STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Sampling methods and estimation

Research articles

Other articles:

Multivariate Statistical Analysis 2015, Łódź. Conference Papers

Classification and data analysis −theory and applications, 2015, Gdańsk. Conference papers

Book review

Conference Announcement

Volume 17, Number 3, September 2016

Editorial Office **ISSN 1234-7655**

Marek Cierpiał-Wolan, Ph.D., Scientific Secretary
m.wolan@stat.gov.pl
Secretary:
Katarzyna Pomianowska, k.pomianowska@stat.gov.pl
Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec, Technical Assistant

Address for correspondence

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

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 359−362

FROM THE EDITOR

Three kinds of papers included in this issue are arranged in a conventional way, starting with sampling methods and estimation issues followed by research papers and other articles, the last being a selection of papers from two conferences: *Multivariate Statistical Analysis 2015*, held in Łódź and *Classification and data analysis − theory and applications, 2015*, held in Gdańsk. Altogether eleven regular articles and a book review constitute this issue.

The first article, ***Small Area Estimation of Income under Spatial SAR Model*** by **J. Kubacki** and **A. Jędrzejczak** presents the method of hierarchical Bayes (HB) estimation under small area models with spatially correlated random effects and a spatial structure implied by the Simultaneous Autoregressive (SAR) process. Calculations are based on the concept of sampling from a posterior distribution under generalized linear mixed models implemented in WinBUGS software along with estimation for small areas by means of the HB method in the case of known model hyperparameters using data from household income data. A three-stage procedure which employs Balanced Repeated Replication, bootstrap and Generalized Variance Function, and additional simulations were conducted to show the influence of the spatial autoregression coefficient on the estimation error reduction. For high spatial correlation between domains, a noticeable MSE reduction showed that the HB-SAR method is in general more efficient than the traditional spatial EBLUP technique.

R. Shanker's paper ***Sujatha distribution and its applications*** proposes a new one-parameter lifetime distribution named “Sujatha Distribution” with an increasing hazard rate for modelling lifetime data. The author presents its first four moments and expressions for coefficient of variation, skewness, kurtosis and index of dispersion. Various mathematical and statistical properties of the proposed distribution are discussed including its hazard rate function, mean residual life function, stochastic ordering, mean deviations, Bonferroni and Lorenz curves, and stress-strength reliability. The applications and goodness of fit of the distribution is also discussed using three real lifetime data sets, and the fit has been compared with one-parameter lifetime distributions including Akash, Shanker, Lindley and exponential distributions.

The next paper, ***Estimation of mean on the basis of conditional simple random sample*** by **J. L. Wywiał,** discusses a problem of estimation of the population mean in a finite and fixed population on the basis of the conditional simple random sampling design dependent on order statistics (quantiles) of an auxiliary variable. The well-known Horvitz-Thompson and ratio type estimators as well as the sample mean are taken into account under the conditional simple random sampling designs. In conclusions following the results of empirical analysis the author stresses that under some additional conditions the proposed estimation strategies based on the conditional simple random sample are in general more accurate than the mean from the simple random sample drawn without replacement. In particular, it can be expected that the sampling design might be useful in the case when there are censored observations of the auxiliary variable and in the presence of outliers.

The research papers section opens with an article by **N. M. Al-Kandari** and **P. Lahiri** on ***Prediction of a function of misclassified binary data.*** The authors start with an interesting observation that the naive predictor, which ignores the misclassification errors, is unbiased even if total misclassification error is high as long as the probabilities of false positives and false negatives are identical. Other than this case, the bias of the naive predictor depends on the misclassification distribution and the magnitude of the bias can be high in certain cases. They correct the bias of the naive predictor using a double sampling idea where both inaccurate and accurate measurements are taken on the binary variable for all the units of a sample drawn from the original data using a probability sampling scheme. Using this additional information and design-based sample survey theory, the authors derive a bias-corrected predictor and examine the cases where the new bias-corrected predictors can

also improve over the naive predictor in terms of mean square error (MSE). They conclude that given availability of additional data that generates the misclassification error, it may be possible to improve on the proposed method, and this direction they plan to explore in the future.

