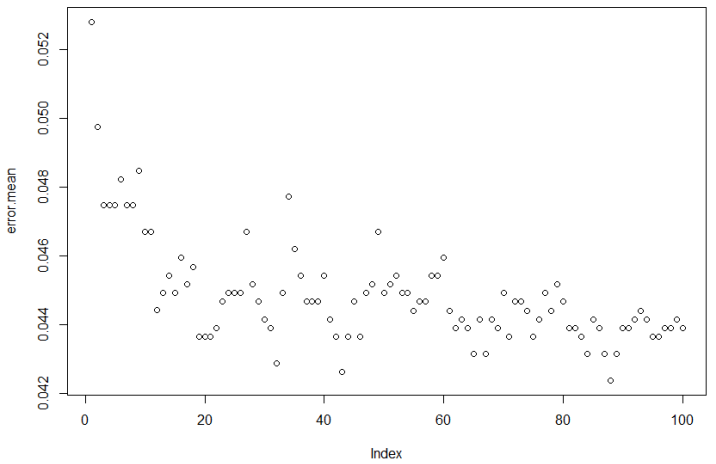

Figure 1. *Error* criteria values for training set $\mathrm{U}_{0.00}^{1}$

Source: Own work.

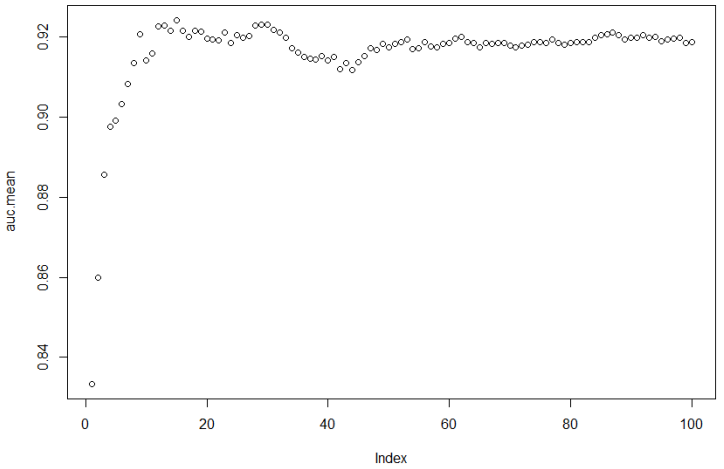

Figure 2. *AUC* criteria values for training set $U_{0.00}^{1}$
Source: Own work.

Figure 1 indicates that the rise in the number of iterations is accompanied by a corresponding decrease in *error* values, which fluctuate around 0.044, whereas the *AUC* criterion value (Figure 2) - initially on the rise - is finally stabilised around 0.92.

As regards the *error* criterion, we selected a model with the lowest share of wrongly classified objects in the total number of objects. As regards the *AUC* criterion, we selected a model with the largest area under the ROC curve.

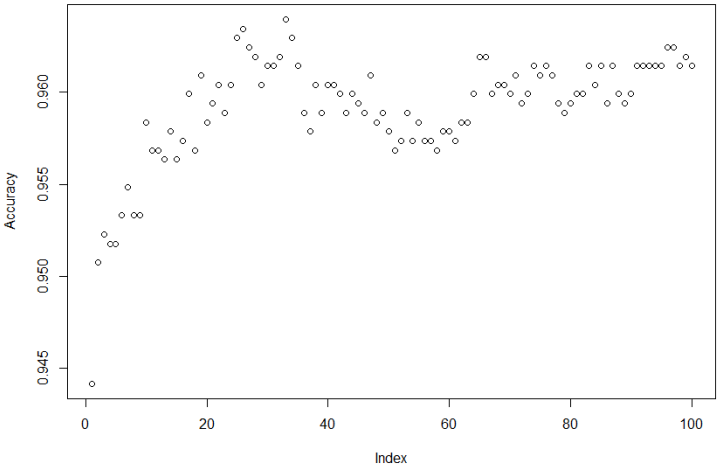

Figure 3. *Accuracy* measure for test set $T_{0.00}^{1}$ (case of the *error* criterion)
Source: Own work.

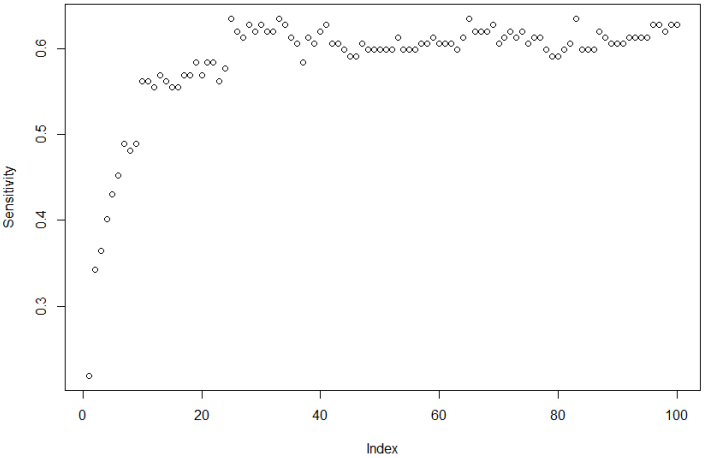

Figure 4. *Sensitivity* measure for test set $T_{0.00}^{1}$ (case of the *error* criterion)
Source: Own work.

The process of model selection may also be based on such prediction accuracy measures as *Accuracy*, *Sensitivity* and *Specificity*. Figures 3-5 show changes in the value of these measures for the selected test set $T_{0.00}^{1}$ by number of iterations (case of the *error* criterion). An analysis of the charts indicates that the *Accuracy* and *Specificity* measures for the test set under consideration stabilised around 0.960 and 0.987 after approx. 70 iterations, whereas the *Sensitivity* measure values fluctuated around 0.60 after approx. 40 iterations.

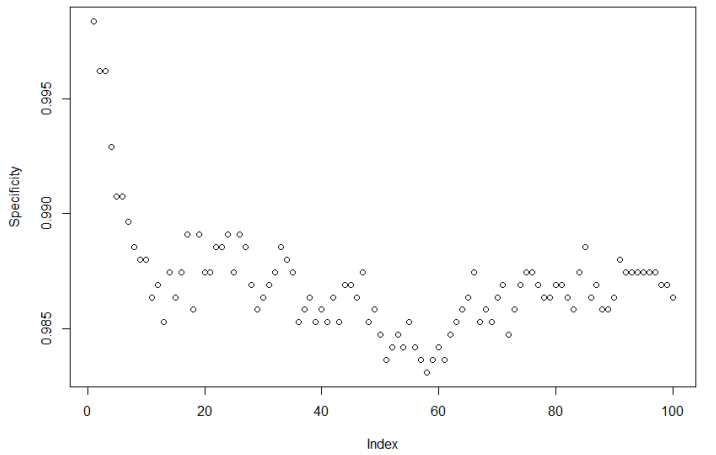

Figure 5. *Specificity* measure for test set $T_{0.00}^{1}$ (case of the *error* criterion)
Source: Own work.

The following section of this document describes the results of empirical research for the two criteria: *error* and *AUC*. For further research purposes, other criteria may also be used for comparison purposes.

4. Empirical results

Table 2 contains results of the Dunn-Bonferroni post hoc test ($\alpha = 0.05$) carried out on the pairs of research approaches based on data free from outliers which have been removed using the quantiles $\{Q_{q_1}, Q_{1-q_1}\}$ and $\{Q_{q_2}, Q_{1-q_2}\}$, where $q_1, q_2 = 0.00, 0.01, 0.05, 0.10$. Due to the fact that the results obtained for all prediction horizons under consideration ($h = 1, 2, 3, 4, 5$) are the same, they were put in a single table.

Table 2. Results of the Dunn-Bonferroni post hoc test ($\alpha = 0.05$) carried out on the pairs of research approached based on data free from outliers which have been removed using the quantiles $\{Q_{q_1}, Q_{1-q_1}\}$ and $\{Q_{q_2}, Q_{1-q_2}\}$ for $h = 1, 2, 3, 4, 5$

	q_1:q_2					
Measure	**0.00:0.01**	**0.00:0.05**	**0.00:0.10**	**0.01:0.05**	**0.01:0.10**	**0.05:0.10**
Criterion: *error*						
Accuracy	S	S	S	S	S	S
Sensitivity	NS	S	S	S	S	S
Specificity	S	S	S	S	S	S
Criterion: *AUC*						
Accuracy	S	S	S	S	S	S
Sensitivity	NS	S	S	S	S	S
Specificity	S	S	S	S	S	S

Significant (S) or non-significant (NS) at $\alpha = 0.05$.

Source: Own work.

At a significant level of 0.05, the test results obtained in most cases under consideration indicate that there is a statistically significant difference between the location parameters for the populations from which relevant prediction accuracy measures are derived, obtained as a result of different variants having been adopted for the training sets $U^h_{0.00}$, $U^h_{0.01}$, $U^h_{0.05}$, $U^h_{0.10}$ ($h = 1, 2, 3, 4, 5$). The results obtained can be interpreted to mean that the proposed removal of outliers from the data sets has a statistically significant impact on the accuracy of the XGBoost method in predicting company bankruptcy.