The next paper, ***An extension of the classical distance correlation coefficient for multivariate functional data with applications*** by **T. Górecki, M. Krzyśko, W. Ratajczak** and **W. Wołyński** is devoted to systematic exploration of the relationship between two sets of real variables defined for the same individuals which can be evaluated by a few different correlation coefficients. For the functional data there is one useful tool, canonical correlations. To extend other similar measures to the context of functional data analysis is not a simple task. The authors show how to use the distance correlation coefficient for a multivariate functional case. The approaches discussed are illustrated with an application to some socioeconomic data. The proposed method has been proved to be useful in investigating the correlation between two sets of variables, especially in the context of the hidden structure of the co-dependence between groups (pillars) of variables representing various fields of socio-economic activity.

A. Szulc's paper ***Changing mortality distribution in developed countries from 1970 to 2010: looking at averages and beyond*** starts with applying methods used in income inequality and poverty research to observe changes in life spans distribution in 35 developed countries. The analyses are performed at two levels, using the same methods when possible: i/ taking the countries as the units with a mean length of life being a single parameter representing the distribution, ii/ utilizing the country life tables (taking people as the units) in order to compare other than mean length of life attributes of mortality distribution. Increasing divergence in the mean length of life across the countries is due to growing distance of the countries below the median, mainly the post-communist ones, to the upper half. The comparisons of the within-country distributions of ages at death by means of the Kullback-Leibler divergence provides similar results. However, poverty and inequality indices calculated at this level yield opposite conclusions. The author concludes that most of the between-country variation might be attributed to the variation in the mean length of life while the changes in within-country inequality reduced this effect.

The third group of articles – all of which are based on conference presentations - starts with ***Quality of institutions and total factor productivity in the European Union*** by **A. P. Balcerzak** and **M. B. Pietrzak.** The authors examine the relationship between the quality of an institutional system and total factor productivity in the EU countries. The quality of the institutional system is defined from the perspective of incentives that influence the use of the potential of KBE and the method for linear ordering of objects was applied in order to determine the level of effectiveness of the institutional system using data from Fraser Institute. The main hypothesis that the quality of the institutional system in the context of KBE has a significant influence on the level of total factor productivity in the EU was verified through estimation of the parameters of the Cobb-Douglas production function. The calculation was based on Eurostat data. In order to identify the relationship between the quality of the institutional system and the level of TFP a panel model was applied using data for years 2000-2010. In conclusions, the authors report that according to the results of the econometric analysis the expectations concerning the positive influence of the quality of institutions on TFP in the EU countries were confirmed. It was also showed that the new EU member states have high potentials for improving their TFP, given however implementation of effective institutional reforms.

In the paper ***Locally regularized linear regression in the valuation of real estate*, M. Kubus** discusses the valuation of real estate in the comparative approach using a data set with similar properties to the data from sales transactions, within a short period of time. Given the large scope of the data sets, a local regression model was preferred over a global model, and a local feature selection via regularization was employed. The empirical analysis confirmed the effectiveness of such an approach, with special attention being paid to the model quality assessment through cross-validation for estimation of the residual standard error.

A. Kozera and **R. Głowicka-Wołoszyn** in the article ***Spatial autocorrelation in assessment of financial self-sufficiency of communes of Wielkopolska province*** employ global and local Moran I statistics using data from two publicly accessible databases - one compiled by the Ministry of Finance *(Indicators for the assessment of financial situation of self-government territorial units)* and the other by the Central Statistical Office (*Local Data Bank*). Calculations were performed in R with packages *spdep, maptools* and *shapefiles*. The study demonstrated that the communes of Wielkopolska province of comparable levels of financial self-sufficiency exhibited a moderate tendency to cluster. Clusters of high levels gathered around larger urban centres, especially around Poznań, while clusters of low levels – in economically underdeveloped agricultural south-eastern and northern part of the province.

The next paper, ***Kernel estimation of cumulative distribution function of a random variable with bounded support*** by **A. Baszczyńska,** presents an attempt to reduce the so-called boundary effects, which appear in the estimation of certain functional characteristics of a random variable with bounded support. The methods of the cumulative distribution function estimation, in particular the kernel method, as well as the phenomenon of increased bias estimation in boundary region are discussed. Using simulation methods, the properties of the modified kernel estimator of the distribution function are investigated and an attempt to compare the classical and the modified estimators is made.

S. Wanat, S. Śmiech, and **M. Papież** in their article ***In search of hedges and safe havens in global financial markets*** explore three instrument classes: assets (represented by the S&P500 index), gold and oil prices, and dollar exchange rates. Weekly series of returns of all the instruments from the period January 1995 – June 2015 are analysed. The study is based on conditional correlations between the instruments in different market regimes obtained with the use of copula-DCC GARCH models. It is assumed that different market regimes will be identified by statistical clustering techniques; however, only conditional variances (without conditional covariances) will be taken into account. The author maintain that such an approach has a considerable advantage because it does not require to determine the number of market regimes as it is established by clustering quality measures. The methodology used in the paper makes it possible to treat the relations between instruments symmetrically. In conclusions, they stress that, according to the results obtained in their study, only dollar exchange rates can be treated as a (strong) hedge and a (strong) safe haven for other instruments, while gold and oil are a hedge for assets.

In the book review section, **J. Kordos** presents ***Microeconometrics in Business Management*** by **J. W. Wiśniewski.** Starting with remark that this book introduces the application of micro-econometric methods for modeling various aspects of economic activity for small- to large-sized enterprises, the reviewer highly recommend the book, especially to experts teaching business management. Basic models used in the modeling of the business (single-equation and multiple-equation systems) are introduced whilst a wide range of economic activity including major aspects of financial management, demand for labour, administrative staff and labour productivity is also explored. The book consists of Preface, Acknowledgments, six chapters which end with Conclusion and Bibliography.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 363

SUBMISSION INFORMATION FOR AUTHORS

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

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

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

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

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

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

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

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

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 365–390

SMALL AREA ESTIMATION OF INCOME UNDER SPATIAL SAR MODEL

Jan Kubacki [1], **Alina Jędrzejczak** [2]

ABSTRACT

The paper presents the method of hierarchical Bayes (HB) estimation under small area models with spatially correlated random effects and a spatial structure implied by the Simultaneous Autoregressive (SAR) process. The idea was to improve the spatial EBLUP by incorporating the HB approach into the estimation algorithm. The computation procedure applied in the paper uses the concept of sampling from a posterior distribution under generalized linear mixed models implemented in WinBUGS software and adapts the idea of parameter estimation for small areas by means of the HB method in the case of known model hyperparameters. The illustration of the approach mentioned above was based on a real-world example concerning household income data. The precision of the direct estimators was determined using own three-stage procedure which employs Balanced Repeated Replication, bootstrap and Generalized Variance Function. Additional simulations were conducted to show the influence of the spatial autoregression coefficient on the estimation error reduction. The computations performed by 'sae' package for R project and a special procedure for WinBUGS reveal that the method provides reliable estimates of small area means. For high spatial correlation between domains, noticeable MSE reduction was observed, which seems more evident for HB-SAR method as compared with the traditional spatial EBLUP. In our opinion, the Gibbs sampler, revealing the simultaneous nature of processes, especially for random effects, can be a good starting point for the simulations based on stochastic SAR processes.

Key words: small area estimation (SAE), SAR model, hierarchical Bayes estimation, spatial empirical best linear unbiased predictor.

1. Introduction

Statistical surveys are often designed to provide data that allow reliable estimation for the whole country and larger administrative units such as regions

[1] Centre of Mathematical Statistics, Statistical Office in Łódź. E-mail: j.kubacki@stat.gov.pl.

[2] Institute of Statistics and Demography, University of Łódź; Centre of Mathematical Statistics, Statistical Office in Łódź. E-mail: jedrzej@uni.lodz.pl.

(in Poland – voivodships). Smaller areas are usually not included into sampling designs mainly because of financial and organizational limitations, and the overall sample size is seldom large enough to yield direct estimates of adequate precision for all the domains of interest. In such cases the inferences are connected with large estimation errors which make them unreliable and useless for decision-makers. The estimation errors can be reduced, however, by means of the model-based approach. Moreover, when an evident correlation exists between survey and administrative data, also the bias of the estimates can be reduced.