An exception to this rule is the *Sensitivity* measure for the pairs of research approaches based on data free from outliers which have been removed using the quantiles $\{Q_{0.00}, Q_{1.00}\}$ and $\{Q_{0.01}, Q_{0.99}\}$. In this case (significant level of 0.05), the removal of outliers from data in the training set has not affected the accuracy of the XGBoost prediction method in a statistically significant way.

To verify that changes in the prediction accuracy of the XGBoost method, occurring as a result of outliers being removed from data sets, constitute a positive trend in bankruptcy prediction, Table 3 provides aggregate information on the arithmetic mean, standard deviation and median values for such measures as *Accuracy*, *Sensitivity* and *Specificity*. For example, Figures 6-8 show results assuming the one-year prediction horizon (prior to bankruptcy).

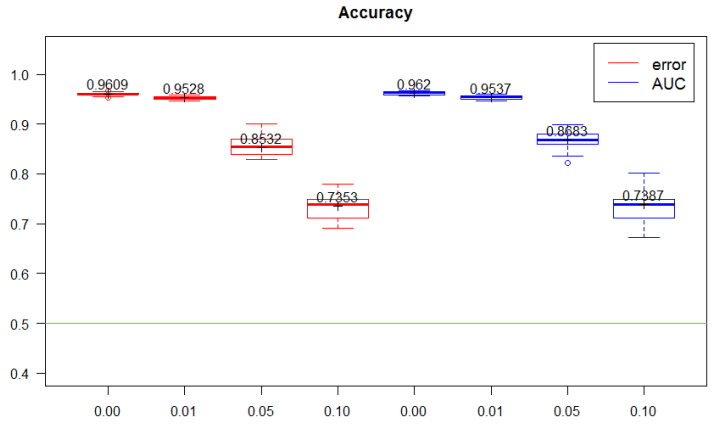

Figure 6. *Accuracy* values obtained assuming the one-year prediction horizon (prior to bankruptcy)

Source: Own work.

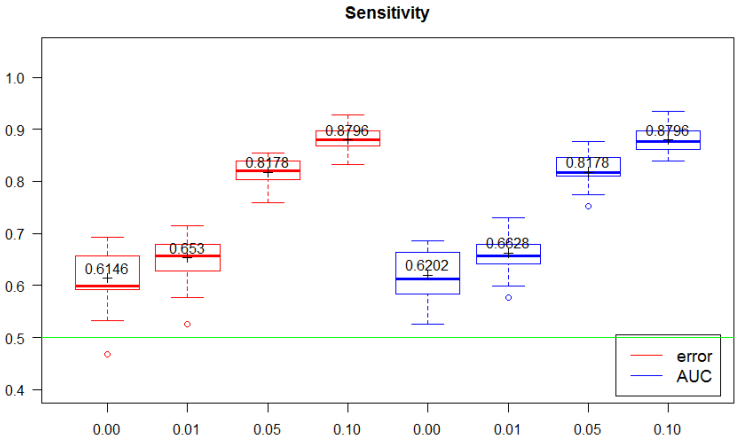

Figure 7. *Sensitivity* values obtained assuming the one-year prediction horizon (prior to bankruptcy)

Source: Own work.

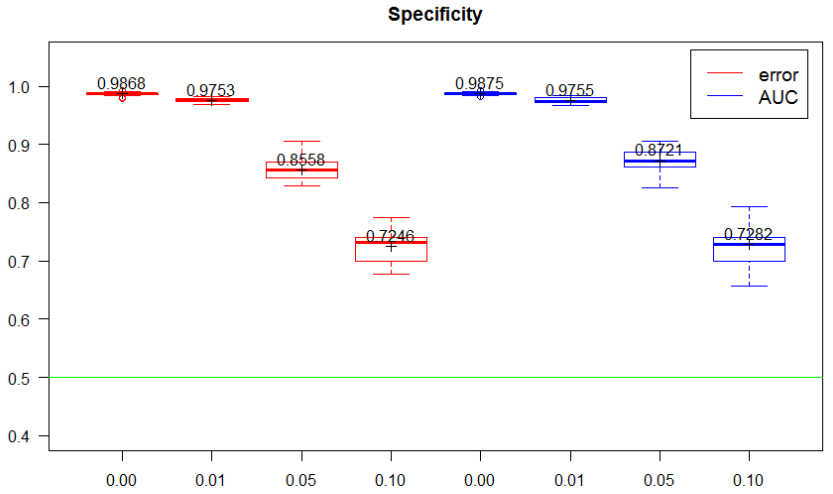

Figure 8. *Specificity* values obtained assuming the one-year prediction horizon (prior to bankruptcy)

Source: Own work.

Figures 6-8 indicate that for the one-year prediction horizon an increase in *q* parameter value (which indicates the level of quantiles used to eliminate outliers from data) is accompanied by corresponding decreases in *Accuracy* and *Specificity* values and increases in *Sensitivity* values. The observed increase in the average value of the *Sensitivity* measure, i.e. from 0.6146 $(q = 0.00)$ to 0.8796 $(q = 0.10)$ (case of the *error* criterion) and from 0.6202 $(q = 0.00)$ to 0.8796 $(q = 0.10)$ (case of the *AUC* criterion), is a positive development, whereas the reported decrease in *Specificity* values, i.e. from 0.9868 $(q = 0.00)$ to 0.7246 $(q = 0.10)$ (case of the *error* criterion) and from 0.9875 $(q = 0.00)$ to 0.7282 $(q = 0.10)$ (case of the *AUC* criterion), is a negative development. From the perspective of credit granting institutions, which commission bankruptcy prediction models, the accuracy of bankruptcy risk prediction is more important than the prediction accuracy for absence of bankruptcy risks. Given the above and the results obtained, the use of quantiles $\{Q_{0.05}, Q_{0.95}\}$ during removal of outliers from the data sets is the most preferred option. The *Sensitivity* measure rose to 0.8178 (case of the *error* criterion) and 0.8178 (case of the *AUC* criterion), while the *Specificity* measure fell only slightly, i.e. to 0.8558 (case of the *error* criterion) and 0.8721 (case of the *AUC* criterion).

The results presented in Tables 3-7 show that the patterns observed for the prediction model with the one-year time horizon can also be identified for the results obtained using the other four databases. The trends reported for the arithmetic mean values of the prediction accuracy measures are also identified for the median values.

Table 3. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures (h = 1)

		Criterion: *error*			Criterion: *AUC*		
Measure	***q***	**Average**	**Standard deviation**	**Median**	**Average**	**Standard deviation**	**Median**
Accuracy	0.00	0.9609	0.0032	0.9609	0.9620	0.0032	0.9629
	0.01	0.9528	0.0036	0.9530	0.9537	0.0034	0.9548
	0.05	0.8532	0.0173	0.8541	0.8683	0.0165	0.8680
	0.10	0.7353	0.0208	0.7381	0.7387	0.0315	0.7391
Sensitivity	0.00	0.6146	0.0487	0.5985	0.6202	0.0431	0.6131
	0.01	0.6530	0.0426	0.6569	0.6628	0.0365	0.6569
	0.05	0.8178	0.0279	0.8212	0.8178	0.0286	0.8175
	0.10	0.8796	0.0265	0.8796	0.8796	0.0245	0.8759
Specificity	0.00	0.9868	0.0026	0.9869	0.9875	0.0023	0.9875
	0.01	0.9753	0.0036	0.9749	0.9755	0.0045	0.9733
	0.05	0.8558	0.0188	0.8560	0.8721	0.0176	0.8723
	0.10	0.7246	0.0226	0.7310	0.7282	0.0332	0.7278

Source: Own work.

Table 4. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures (h = 2)

		Criterion: *error*			Criterion: *AUC*		
Measure	***q***	**Average**	**Standard deviation**	**Median**	**Average**	**Standard deviation**	**Median**
Accuracy	0.00	0.9645	0.0020	0.9645	0.9641	0.0023	0.9635
	0.01	0.9465	0.0064	0.9459	0.9490	0.0050	0.9505
	0.05	0.7748	0.0241	0.7713	0.7882	0.0202	0.7874
	0.10	0.6175	0.0251	0.6131	0.6269	0.0331	0.6176
Sensitivity	0.00	0.5060	0.0389	0.5000	0.5062	0.0372	0.5058
	0.01	0.5554	0.0507	0.5698	0.5665	0.0433	0.5901
	0.05	0.7556	0.0630	0.7791	0.7653	0.0439	0.7791
	0.10	0.8762	0.0279	0.8779	0.8624	0.0247	0.8605
Specificity	0.00	0.9900	0.0017	0.9897	0.9896	0.0017	0.9897
	0.01	0.9682	0.0055	0.9683	0.9703	0.0045	0.9702
	0.05	0.7758	0.0266	0.7705	0.7895	0.0218	0.7908
	0.10	0.6031	0.0266	0.5983	0.6138	0.0353	0.6025

Source: Own work.