Small area estimation offers a wide range of methods that can be applied when a sample size is insufficient to obtain high precision by means of conventional direct estimates. The techniques based on small area models - empirical best linear unbiased prediction (EBLUP) as well as empirical and hierarchical Bayes (EB and HB), seem to have distinct advantage over other methods. The model-based approach treats the population values as random and the associated inferences are based on the probability distribution induced by the assumed superpopulation model. One of these techniques is the Spatial EBLUP (Spatial Empirical Best Linear Unbiased Prediction). It is usually based on the assumption that the spatial relationships between domains can be modelled by the simultaneous autoregressive process SAR (see Pratesi and Salvati (2008), p. 114 for better explanation of this term). The method was introduced by Cressie (1991) and is explained in detail in the publications of Saei and Chambers (2003), Pratesi and Salvati (2004, 2005, 2008), Singh et al. (2005), Petrucci and Salvati (2006). The spatial SAR estimation was also applied in SAMPLE project (2010) for the purpose of bootstrap estimation of the MSE for the populations having various spatial autocorrelation levels. Recently, the Spatial EBLUP technique was used in ‘sae’ package (Molina, Marhuenda (2013)) for R-project environment published in CRAN resources. Moreover, some spatial econometric models were discussed in Griffith, D.A., Paelinck, J.H.P. (2011), where MCMC (Markov Chain Monte Carlo) applications for spatial models are presented.

In the paper we compare two approaches to the spatial SAR modelling implemented for small area estimation. Besides the above-mentioned ordinary Spatial EBLUP, we develop a HB model, which is based on the spatial autoregressive structure of random effects incorporated into Bayesian inference. The model will be called the SAR HB model.

In our opinion, HB estimation can be practically appealing with respect to the traditional EBLUP approach. First, the most common method used to fit EBLUP models was the ML method, although maximum likelihood estimators are asymptotic in nature and little is known about their behaviour in small samples. Moreover, when using the HB approach it is possible not only to obtain the point estimates of the parameters, but also approximate their distributions (including the distributions of model variance and random effects). For SAR process, one can also obtain the approximation of spatial autoregression coefficient distribution. It may be helpful in obtaining the model diagnostics, which is a non-trivial problem in the case of linear mixed models.

2. Small area model for spatially correlated random effects based on SAR process.

In the paper a special case of the area level model (type-A model) is discussed, where the parameter of interest is a vector $\mathbf{\theta}$ of size m (where m is the number of small areas), which is related to the direct estimator $\hat{\theta}$ of this quantity by means of the following relationship

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{e} \tag{1}$$

where **e** is a vector of independent sampling errors having mean 0 and diagonal variance matrix $\mathbf{\Psi}$. The parameter $\mathbf{\theta}$ also satisfies the common relationship connected with linear mixed models, which incorporates the spatial correlation between areas. This relationship is as follows

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} \tag{2}$$

where **X** is the matrix of area-dependent auxiliary variables of size $m \times p$, $\boldsymbol{\beta}$ is the vector of regression parameters of size $p \times 1$, **Z** is the matrix ($m \times m$) of known positive constants and **v** is the $m \times 1$ vector of the second order variation. Within the scope of the study it is assumed that the random effects are described by the SAR process. In such a case the vector **v** can be described as

$$\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u} \;\Rightarrow \mathbf{v} = (\mathbf{I}_{\boldsymbol{m}} - \rho\mathbf{W})^{-1}\mathbf{u} \tag{3}$$

where ρ is the parameter of the spatial autoregression and **W** is the spatial weight matrix (of size $m \times m$), which can be defined in many different ways. In the paper the entries of the spatial weight matrix take values in the interval <0,1> and indicate whether the row and column domains are neighbours or not. The additional restriction imposed on **W** is that row elements add up to 1, **u** is the vector of independent error term with zero mean and constant variance σ_u^2 and $\mathbf{I_m}$ is the identity matrix of size $m \times m$. The random effects have the following covariance matrix **G** (also called SAR dispersion matrix)

$$\mathbf{G} = \sigma_u^2[(\mathbf{I}_{\boldsymbol{m}} - \rho\mathbf{W})^T(\mathbf{I}_{\boldsymbol{m}} - \rho\mathbf{W})]^{-1} \tag{4}$$

and the sampling error **e** has the following covariance matrix

$$\mathbf{R} = \mathbf{\Psi} = diag(\psi_i) \tag{5}$$

Further, we will assume that the matrix **Z** is equal to $\mathbf{I_m}$. Thus, using (1), (2) and (3) the model can be described as follows

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_{\boldsymbol{m}} - \rho\mathbf{W})^{-1}\mathbf{u} + \mathbf{e} \tag{6}$$