Table 5. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures (h = 3)

		Criterion: *error*			Criterion: *AUC*		
Measure	***q***	**Average**	**Standard deviation**	**Median**	**Average**	**Standard deviation**	**Median**
Accuracy	0.00	0.9642	0.0017	0.9640	0.9645	0.0028	0.9646
	0.01	0.9512	0.0042	0.9514	0.9529	0.0027	0.9520
	0.05	0.7675	0.0240	0.7775	0.7876	0.0210	0.7832
	0.10	0.5648	0.0199	0.5701	0.5879	0.0361	0.5920
Sensitivity	0.00	0.3582	0.0978	0.3758	0.4154	0.0439	0.4182
	0.01	0.4317	0.0379	0.4303	0.4564	0.0555	0.4303
	0.05	0.6919	0.0165	0.6909	0.6846	0.0193	0.6848
	0.10	0.8220	0.0204	0.8182	0.8265	0.0187	0.8242
Specificity	0.00	0.9941	0.0038	0.9922	0.9917	0.0021	0.9912
	0.01	0.9768	0.0051	0.9769	0.9775	0.0024	0.9781
	0.05	0.7713	0.0255	0.7830	0.7927	0.0221	0.7891
	0.10	0.5520	0.0205	0.5579	0.5761	0.0381	0.5808

Source: Own work.

Table 6. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures (h = 4)

		Criterion: *error*			Criterion: *AUC*		
Measure	***q***	**Average**	**Standard deviation**	**Median**	**Average**	**Standard deviation**	**Median**
Accuracy	0.00	0.9730	0.0019	0.9732	0.9727	0.0021	0.9729
	0.01	0.9643	0.0031	0.9653	0.9647	0.0033	0.9655
	0.05	0.7810	0.0276	0.7846	0.8069	0.0311	0.8178
	0.10	0.5527	0.0306	0.5520	0.5789	0.0478	0.5619
Sensitivity	0.00	0.4341	0.0277	0.4361	0.4368	0.0289	0.4361
	0.01	0.4476	0.0291	0.4511	0.4574	0.0254	0.4586
	0.05	0.6629	0.0536	0.6692	0.6634	0.0463	0.6692
	0.10	0.8253	0.0338	0.8346	0.8183	0.0350	0.8158
Specificity	0.00	0.9950	0.0015	0.9952	0.9946	0.0016	0.9945
	0.01	0.9854	0.0032	0.9865	0.9855	0.0036	0.9863
	0.05	0.7858	0.0293	0.7893	0.8128	0.0325	0.8241
	0.10	0.5416	0.0319	0.5428	0.5691	0.0502	0.5525

Source: Own work.

Table 7. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures ($h = 5$)

Measure	q	Criterion: *error*			Criterion: *AUC*		
		Average	Standard deviation	Median	Average	Standard deviation	Median
Accuracy	0.00	0.9793	0.0014	0.9791	0.9789	0.0012	0.9795
	0.01	0.9730	0.0021	0.9718	0.9736	0.0021	0.9735
	0.05	0.8898	0.0177	0.8915	0.9018	0.0171	0.9065
	0.10	0.7350	0.0234	0.7400	0.7476	0.0191	0.7447
Sensitivity	0.00	0.5411	0.0465	0.5556	0.5426	0.0392	0.5556
	0.01	0.5715	0.0136	0.5667	0.5707	0.0169	0.5778
	0.05	0.6756	0.0520	0.6889	0.6711	0.0421	0.6778
	0.10	0.7967	0.0349	0.8222	0.7911	0.0402	0.8000
Specificity	0.00	0.9968	<0.0001	0.9969	0.9964	<0.0001	0.9964
	0.01	0.9890	0.0021	0.9880	0.9897	0.0019	0.9893
	0.05	0.8984	0.0188	0.9028	0.9111	0.0168	0.9156
	0.10	0.7326	0.0249	0.7393	0.7459	0.0211	0.7411

Source: Own work.

In the case of standard deviation values, no pattern that would be common to all cases under consideration has been observed. It is often the case that an increase in q parameter values is accompanied by corresponding increases in standard deviation values for *Accuracy* and *Specificity* measures. Further research is needed to assess how the elimination of outliers from data sets could affect measures other than the arithmetic mean and median of the data set.

5. Conclusions

The above considerations can be summed up by stating that in most cases where the significant level was set at 0.05, an analysis of the prediction accuracy measures resulted in the null hypothesis being rejected in favour of the alternative hypothesis stating that the resulting data sets were not derived from the populations with the same location parameters. An exception to this rule is the pair of research approaches based on the training sets free from outliers which have been removed using the quantiles $\{Q_{0.00}, Q_{1.00}\}$ and $\{Q_{0.01}, Q_{0.99}\}$ for the *Sensitivity* measure.

It can therefore be concluded that the removal of the outliers reported for companies which continue to operate as a going concern from data sets affects the accuracy of the extreme gradient boosting method in predicting company bankruptcy.

The results show that the use of quantiles for the removal of the outliers reported for companies which continue to operate as a going concern from training sets increases the accuracy of the extreme gradient boosting method in detecting bankrupt companies (*Sensitivity*), while reducing the prediction

accuracy of that method when measured as total (*Accuracy*) and for a group of non-bankrupt companies (*Specificity*). In addition, the following pattern was observed: the more the accuracy is affected, the higher the *q* parameter $(Q_q \text{ and } Q_{1-q})$.

The results of the empirical research are consistent with the statement that the longer the prediction horizon (*h*), the less accurate the bankruptcy detection model (*Sensitivity*).

Among the quantiles examined, the pair $Q_{0.05}$ and $Q_{0.95}$ should be highlighted due to (among others) the fact that when it was used to remove the outliers reported for companies which continue to operate as a going concern from the training set, the average value of the *Sensitivity* measure for *h* = 3, 4 rose above 0.50.

Acknowledgements

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

REFERENCES

BAESENS, B., VAN GESTEL, T., VIAENE, S., STEPANOVA, M., SUYKENS, J., VANTHIE, J., (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54, pp. 627–635. DOI: http://dx.doi.org/10.1057/palgrave.jors.2601545.

BROWN, I., MUES, Ch., (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39 (3), pp. 3446–3453. DOI: https://doi.org/10.1016/j.eswa.2011.09.033.

CHEN, T., GUESTRIN, C., (2016). XGBoost: A Scalable Tree Boosting System. DOI: http://dx.doi.org/10.1145/2939672.2939785.

GARCIA, V., MARQUES, A. I., SANCHEZ, J. S., (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. Information Fusion, 47, pp. 88–101. DOI: https://doi.org/10.1016/j.inffus.2018.07.004.

FRIEDMAN, J. H., (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29 (5), pp. 1189–1232.

FRIEDMAN, J. H., HASTIE, T., TIBSHIRANI, R., (2000). Additive logistic regression: a statistical view of boosting. The Annals of Statistics, 28 (2), pp. 337–407.

LESSMANN, S., BAESENS, B., SEOW, H. V., THOMAS, L. C., (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. European Journal of Operational Research, 247 (1), pp. 124–136. DOI: https://doi.org/10.1016/j.ejor.2015.05.030.

NANNI, L., LUMINI, A., (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications, 36, pp. 3028–3033.
DOI: http://dx.doi.org/10.1016/j.eswa.2008.01.018.

PAWEŁEK, B., GAŁUSZKA, K., KOSTRZEWSKA, J., KOSTRZEWSKI, M., (2017). Classification Methods in the Research on the Financial Standing of Construction Enterprises After Bankruptcy in Poland. In: Palumbo, F. et al. (Eds.), Data Science, Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Switzerland, pp. 29–42.
DOI: http://dx.doi.org/10.1007/978-3-319-55723-6_3.

PAWEŁEK, B., (2017). Prediction of Company Bankruptcy in the Context of Changes in the Economic Situation. In: Papież, M., Śmiech, S. (Eds.), The 10th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings. Cracow: Foundation of the Cracow University of Economics, pp. 290–299.

WU, Y., GAUNT, C., GRAY, S., (2010). A comparison of alternative bankruptcy prediction models. Journal of Contemporary Accounting & Economics, 6, pp. 34–45. DOI: http://dx.doi.org/10.1016/j.jcae.2010.04.002.

XIA, Y., LIU, Ch., LI, Y., LIU, N., (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems With Applications, 78, pp. 225–241.

ZIĘBA, M., TOMCZAK, S. K., TOMCZAK, J. M., (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems With Applications, 58, pp. 93–101.
DOI: http://dx.doi.org/10.1016/j.eswa.2016.04.001.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 173–185, DOI 10.21307/stattrans-2019-021

EFFICIENT TWO-PARAMETER ESTIMATOR IN LINEAR REGRESSION MODEL

Ashok V. Dorugade[1]

ABSTRACT

In this article, two-parameter estimators in linear model with multicollinearity are considered. An alternative efficient two-parameter estimator is proposed and its properties are examined. Furthermore, this was compared with the ordinary least squares (OLS) estimator and ordinary ridge regression (ORR) estimators. Also, using the mean squares error criterion the proposed estimator performs more efficiently than OLS estimator, ORR estimator and other reviewed two-parameter estimators. A numerical example and simulation study are finally conducted to illustrate the superiority of the proposed estimator.

Key words: multicollinearity, ridge regression, two-parameter estimator, mean squared error.

1. Introduction

The ordinary least squares (OLS) method is one of the most important ways for estimating the parameters of the general linear model. Because of its simplicity and rationality, the results are obtained when specific assumptions are achieved. But if these assumptions are violated, OLS method does not assure the desirable results. Multicollinearity occurs when two or more than two explanatory variables are correlated with each other. To solve this problem, various biased estimators were put forward in the literature. The ordinary ridge regression (ORR) proposed by Hoerl and Kennard (1970a) is the most popular biased estimator. However, ORR estimator has some disadvantages; mainly it is a nonlinear function of the ridge parameter k. This leads to complicated equations when selecting k. To solve such difficulty, Liu (1993) then proposed the estimator called Liu estimator (LE). As seen, LE is a linear function of the ridge parameter d and thus it is more convenient to choose d than k. Liu (2003) suggested two Liu-type estimators and proved that these estimators have some superior properties over RR estimator under the mean squared error (MSE) criterion. However, it is difficult to determine which is better between them.

Some other popular numerical techniques to deal with multicollinearity are the ridge regression due to Singh and Chaubey (1987), Liu (1993), Akdeniz and

[1] Y C Mahavidyalaya Halkarni, Tal-Chandgad, Kolhapur, Maharashtra India – 416552.
E-mail: adorugade@rediffmail.com.

Kaciranlar (1995), Crouse *et al.,* (1995), Kaciranlar *et al.*, (1999), Ozkale and Kaciranlar (2007), Yang and Chang (2010), Wu and Yang (2011), Dorugade (2014) and others.

In this paper, a new method for estimating the parameters in linear regression model with multicollinearity problem is proposed. The rest of this paper is organized as follows. The model and some well-known estimators are reviewed in Section 2. The efficient two parameter estimator is introduced in Section 3. Performances of the proposed estimator with respect to the scalar MSE criterion are discussed in Section 4. In Section 5, the methods of choosing the parameters were discussed. A simulation study to justify the superiority of the suggested estimator is given in Section 6. Some concluding remarks are given in Section 7.

2. Model and estimators

Consider the linear regression model

$$Y = X\beta + \varepsilon, \tag{1}$$

where *Y* is a n×1 random vector of response variables, ***X*** is a known n×p matrix with full column rank, ε is the vector of errors E(ε) = 0 and Cov(ε) = $\sigma^2 I_n$. , β is a p×1 vector of unknown regression parameters and σ^2 is the unknown variance parameter. For the sake of convenience, it was assumed that the matrix *X* and the response variable *Y* are standardized in such a way that $X'X$ is a non-singular correlation matrix and $X'Y$ is the correlation between *X* and *Y*.

Let $\wedge$ and T be the matrices of eigen values and eigen vectors of $X'X$, respectively, satisfying $T'X'XT = \wedge$ = diagonal ($\lambda_1, \lambda_2, ..., \lambda_p$), with λ_i being the i[th] eigen value of $X'X$ and $T'T = TT' = I_p$ We obtain the equivalent model

$$Y = Z\alpha + \varepsilon, \tag{2}$$

where Z = *X*T , it implies that $Z'Z = \wedge$, and α = $T'\beta$ (see Montgomery *et al*., 2006).

Then, OLS estimator of α is given by

$$\hat{\alpha}_{OLS} = (Z'Z)^{-1}Z'Y = \wedge^{-1} Z'Y. \tag{3}$$

Therefore, LS estimator of β is given by

$$\hat{\beta}_{OLS} = T\hat{\alpha}_{OLS}$$

However, it is well-known that OLS estimator performs poorly when multicollinearity exists. In order to control the instability in least squares estimates, Hoerl (1962), Hoerl and Kennard (1968) and then Hoerl and Kennard (1970b) suggested an alternative estimate of the regression coefficients namely ridge

regression as obtained by adding a positive constant (or ridge parameter) k to the diagonal elements of the ordinary least square estimator. It is given as:

$$\hat{\alpha}_{ORR} = \left[I - k\left(\wedge + kI\right)^{-1}\right]\hat{\alpha}_{OLS} \tag{4}$$

Therefore, ORR estimator of β is given by

$$\hat{\beta}_{ORR} = T\hat{\alpha}_{ORR}$$

The literature has shown that some ridge estimators are based on a single ridge parameter while some are based on two ridge parameters. Some of the well-known methods used for estimation are listed below.

The Jackknifed ridge regression estimator introduced by Singh and Chaubey (1987) is defined by

$$\hat{\alpha}_{JRR} = \left[I - k^2\left(\wedge + kI\right)^{-2}\right]\hat{\alpha}_{OLS} \tag{5}$$

Liu (1993) introduced a biased estimator, which is defined by

$$\hat{\alpha}_{Liu} = \left(\wedge + I\right)^{-1}\left(\wedge + dI\right)\hat{\alpha}_{OLS} \tag{6}$$

The almost unbiased Liu estimator introduced by Akdeniz and Kaciranlar (1995) is defined by

$$\hat{\alpha}_{AUL} = \left[I - \left(\wedge + I\right)^{-2} + (1-d)^2\right]\wedge^{-1} X'Y\,. \tag{7}$$

Crouse et al., (1995) defined unbiased ridge estimator given by

$$\hat{\alpha}_{URR} = \left(\wedge + kI\right)^{-1}\left(Z'Y + kJ\right) \;\; where \quad J = \sum_{i=1}^{p}\hat{\alpha}_i / p\,. \tag{8}$$

Ozkale and Kaciranlar (2007) introduced a two-parameter estimator, which is defined by

$$\hat{\alpha}_{TP} = \left(\wedge + kI\right)^{-1}\left(\wedge + kdI\right)\hat{\alpha}_{OLS}\,. \tag{9}$$

The ridge parameter $k = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p}\hat{\alpha}_i^2$ given by Hoerl et al. (1975) performs fairly well and the well-known estimate of 'd'proposed by Liu (1993) is given as:

$$d = \sum_{i=1}^{p}(\hat{\alpha}_i^2 - \hat{\sigma}^2)/(\lambda_i + 1)^2 \Big/ \sum_{i=1}^{p}(\hat{\sigma}^2 + \lambda_i\hat{\alpha}_i^2)/(\lambda_i + 1)^2\lambda_i\,.$$

The above calculated values of k and d are used to determine estimators given in (5) to (9), where, $\hat{\alpha}_i$ is the i[th] element of $\hat{\alpha}_{OLS}$, $i = 1,2,...,p$ and $\hat{\sigma}^2$ is the OLS estimator of σ^2 i.e. $\hat{\sigma}^2 = (Y'Y - \hat{\alpha}'Z'Y)/(n-p-1)$.

In the context of two-parameter estimator, Yang and Chang (2010), Wu and Yang (2011) have recently suggested two-parameter estimator's alternative to LS estimator in the presence of multicollinearity. These estimators are given as:

Yang and Chang (2010) suggested new two-parameter (NTP) estimator, given by

$$\hat{\alpha}_{NTP} = \left(\wedge + I\right)^{-1}\left(\wedge + dI\right)\left(\wedge + kI\right)^{-1} Z^{'}Y \tag{10}$$

where, $k = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p} \hat{\alpha}_i^2$ and

$$d = \frac{\sum_{i=1}^{p} \left\{ \left[(k+1)\lambda_i + k\right]\lambda_i \hat{\alpha}_i^2 - \lambda_i^{\,2}\hat{\sigma}^2 \right\} \Big/ \left[(\lambda_i+1)^2(\lambda_i+k)^2\right]}{\sum_{i=1}^{p} (\hat{\sigma}^2 + \lambda_i\hat{\alpha}_i^2) \Big/ \left[(\lambda_i+1)^2(\lambda_i+k)^2\right]}.$$

Also, MSE of $\hat{\alpha}_{NTP}$ is given as:

$$MSE\left(\hat{\alpha}_{NTP}\right) = \sigma^2 \sum_{i=1}^{p} \left[\frac{\lambda_i(\lambda_i+d)^2}{(\lambda_i+1)^2(\lambda_i+k)^2} \right] + \sum_{i=1}^{p} \left\{ \frac{\left[(k+1-d)\lambda_i + k\right]^2}{(\lambda_i+1)^2(\lambda_i+k)^2} \right\} \alpha_i^2 . \tag{11}$$

Wu and Yang (2011) introduced always unbiased two-parameter (AUTP) estimator, which is defined by

$$\hat{\alpha}_{AUTP} = \hat{\alpha}_{TP} + k(1-d)\left(\wedge + kI\right)^{-1}\hat{\alpha}_{TP} \tag{12}$$

where, $d < 1 - \min\left(\hat{\sigma} \Big/ \sqrt{\lambda_i\hat{\alpha}_i^2 + \hat{\sigma}^2} \right)$ and $k = \lambda_i\hat{\sigma} \Big/ \left[(1-d)\sqrt{\lambda_i\hat{\alpha}_i^2 + \hat{\sigma}^2} - \hat{\sigma} \right]$.