The covariance matrix for $\hat{\theta}$ is equal to

$$\mathbf{V} = \mathbf{G} + \mathbf{\Psi} \tag{7}$$

Under the model (8) the Spatial EBLUP estimator is equal to (see for example formula (8) in Pratesi and Salvati (2008))

$$\widetilde{\boldsymbol{\theta}}_i = \mathrm{x}_\mathbf{i}^\mathbf{T}\widetilde{\boldsymbol{\beta}} + \boldsymbol{b}_i^T\boldsymbol{G}\boldsymbol{V}^{-\mathbf{1}}(\widehat{\boldsymbol{\theta}} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) \tag{8}$$

where $\widetilde{\boldsymbol{\beta}} = [\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{V}^{-1}\widehat{\boldsymbol{\theta}}$ is the generalized least squares estimator of the regression parameter and $\boldsymbol{b}_i^{\boldsymbol{T}}$ is the *1*×*m* vector (0,…,0,1,0,…0) with 1 in the *i*-th position. This estimator is dependent on σ_u^2 and ρ. These parameters can be obtained by means of the Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) method (where a log likelihood function is used), both applying the Fisher scoring algorithm. The method was implemented, for example, in ‘sae’ package for R-project. More details on this procedure can be found in the SAMPLE project deliverable 22 (2011) in part 2.1.2.

The mean square error for the Spatial EBLUP estimator (8) can be expressed as the sum of four components, which can be given by (see for example formula (2.17) in Singh et al. (2005) or formula (43) in Molina and Marhuenda (2015))

$$mse[\tilde{\theta}_i] = g_{1i} + g_{2i} + 2g_{3i} - g_{4i} \tag{9}$$

where g_1 is connected with uncertainty about the small area estimate and is of order *O(1)*, g_2 is connected with uncertainty about $\widetilde{\boldsymbol{\beta}}$ and is of order $O(m^{-1})$ for large *m*, g_3 is connected with uncertainty about σ_u^2 (or variance components) and g_4 is connected with uncertainty of spatial autocorrelation parameter ρ. The first two components of MSE are given by

$$g_{1i} = \mathbf{b}_\mathbf{i}^\mathbf{T}[\mathbf{G} - \mathbf{G}\mathbf{V}^{-\mathbf{1}}\mathbf{G}]\mathbf{b}_\mathbf{i} \tag{10}$$

$$g_{2i} = \mathbf{b}_\mathbf{i}^\mathbf{T}[\mathbf{I}_\mathbf{m} - \mathbf{G}\mathbf{V}^{-\mathbf{1}}]\mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{V}^{-\mathbf{1}}\mathbf{X})^{-\mathbf{1}}\,\mathbf{X}^\mathbf{T}[\mathbf{I}_\mathbf{m} - \mathbf{V}^{-\mathbf{1}}\mathbf{G}]\mathbf{b}_\mathbf{i} \tag{11}$$

The third element has a more complicated form and for the spatial EBLUP estimator can be expressed by the following equation

$$g_{3i} = trace\{\mathbf{L}_\mathbf{i}\mathbf{V}\mathbf{L}_\mathbf{i}^\mathbf{T}I^{-1}\} \tag{12}$$

where

$$\mathbf{L}_i = \begin{pmatrix} \mathbf{b}_\mathbf{i}^\mathbf{T}[\mathbf{C}^{-\mathbf{1}}\mathbf{V}^{-\mathbf{1}} - \sigma_u^2\mathbf{C}^{-\mathbf{1}}\mathbf{V}^{-\mathbf{1}}\mathbf{C}^{-\mathbf{1}}\mathbf{V}^{-\mathbf{1}}] \\ \mathbf{b}_\mathbf{i}^\mathbf{T}[\mathbf{A}\mathbf{V}^{-\mathbf{1}} - \sigma_u^2\mathbf{C}^{-\mathbf{1}}\mathbf{V}^{-\mathbf{1}}\mathbf{A}\mathbf{V}^{-\mathbf{1}}] \end{pmatrix} \tag{13}$$

$$\mathbf{C} = (\mathbf{I}_\mathbf{m} - \rho\mathbf{W})^T(\mathbf{I}_\mathbf{m} - \rho\mathbf{W})$$

$$\mathbf{A} = -\sigma_u^2\mathbf{C}^{-\mathbf{1}}\frac{\partial\mathbf{C}}{\partial\rho}\mathbf{C}^{-\mathbf{1}} = -\sigma_u^2\mathbf{C}^{-\mathbf{1}}(-\mathbf{W} - \mathbf{W}^\mathbf{T} + 2\rho\mathbf{W}^\mathbf{T}\mathbf{W})\mathbf{C}^{-\mathbf{1}}$$