Also, MSE of $\hat{\alpha}_{AUTP}$ is given as:

$$MSE\left(\hat{\alpha}_{AUTP}\right) = \sigma^2 \sum_{i=1}^{p} \left\{ \frac{\left[\lambda_i(\lambda_i+2k) + d\,k^2(2-d)\right]^2}{\lambda_i\left(\lambda_i+k\right)^4} \right\} + \sum_{i=1}^{p} \left[\frac{k^4(1-d)^4}{\left(\lambda_i+k\right)^4} \right] \alpha_i^2 . \tag{13}$$

Recently, Dorugade (2014) introduced a modified two-parameter (MTP) estimator, which is defined by

$$\hat{\alpha}_{MTP} = \left[I + k(1-d)(\Lambda + kdI)^{-1}\right]\left[I - kd\,(\Lambda + kdI)^{-1}\right]\hat{\alpha}_{OLS} \tag{14}$$

Where, $k = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p} \hat{\alpha}_i^2$ and $\hat{d} = \sum_{i=1}^{p} \left[\frac{\left(\lambda_i + k\right)\left(\hat{\sigma}^2 + \lambda_i\hat{\alpha}_i^2\right) - \lambda_i}{k\hat{\alpha}_i^2} \right]$.

Also, MSE of $\hat{\alpha}_{MTP}$ is given as:

$$MSE\left(\hat{\alpha}_{MTP}\right) = \sigma^2 \sum_{i=1}^{p} \left[\frac{\lambda_i\left(\lambda_i+k\right)^2}{\left(\lambda_i+kd\right)^4} \right] + \sum_{i=1}^{p} \left\{ \frac{k^2\left[(1-2d)\lambda_i - kd^2\right]^2}{\left(\lambda_i+kd\right)^4} \right\} \alpha_i^2 . \tag{15}$$

All the above methods of estimating α are used in Section 6.

3. Proposed estimator

Hoerl and Kennard (1962) first suggested that to control the inflation and general instability associated with the least squares estimates. The relationship of a ridge estimate to an ordinary estimate is given by the alternative form. It is observed that OLS is unbiased but has inflated variances under multicollinearity. Due to the complicated nature of the ridge parameter k, in the ridge regression method, Liu (2003) proposed a two-parameter estimator. In this article we introduce the efficient two-parameter estimator, which can be computed in two steps. Initially, following the similar method proposed by Liu (1993), Kaciranlar *et al.,* (1999) and Yang and Chang (2010), we introduce the two-parameter estimator as:

$$\hat{\alpha}^{*} = \left[\wedge + k(1-d)I\right]^{-1} Z'Y . \tag{16}$$

Hoerl and Kennard (1970a) pointed that ORR avoids inflating the variances at the cost of bias by pre-multiplying $\hat{\alpha}_{OLS}$ with the matrix $[I-k(\Lambda+kI)^{-1}]$ to reduce the inflated variances in OLS. The proposed estimator is based on this same logic by pre-multiplying $\hat{\alpha}^{*}$ with the matrix $[I-k(\Lambda+kI)^{-1}]$. This is defined in equation (17) as :

$$\hat{\alpha}_{ETP} = [I-k(\Lambda+kI)^{-1}]\hat{\alpha}^{*}, \tag{17}$$

or

$$\hat{\alpha}_{ETP} = \Lambda^{2}[(\Lambda+kI)^{-1}(\Lambda+k(1-d)I)^{-1}]\,\hat{\alpha}_{OLS} .$$

Equation (17) is termed as Efficient Two-Parameter (ETP) estimator of α. Thus, the coordinate wise estimators can be written as

$$\hat{\alpha}_{iETP} = \left[\frac{\lambda_i^{2}}{(\lambda_i+k)(\lambda_i+k(1-d))}\right]\hat{\alpha}_i \quad i = 1,2,...,\, p \tag{18}$$

where $\hat{\alpha}_i$ are the individual components of $\hat{\alpha}_{OLS}$.

We can see that it is a general estimator which includes the OLS and ORR estimators as special cases:

at $(k=0,\ d)$ $\hat{\alpha}_{ETP} = \hat{\alpha}_{OLS}$ the OLS estimator,

at $(k,\ d=1)$ $\hat{\alpha}_{ETP} = [I-k(\wedge+kI)^{-1}]\hat{\alpha}_{OLS}$ the ORR estimator,

at $(k=d=1)$ $\hat{\alpha}_{ETP} = [I-(\wedge+I)^{-1}]\hat{\alpha}_{OLS}$,

at $(k,\ d=0)$ $\hat{\alpha}_{ETP} = [I-k(\wedge+kI)^{-1}]\hat{\alpha}_{ORR}$.

3.1. Bias, Variance and MSE of ETP estimator

It is clear that $\hat{\alpha}_{ETP}$ is biased estimator, the bias of the ETP estimator is given by:

$$Bias(\hat{\alpha}_{ETP}) = E[\hat{\alpha}_{ETP}] - \alpha$$

$$= [\Lambda^2(\Lambda + kI)^{-1}(\Lambda + k(1-d)I)^{-1} - I]\,\alpha$$

$$= \sum_{i=1}^{p} \left\{ \frac{-\left[k(2-d)\lambda_i + k^2(1-d)\right]}{(\lambda_i + k)(\lambda_i + k(1-d))} \right\} \hat{\alpha}_i \tag{19}$$

$V(\hat{\alpha}_{ETP})$ = $\sigma^2 V \wedge^{-1} V'$ where $V = \Lambda^2[(\Lambda + kI)^{-1}(\Lambda + k(1-d)I)^{-1}]$

$$= \hat{\sigma}^2 \sum_{i=1}^{p} \left[\frac{\lambda_i^{\,3}}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right]. \tag{20}$$

The MSE of ETP estimator is

$$MSE(\hat{\alpha}_{ETP}) = V(\hat{\alpha}_{ETP}) + [Bias(\hat{\alpha}_{ETP})]\,[Bias(\hat{\alpha}_{ETP})]'$$

$$MSE(\hat{\alpha}_{ETP}) = \hat{\sigma}^2 \sum_{i=1}^{p} \left[\frac{\lambda_i^{\,3}}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right] + \sum_{i=1}^{p} \left\{ \frac{\left[k(2-d)\lambda_i + k^2(1-d)\right]^2}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right\} \hat{\alpha}_i^2 . \tag{21}$$

Setting $k = 0$ in (21), we obtain

$$MSE(\hat{\alpha}_{OLS}) = \hat{\sigma}^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} . \tag{22}$$

Also, setting $d = 1$ in (21), we obtain

$$MSE(\hat{\alpha}_{ORR}) = \hat{\sigma}^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{(\lambda_i + k)^2} . \tag{23}$$

4. Performance of the proposed estimator

This section compares the performance of $\hat{\alpha}_{ETP}$ with $\hat{\alpha}_{OLS}$ and $\hat{\alpha}_{ORR}$ using MSE criteria.

4.1. Comparison between the $\hat{\alpha}_{ETP}$ and $\hat{\alpha}_{OLS}$ using MSE criterion

The difference between $\hat{\alpha}_{OLS}$ and $\hat{\alpha}_{ETP}$ in the MSE sense is as follows:

$MSE(\hat{\alpha}_{OLS}) - MSE(\hat{\alpha}_{ETP})$

$$= \hat{\sigma}^2 \sum_{i=1}^{p} \left[\frac{1}{\lambda_i} - \frac{{\lambda_i}^3}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right] - \sum_{i=1}^{p} \left\{ \frac{\left[k(2-d)\lambda_i + k^2(1-d) \right]^2}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right\} \hat{\alpha}_i^2$$

$$= \sum_{i=1}^{p} \left\{ \frac{\hat{\sigma}^2 \left[(\lambda_i + k)^2 (\lambda_i + k(1-d))^2 - {\lambda_i}^4 \right] - \lambda_i \hat{\alpha}_i^2 \left[k(2-d)\lambda_i + k^2(1-d) \right]^2}{\lambda_i (\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right\}.$$

From the above equation, it can be shown that the difference $MSE(\hat{\alpha}_{OLS})$- $MSE(\hat{\alpha}_{ETP})$ will be positive if and only if

$$\hat{\sigma}^2 \left[(\lambda_i + k)^2 (\lambda_i + k(1-d))^2 - {\lambda_i}^4 \right] \geq \lambda_i \hat{\alpha}_i^2 \left[k(2-d)\lambda_i + k^2(1-d) \right]^2$$

Thus, $MSE(\hat{\alpha}_{OLS}) \geq MSE(\hat{\alpha}_{ETP})$

if and only if

$$\hat{\sigma}^2 \left[(\lambda_i + k)^2 (\lambda_i + k(1-d))^2 - {\lambda_i}^4 \right] \geq \lambda_i \hat{\alpha}_i^2 \left[k(2-d)\lambda_i + k^2(1-d) \right]^2 . \quad (24)$$

4.2. Comparison between the $\hat{\alpha}_{ETP}$ and $\hat{\alpha}_{ORR}$

The difference between $\hat{\alpha}_{ORR}$ and $\hat{\alpha}_{ETP}$ in the MSE sense is as follows:

$$MSE(\hat{\alpha}_{ORR}) - MSE(\hat{\alpha}_{ETP}) = \hat{\sigma}^2 \sum_{i=1}^{p} \left[\frac{\lambda_i}{(\lambda_i + k)^2} - \frac{{\lambda_i}^3}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right]$$

$$+ \sum_{i=1}^{p} \left\{ \frac{k^2}{(\lambda_i + k)^2} - \frac{\left[k(2-d)\lambda_i + k^2(1-d) \right]^2}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right\} \hat{\alpha}_i^2$$

$$= \sum_{i=1}^{p} \left\{ \frac{\hat{\sigma}^2 k(1-d)\lambda_i \left[2\lambda_i + k(1-d) \right] + \hat{\alpha}_i^2 \left\{ k^2 (\lambda_i + k(1-d))^2 - \left[k(2-d)\lambda_i + k^2(1-d) \right]^2 \right\}}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right\}$$

From above equation, it can be shown that the difference $MSE(\hat{\alpha}_{ORR})$- $MSE(\hat{\alpha}_{ETP})$ will be positive if and only if $k^2 (\lambda_i + k(1-d))^2 \geq \left[k(2-d)\lambda_i + k^2(1-d) \right]^2$

Thus, $MSE(\hat{\alpha}_{ORR}) \geq MSE(\hat{\alpha}_{ETP})$

if and only if

$$k^2 (\lambda_i + k(1-d))^2 \geq \left[k(2-d)\lambda_i + k^2(1-d) \right]^2 . \quad (25)$$

5. Determination of ridge parameter k and d

A very important issue in the study of the ridge and Liu regression is how to find appropriate ridge and Liu parameters, k and d respectively. These

parameters may either be nonstochastic or may depend on the observed data. The choice of values for these ridge parameters has been one of the most difficult problems confronting the study of the generalized ridge regression.

In order to determine and evaluate the performance of our proposed estimator $\hat{\alpha}_{ETP}$ as compare to OLS estimator and others, we will find the optimal values of ridge parameters k and d. Let $\hat{k}$ is the optimal value of the k determined by well-known method of determining the ridge parameter, the optimal value of the d can be considered to be this d that minimizes $MSE(\hat{\alpha}_{ETP})$.

Let $g(k,d) = MSE(\hat{\alpha}_{ETP})$

$$= \sigma^2 \sum_{i=1}^{p} \left[\frac{\lambda_i^3}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right] + \sum_{i=1}^{p} \left\{ \frac{\left[k(2-d)\lambda_i + k^2(1-d) \right]^2}{(\lambda_i + k)^2 (\lambda_i + k(1-d))^2} \right\} \alpha_i^2 .$$

Then, by differentiating $g(k,d)$ w.r.t. d and equating to 0, we have

$$d = \sum_{i=1}^{p} \left\{ \frac{k \alpha_i^2 \lambda_i^2 \left[(\lambda_i + k) k + \lambda_i \right] - \sigma^2 \lambda_i^3}{k^2 \alpha_i^2 \lambda_i^2 (\lambda_i + k)} \right\}, \tag{26}$$

d is the function of k depends on the σ^2 and α_i. For practical purposes, they are replaced by their unbiased estimator $\hat{\sigma}^2$ and $\hat{\alpha}_i$. Hence,

$$\hat{d} = \sum_{i=1}^{p} \left\{ \frac{k \hat{\alpha}_i^2 \lambda_i^2 \left[(\lambda_i + k) k + \lambda_i \right] - \hat{\sigma}^2 \lambda_i^3}{k^2 \hat{\alpha}_i^2 \lambda_i^2 (\lambda_i + k)} \right\}. \tag{27}$$

6. Comparative study

6.1. A simulation study

The performance of the proposed estimator and the existing estimators is examined via a simulation study. The simulation is carried out under different degrees of multicollinearity. The average MSE (AMSE) ratios of the $\hat{\alpha}_{ETP}$ and other ridge estimators over OLS estimator are evaluated. The true model is considered as $Y = X\beta + \varepsilon$. Here ε follows a normal distribution $N(0, \sigma^2 I_n)$ and following McDonald and Galerneau (1975) the explanatory variables are generated by

$$x_{ij} = (1-\rho^2)^{1/2} u_{ij} + \rho u_{ip} \quad i = 1,2,..., n \qquad j = 1,2,..., p$$

where u_{ij} are independent standard normal pseudo-random numbers and ρ is specified so that the theoretical correlation between any two explanatory variables is given by ρ^2. In this study, to investigate the effects of different degrees of multicollinearity on the estimators, we consider two different correlations, ρ = 0.95, 0.99. β parameter vectors are chosen arbitrarily such that

$\beta = (2, 1, 6, 2)'$, $\beta = (1, 1, 3)'$ for p = 4 and 3, respectively. The sizes of samples are 20, 50 and 100. The variance of the error terms is taken as σ^2 = 1, 5 and 10.

The well-known ridge parameter $k = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p} \hat{\alpha}_i^2$ suggested by Hoerl *et al.*, (1975) was used. d is determined as defined in equation (27). Efficient two-parameter estimator and estimators given in (3) to (5), (7) to (10), (12) and (14) are computed. The experiment is repeated 1000 times and the average MSE (AMSE) of estimators is obtained using the following expression:

$$AMSE(\hat{\alpha}) = \frac{1}{1000} \sum_{i=1}^{p} \sum_{j=1}^{1000} (\hat{\alpha}_{ij} - \alpha_i)^2 \tag{28}$$

where $\hat{\alpha}_{ij}$ denote the estimator of the ith parameter in the jth replication and α_i, $i = 1,2,...,p$ are the true parameter values.

Firstly, we computed the AMSE ratios ($AMSE(\hat{\alpha}_{OLS})/AMSE(\hat{\alpha})$) of OLS estimator over different estimators for various values of triplet (ρ, n, σ^2) and reported in Tables 1-4. We consider the method that leads to the maximum AMSE ratio to the best from the MSE point of view.

From Tables 1-4, we observe that the performance of our proposed efficient two-parameter estimator ($\hat{\alpha}_{ETP}$) is better than $\hat{\alpha}_{OLS}$. Also, $\hat{\alpha}_{ETP}$ is more efficient in terms of MSE than other biased estimators $\hat{\alpha}_{JRR}$, $\hat{\alpha}_{AUL}$, $\hat{\alpha}_{URR}$, $\hat{\alpha}_{TP}$, $\hat{\alpha}_{NTP}$, $\hat{\alpha}_{AUTP}$ and $\hat{\alpha}_{MTP}$ including $\hat{\alpha}_{ORR}$ for various values of triplet (ρ, n, σ^2). The results agree with our theoretical findings in Section 4.

Table 1. Ratio of AMSE of OLS over various ridge estimators
(p = 4, $\beta = (2, 1, 6, 2)'$ and ρ = 0.95)

Estimator	n = 20			50			100		
	$\hat{\sigma}^2$ = **1**	**5**	**10**	**1**	**5**	**10**	**1**	**5**	**10**
$\hat{\alpha}_{ORR}$	1.0209	1.3800	1.4746	1.5603	1.7090	2.1274	2.1833	2.2644	2.3872
$\hat{\alpha}_{JRR}$	1.0019	1.3525	1.3624	1.5186	1.5745	1.8444	1.7952	1.4780	1.4779
$\hat{\alpha}_{AUL}$	1.0326	1.2966	1.0040	1.0793	1.2278	1.0487	1.0013	1.0125	1.0219
$\hat{\alpha}_{URR}$	1.4067	1.3184	1.3363	1.5621	1.7799	1.6437	1.8530	2.4037	3.2057
$\hat{\alpha}_{TP}$	0.9909	0.8698	0.9730	0.8516	0.7984	0.8615	0.9675	0.9390	1.0504
$\hat{\alpha}_{NTP}$	0.7030	0.3278	1.2536	1.5321	0.8943	1.9981	2.4686	2.0685	2.3089
$\hat{\alpha}_{AUTP}$	1.0004	1.0695	1.0017	1.0697	1.1131	1.0328	1.0014	1.0032	1.0017
$\hat{\alpha}_{MTP}$	1.0210	1.3800	1.4748	1.5600	1.7094	2.1270	2.1836	2.2648	2.3870
$\hat{\alpha}_{ETP}$	1.4388	1.4078	1.5967	1.6028	1.8579	2.4492	2.6313	3.0779	3.3611

Table 2. Ratio of AMSE of OLS over various ridge estimators (p = 4, $\beta = (2, 1, 6, 2)'$ and ρ = 0.99)

Estimator	n = 20			50			100		
	$\hat{\sigma}^2 = 1$	5	10	1	5	10	1	5	10
$\hat{\alpha}_{ORR}$	1.2365	1.4166	1.9591	2.2037	2.7241	2.8145	2.2735	2.7244	2.9822
$\hat{\alpha}_{JRR}$	1.1365	1.1309	1.4754	1.2585	2.4739	1.7537	2.2602	2.3833	1.6211
$\hat{\alpha}_{AUL}$	1.0873	1.0061	1.052	1.068	1.0458	1.02	1.0022	1.0142	1.1438
$\hat{\alpha}_{URR}$	1.2691	1.4654	0.9704	1.7314	2.281	2.8455	1.5181	1.742	2.257
$\hat{\alpha}_{TP}$	1.0435	0.9784	0.8843	0.9211	1.2129	0.8857	1.0332	1.1264	0.8024
$\hat{\alpha}_{NTP}$	1.1173	1.1683	1.5396	0.516	2.587	2.9986	2.5124	2.1497	2.0449
$\hat{\alpha}_{AUTP}$	1.0054	1.0006	1.0176	1.0036	1.0409	1.0119	1.0016	1.0153	1.0336
$\hat{\alpha}_{MTP}$	1.2300	1.4168	1.8960	2.2039	2.7245	2.8150	2.2740	2.7250	2.9820
$\hat{\alpha}_{ETP}$	1.34	1.5982	2.5048	2.6136	3.0088	4.0831	2.2873	3.1294	4.2129