and I^{-1} is the Fisher information matrix inverse. It depends on σ_u^2 and ρ and its elements can be expressed as

$$I(\sigma_u^2, \rho) = \begin{pmatrix} I_{\sigma_u^2\sigma_u^2} & I_{\sigma_u^2\rho} \\ I_{\rho\sigma_u^2} & I_{\rho\rho} \end{pmatrix} \tag{14}$$

and their elements are given by

$$I_{\sigma_u^2\sigma_u^2} = \tfrac{1}{2} trace\{\mathbf{V^{-1}C^{-1}V^{-1}C^{-1}}\} \tag{15}$$

$$I_{\sigma_u^2\rho} = I_{\rho\sigma_u^2} = \tfrac{1}{2} trace\{\mathbf{V^{-1}AV^{-1}C^{-1}}\} \tag{16}$$

$$I_{\rho\rho} = \tfrac{1}{2} trace\{\mathbf{V^{-1}AV^{-1}A}\} \tag{17}$$

The last term g_4 can be expressed by

$$g_{4i} = \frac{1}{2}\sum_{k=1}^{2}\sum_{l=1}^{2} \mathbf{b_i^T} \mathbf{\Psi} \boldsymbol{V}^{\mathbf{-1}} \frac{\partial^2 V(\omega)}{\partial\omega_k \partial\omega_l} \boldsymbol{V}^{\mathbf{-1}}\mathbf{\Psi} I_{\mathrm{kl}}^{-1}(\omega)\mathbf{b_i} \tag{18}$$

where $\omega_1 = \sigma_u^2$, $\omega_2 = \rho$ and the second derivatives can be expressed as

$$\frac{\partial^2 \mathbf{V}(\omega)}{\partial(\sigma_u^2)^2} = 0_{m\times m}$$

$$\frac{\partial^2 \mathbf{V}(\omega)}{\partial\sigma_u^2\partial\rho} = \frac{\partial^2 \mathbf{V}(\omega)}{\partial\rho\partial\sigma_u^2} = -\mathbf{C^{-1}} \frac{\partial \mathbf{C}}{\partial\rho}\mathbf{C^{-1}}$$

$$\frac{\partial^2 \mathbf{V}(\omega)}{\partial\rho^2} = 2\sigma_u^2 \mathbf{C^{-1}} \frac{\partial \mathbf{C}}{\partial\rho}\mathbf{C^{-1}} \frac{\partial \mathbf{C}}{\partial\rho}\mathbf{C^{-1}} - 2\sigma_u^2 \mathbf{C^{-1}W^TWC^{-1}}$$

The above relationships were obtained under the assumption that the variance components can be described by the spatial SAR process. The spatial hierarchical Bayes model can be formulated in the manner analogous to the model (10.3.1) from Rao (2003), but in the model definition the spatial dependence between domains, determining the structure of the SAR process, should be specified (via ρ and spatial weight matrix **W**). Contrary to the other parameters, for the parameter of spatial autoregression ρ it is difficult to elicit an informative prior, either subjectively or from previous data. A uniform prior which assigns equal weight to all values of the spatial parameter seems unreasonable, as most of the SAR models based on real data sets reported in the literature have yielded moderate or large (positive) ρ estimates. When the values of ρ coefficients are treated as constants (they can be obtained from the previous Spatial EBLUP estimation), the model can be expressed as follows:

(i) $\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_u^2 \sim N(\boldsymbol{\theta}, \mathbf{\Psi})$

(ii) $\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma_u^2, \rho \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_u^2[(\mathbf{I_m} - \rho\mathbf{W})^{\boldsymbol{T}}(\mathbf{I_m} - \rho\mathbf{W})]^{\mathbf{-1}})$

(iii) $f(\boldsymbol{\beta}) \propto 1$

(iv) $\sigma_u^2 \sim G^{-1}(a, b)$ (19)

It has also been assumed that the initial parameters a and b of the Gamma prior describing the σ_u^2 distribution are both known. In the study they were obtained on the basis of the EBLUP models estimated for 16 regions and two time periods, which may be a good approximation of σ_u^2 variability. A similar approach was used in the previous work by Kubacki (2012) in the context of the traditional Fay-Heriott model, where the distribution of σ_u^2 was obtained from the ordinary regression. The model (19) applied in the paper is somewhat similar to that presented in Gharde, Rai and Jaggi (2013), but is an area-level (not unit-level) model and includes additional assumptions on σ_u^2 prior distribution as well as on direct relationships between the model estimates and the values $\mathbf{W}$, ρ and σ_u^2 (to be discussed further).