Table 3. Ratio of AMSE of OLS over various ridge estimators (p = 3, $\beta = (1, 1, 3)'$ and ρ = 0.95)

Estimator	n = 20			50			100		
	$\hat{\sigma}^2 = 1$	5	10	1	5	10	1	5	10
$\hat{\alpha}_{ORR}$	1.5559	2.0183	2.7696	1.8432	1.923	2.3064	2.0479	2.0930	3.0024
$\hat{\alpha}_{JRR}$	1.5030	1.3665	1.4734	1.7442	1.5766	1.7138	1.3825	1.3013	1.9361
$\hat{\alpha}_{AUL}$	1.0061	1.0577	1.0181	1.0044	1.0591	1.0157	1.0690	1.0469	1.0214
$\hat{\alpha}_{URR}$	1.1864	1.7946	2.5463	1.8317	1.8465	0.7616	1.7940	1.8067	2.5083
$\hat{\alpha}_{TP}$	0.9606	0.8749	0.9631	1.0025	0.8328	0.8223	0.8684	0.8958	0.9171
$\hat{\alpha}_{NTP}$	1.2280	1.7430	2.1870	1.8514	1.5227	1.6014	1.6203	1.8242	2.7112
$\hat{\alpha}_{AUTP}$	1.0023	1.0176	1.0044	1.003	1.0301	1.027	1.0118	1.0090	1.0063
$\hat{\alpha}_{MTP}$	1.5550	2.0189	2.7755	1.8440	1.933	2.3164	2.0586	2.0980	3.0124
$\hat{\alpha}_{ETP}$	1.6037	2.3405	3.5764	1.9354	2.2614	2.8938	2.3319	2.4589	4.0607

Table 4. Ratio of AMSE of OLS over various ridge estimators (p = 3, $\beta = (1, 1, 3)'$ and ρ = 0.99)

Estimator	n = 20			50			100		
	$\hat{\sigma}^2 = 1$	5	10	1	5	10	1	5	10
$\hat{\alpha}_{ORR}$	1.5622	2.8334	3.0927	1.9421	2.2077	2.4474	2.5685	2.6181	3.5616
$\hat{\alpha}_{JRR}$	1.2507	1.8883	2.2305	1.2062	1.7607	1.4847	1.7945	1.4024	2.1656
$\hat{\alpha}_{AUL}$	1.0227	1.0308	1.0204	1.0817	1.0179	1.1277	1.0273	1.1743	1.0506
$\hat{\alpha}_{URR}$	1.4735	1.8497	2.1001	1.5143	1.4785	2.1487	2.3512	2.0673	2.2603
$\hat{\alpha}_{TP}$	0.8635	0.8921	0.8776	0.8935	0.915	0.8246	0.9067	0.8155	0.8338
$\hat{\alpha}_{NTP}$	1.4160	2.0758	2.5563	1.3947	1.9748	1.7623	0.5709	1.6083	3.438
$\hat{\alpha}_{AUTP}$	1.0032	1.0115	1.0058	1.0076	1.0096	1.0288	1.0111	1.0246	1.0251
$\hat{\alpha}_{MTP}$	1.5620	2.8634	3.2927	2.1011	2.3077	2.8474	3.1685	2.9182	3.7521
$\hat{\alpha}_{ETP}$	1.8216	3.9784	3.6171	2.1426	2.7185	3.2002	3.3286	3.1803	4.4847

6.2. Numerical example

To validate the theoretical results, the numerical example used by Gruber (1988) was adopted. It was established that this data set suffers multicollinearity. Data shows Total National Research and Development Expenditures as a Percent of Gross National Product by Country: 1972-1986. It represents the relationship between the dependent variable Y, the percentage spent by the United States and four other independent variables X_1, X_2, X_3, and X_4.

The estimated MSE values for $\hat{\alpha}_{OLS}, \hat{\alpha}_{ORR}, \hat{\alpha}_{NTP}, \hat{\alpha}_{AUTP}, \hat{\alpha}_{MTP}$ and $\hat{\alpha}_{ETP}$ estimators, are obtained and reported in Table 5.

Table 5. Values of MSE

Estimator	$\hat{\alpha}_{OLS}$	$\hat{\alpha}_{ORR}$	$\hat{\alpha}_{NTP}$	$\hat{\alpha}_{AUTP}$	$\hat{\alpha}_{MTP}$	$\hat{\alpha}_{ETP}$
MSE	0.2833	0.1256	0.1255	0.2832	0.1257	0.1251

From Table 5, it was observed that the estimated MSE value of the efficient two-parameter estimator ($\hat{\alpha}_{ETP}$**)** is always smaller than those of the $\hat{\alpha}_{OLS}$, $\hat{\alpha}_{ORR}$, $\hat{\alpha}_{NTP}$, $\hat{\alpha}_{AUTP}$ and $\hat{\alpha}_{MTP}$ estimators. The results agree with the theoretical findings in Section 4. Finally, $\hat{\alpha}_{ETP}$ is meaningful in practice.

7. Conclusion

A biased efficient two-parameter estimator has been proposed for estimating the parameter of the linear regression model with multicollinearity. The proposed estimator is examined against OLS and ORR estimator in terms of scalar MSE criterion. Finally, from the simulation study and numerical example, the performance of the proposed estimator is satisfactory in the presence of multicollinearity over other estimators reviewed in this article.

Acknowledgements

The author would like to thank the referee for his valuable comments, which resulted in the present version of this article. In the part Research Award Scheme, the present studies were supported by UGC, India Project No. F. 30-1/2014/RA-2014-16-GE-MAH-5958 (SA-II).

REFERENCE

AKDENIZ, F., KACIRANLAR, S., (1995). On the almost unbiased generalized Liu estimator and unbiased estimation of the bias and MSE, Commun. Statist. Theor. Meth., 24, pp. 1789–1797.

CROUSE, R. H., JIN, C., HANUMARA, R. C., (1995). Unbiased ridge estimation with prior information and ridge trace, Commun. Statist. Theor. Meth., 24, pp. 2341–2354.

DORUGADE, A. V., (2014). Modified Two Parameter Estimator in Linear regression, Statistics in Transition - new series, 15 (1), pp. 23–36.

GRUBER, M. H. J., (1998). Improving efficiency by shrinkage the James–Stein and Ridge regression estimators. Marcell Dekker, NewYork.

HOERL, A. E., (1962). Application of ridge analysis to regression problems, Chemical Engineering Progress., 68, pp. 54–59.

HOERL, A. E., KENNARD, R. W., (1968). On regression analysis and biased estimation, Technometrics, 10, pp. 422–423.

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

HOERL, A. E., KENNARD, R. W., (1970b). Ridge regression: Applications to Nonorthogonal problems, Technometrics, 12, pp. 69–82.

HOERL, A. E., KENNARD, R. W., BALDWIN, K. F., (1975). Ridge regression: Some Simulations, Commun. Statist., 4, pp. 105–123.

KACIRANLAR, S., SAKALLIOGLU, S., AKDENIZ, F., STYAN, G. P. H., WERNER, H. J., (1999). A new biased estimator in linear regression and a detailed analysis of the widely-analysed dataset on Portland Cement, Sankhya Ind. J. Statist., 61, pp. 443–459.

LIU, K., (1993). A new class of biased estimate in linear regression, Commun. Statist. Theor. Meth., 22, pp. 393–402.

LIU, K., (2003). Using Liu-type estimator to combat Collinearity. Commun, Statist. Theor. Meth., 32, pp. 1009–1020.

MCDONALD G. C., GALARNEAU, D. I., (1975). A Monte Carlo evaluation of some ridge-type estimators, J Am Stat Assoc., 20, pp. 407–416.

MONTGOMERY, D. C., PECK, E. A., VINING, G. G., (2006). Introduction to linear regression analysis. John Wiley and Sons, New York.

OZKALE, M. R., KACIRANLAR , S., (2007). The restricted and unrestricted two-parameter estimators, Commun. Statist. Theor. Meth., 36, pp. 2707–2725.

SINGH, B., CHAUBEY, Y. P., (1987). On some improved ridge estimators, Stat Papers, 28, pp. 53–67.

YANG, H., CHANG, X., (2010). A New Two-Parameter Estimator in Linear Regression, Commun. Statist. Theor. Meth., 39, pp. 923–934.

WU, J., YANG, H., (2011). Efficiency of an almost unbiased two-parameter estimator in linear regression model, Statistics., 47 (3), pp. 535–545.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 187

SATELLITE WORKSHOP ON SOME TOPICS IN THE PRACTICE OF SURVEY SAMPLING

Instructor: **Graham Kalton**

Date: **2 July 2019**, the pre-conference day of the MET2019.

Venue: Statistics Poland/GUS, Al. Niepodległości 208, Room 149

A satellite training workshop will take place on Tuesday, 2 July, the day preceding the MET2019 (Methodology of Statistical Research) conference. The workshop instructor will be Graham Kalton, one of the leading authorities in the field of survey research, author of textbooks and papers on survey sampling and survey methodology.

The day-long program will focus on selected topics often encountered in survey practice—such as designs for surveys over time, methods for dealing with deficiencies in sampling frames, dealing with problems caused by inaccurate measures of size in multistage samples, and future trends in survey research.

This free workshop is addressed primarily to employees in national and local statistical agencies who are interested in the methodology of survey sampling practice.

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 189–192

ABOUT THE AUTHORS

Adepoju Abosede Adedayo is a former Acting Head of Department in the Department of Statistics, Faculty of Science, University of Ibadan. She has over twenty years of teaching and research experience. Her research interests include econometric modelling, applied Bayesian analysis, economic/financial statistics and environmental statistics. She was appointed the Educational Ambassador by the American Statistical Association (ASA) for the year 2016. She is a member of many professional bodies like International Biometric Society (IBS) Group Nigeria, International Statistical Institute (ISI), Caucus of Women in Statistics of the American Statistical Association (ASA) and Nigerian Statistical Association (NSA). She is a Research Fellow with the Centre for Econometric and Allied Research (CEAR), University of Ibadan.

Chaturvedi Anoop is a Professor in the Department of Statistics, University of Allahabad. He has worked on diversified fields including Bayesian unit root testing, developing improved estimators for linear models, Bayesian and classical analysis of various econometric models including panel data models, spatial autoregressive and dynamic models, and model averaging. He is a fellow of Royal Statistical Society. He has visited Université Paris II, Paris, University of Wollongong, Australia; City University of Hong Kong; CentER, Tilburg University; Chinese Academy of Sciences. He has authored two books and 100 research papers.

Cibulková Jana is a third year PhD student of Statistics study program at the University of Economics, Prague (Department of Statistics and Probability), Czech Republic. She has received her Master's degree from the Masaryk University in Brno in a study program of Statistics and Data Analysis. Her current area of interest is cluster analysis and its application in the field of marketing. In her research, she focuses on cluster analysis, mainly on clustering of categorical and binary data. Besides her studies, she works as a data scientist, where she works closely with a marketing department.

Dalei Narendra Nath is working as an Assistant Professor in the Department of Economics, School of Business, UPES, Dehradun. His areas of interest are energy economics, applied economics, econometrics and environmental economics. He has published over 15 research papers in renowned international/national journals along with two chapters in the books.

Dehnel Grażyna is an Associate Professor at the Department of Statistics, Poznań University of Economics and Business. Her main research domain is small area estimation, survey sampling, short-term and structural business statistics. She is also interested in outlier robust regression applied to business data, business demography and data integration.

Dorugade Ashok is a Assistant Professor in the Department of Statistics in Yashwantrao Chavan Mahavidyalaya, Halkarni, Dist-Kolhapur, India. His research intersts are multivariate statistical analysis, Applied Statistics, and Biostatistics in particular. Assistant Professor Dorugade has published 11 research papers in internationals. He has also completed 3 research projects. He is an active meber of professional bodies.

Górecki Tomasz received his MSc in mathematics from the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznań, Poland, in 2001. There he received his PhD in 2005. He obtained habilitation in computer science from Systems Research Institute Polish Academy of Sciences in 2015. Currently, he is an Assistant Professor at this University. His research interests include machine learning, times series classification and data mining.

Janusz Bartłomiej is working as a manager in Internal Audit Unit in TAURON Polska Energia S.A. He completed his PhD in Statistics from University of Economics in Katowice, Poland, in 2016. His research interests are applying statistics in auditing and audit sampling.

Kozioł Arkadiusz is a PhD at the Faculty of Mathematics, Computer Science and Econometrics in University of Zielona Góra, Poland. His PhD thesis was defended in 2019 and the PhD dissertation was awarded by the scientific council of the faculty. His research interests are multivariate statistical analysis, linear models, statistical inference and statistical applications, for example in medicine. He has already published 5 papers in international journals. He was an invited speaker in 20th European Young Statisticians Meeting in 2017.

Krzyśko Mirosław is a Full Professor of Mathematics and Statistics. His research interests are multivariate statistical analysis, analysis of multivariate functional data, statistical inference and data analysis in particular. Professor Krzyśko has published over 150 research papers in international/national journals and conferences. He has also published five books/monographs. Professor Krzyśko is an active member of many scientific professional bodies.

Mishra Sandeep received his PhD degree from University of Allahabad, Prayagraj, India, in 2018. He is currently a Guest Faculty in the Department of Statistics, University of Allahabad. His research interest includes Bayesian econometrics, spatial modelling and demography.

Ogundunmade Tayo Peter is a Tutorial Assistant in the Department of Statistics, University of Ibadan, Nigeria. He simultaneously works as a Research Assistant & Senior Collaborator at the University of Ibadan Laboratory for Interdisciplinary Statistical Analysis (UI-LISA), Department of Statistics, and a Volunteer - Data Management Specialist at Centre for Petroleum Energy Economic and Law (CPEEL), University of Ibadan. At UI-LISA, he is serving as a research consultant and conducts daily statistical consulting and research advice to non-statisticians and domain experts from other faculties and departments within and outside the University of Ibadan community, who are in need of statistical advice for their various research. He has facilitated series of trainings and workshops on statistical research, statistical software usage and statistical methodology. Tayo is currently undergoing his PhD study in Statistics with research focus in Bayesian

neural network and data mining. He enjoys coding with R and Python. He loves to impact knowledge and enjoys sharing ideas. He has mentored several undergraduate students in the Department. He wants to be a renowned academic scholar known for quality statistical research in Africa.

Pawełek Barbara has received the PhD and Habilitation in economics from the Cracow University of Economics (Poland). At present, she is an Associate Professor in the Department of Statistics at the same university. She has done extensive research on data analysis, classification methods, data mining, modelling and forecasting of socio-economic phenomena, prediction of company bankruptcy.

Řezanková Hana is a Professor at the University of Economics, Prague (Faculty of Informatics and Statistics, Department of Statistics and Probability). She received her Master and Ph.D. degrees from the University of Economics, Prague. From 2008, she is a full Professor of statistics. Her main interests are multivariate statistical methods, mainly cluster analysis and categorical data analysis. She is a vice-president of the Czech Statistical Society (past president in the period 2013–2017), and a member of the Czech Statistical Council of the Czech Statistical Office.

Roy Hiranmoy is working as an Associate Professor and Head of the Department of Economics, School of Business, UPES, Dehradun. His areas of interest are energy economics, development economics, power economics and environmental economics. He has published over 20 research papers in renowned international/national journals along with two books and two chapters in books. He has organized several workshops/FDPs/conferences, etc.

Sharma Anukriti is a Doctoral Research Fellow in the Department of Economics, School of Business, UPES, Dehradun. Her areas of interest are energy economics, applied economics and environmental economics. She has published five research paper in renowned international/national journals. She has completed MSc (Applied Economics). She has received the Erasmus+ scholarship of the European Union under the PhD exchange program and spent three months in the University of Maribor, Slovenia for the research purpose.

Šulc Zdeněk is a Assistant Professor at the University of Economics, Prague. His primary field of interest is categorical data classification. In unsupervised classification, he deals with similarity measures for categorical data with more than two categories for purposes of hierarchical cluster analysis. He is the author of the R package ”nomclust“ for hierarchical clustering of data characterized by nominal variables. In supervised classification, he currently deals with the Delta Machine classification method in cooperation with the Leiden University, which is considered as an alternative to logistic regression. This method uses distances between objects as predictors for a class variable. Furthermore, he deals with processing medical data, especially with creating norms for the healthy people and evaluating neuropsychological test quality.

Sirota Sergej is a doctoral student at the University of Economics, Prague (Department of Statistics and Probability), where he started his research in 2015 after Master’s degree Statistics from the same university (the title of the diploma

thesis is Scoring methods used in cluster analysis). The area of his doctoral research is bank's propensity models (this is also the title of the doctoral thesis) with the main focus on using cluster analysis to improve propensity models.

Walesiak Marek is a Professor at Wroclaw University of Economics in Department of Econometrics and Computer Science. He is a member of the Methodological Commission in Statistics Poland (GUS) and an active member of many scientific professional bodies. His main areas of interest include: classification and data analysis, multivariate statistical analysis, marketing research, computational techniques in R. Currently, he is a member of editorial boards of two journals (Przegląd Statystyczny. Statistical Review and Econometrics. Advances in Applied Data Analysis) and a scientific council of one journal (Wiadomości Statystyczne. The Polish Statistician).

Wołyński Waldemar is a Professor at the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznań, Poland. His major research interests focus on various aspects of multivariate statistical analysis. He is an author or co-author of over 50 research papers.

Zmyślony Roman is a Professor at the Faculty of Mathematics and Computer Sciences in University of Zielona Góra, Poland. His research interests are multivariate statistical analysis, linear models, completeness of minimal sufficient statistics, statistical inference and data analysis in chemistry and agriculture. Professor Zmyślony has published more than 80 research papers in international/national journals and conferences. Moreover, he has two patents in chemical industry. He has also published five books as a co-author. Professor Zmyślony is an active member of many scientific professional bodies and a member of the advising committee of CMA in Portugal. He has organized a lot of national and international conferences in statistics.

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).