3. Illustration

The application of the proposed procedures to the Polish income data consisted of the following steps:

1. Direct estimation of average per capita income for counties.
2. Estimation of standard errors of the direct estimates.
3. Specification of small area with spatially correlated random effects for counties *(powiats)* [formulas (2).(3)].
4. Model-based estimation for counties based on EBLUP and Spatial EBLUP procedures [formula (9)].
5. Formulation of hierarchical Bayes model incorporating spatially correlated random effects for counties [formula (19)].
6. Implementation of computations for HB spatial model (to be described as a separate paragraph).

The variable of interest was household available income. We were particularly interested in the estimation of its average per capita value for counties, i.e. NUTS-4 areas according to the Eurostat classification. The basis of the direct estimation was the individual data coming from the Polish Household Budget Survey (HBS).

The precision of direct estimates is usually computed by means of the Balanced Repeated Replication (BRR) technique. This method is valid when the sample for each county is composed of two subsamples that allow constructing the replications called half-samples. However, in the case of extremely small samples there might not be two sub-samples for each county, so the simple bootstrap method must be used instead. Another difficulty arises when for a particular county there is no information available about the variable of interest. In such a case the Generalized Variance Function (GVF), traditionally used to smooth out the uncertainty of the design-based variance estimates, can be helpful. It is worth mentioning that the previous investigations of the authors revealed no underestimation in the bootstrap and GVF-based estimates of precision (see Kubacki and Jędrzejczak (2012). Therefore, using such an approximation may properly reflect the precision for all counties.

In the applied small area models, two auxiliary variables coming from the Polish tax register POLTAX were specified as covariates. They include: the average salary and the average universal health insurance premium contribution, both determined by dividing the sums of the respective totals by de facto population sizes for particular NUTS-4 units.

To provide the entries of the spatial weight matrix W, necessary for the spatial model specification, the digital maps for Polish counties were used. During the computations (using 'spdep' package for R-project environment) sub-maps for regions were automatically generated, which simplified the visualization of the results.

When formulating the HB model for counties one should determine the prior distribution for the model variance σ_u^2. In the paper, ordinary EBLUP and Spatial EBLUP estimation results were used to obtain the empirical distribution of this variance. The results of these computations were summarized in Figure 1.

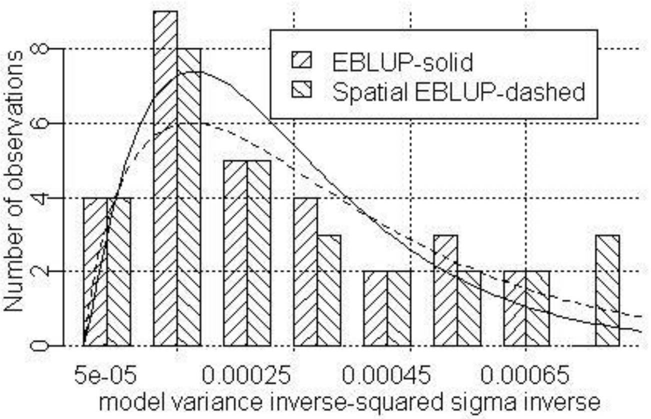


Figure 1. Empirical distributions of model variance inverse σ_u^{-2}obtained for small area models (EBLUP and Spatial EBLUP) of household per capita available income by NUTS-4 - counties in Poland.

The distributions of model errors were found similar for both EBLUP and Spatial EBLUP models. The distributions of inverse model variances (Figure 1) are both positively skew and can be approximated by gamma priors as it was assumed in the hierarchical model (19). Slight differences, which can be observed in Figure 1, seem not to have significant influence on the variability of the estimates which were obtained using EBLUP and Spatial EBLUP techniques. On the other hand, one should be careful while dealing with more specific types of income and some further simulations should be made for them.

4. Results and discussion

The results presented for the Silesian region and for the year 2004 (Tables 1, 2 and 3) show that in the case of ρ values significantly different from zero (in the case presented here it is equal to 0.681), estimation based on spatial models may significantly reduce the estimation error. This effect is more evident for HB estimator. For non-spatial models some consistency between relative estimation errors and random effect values is observed (see Figure 3). Please note that the random effects arise from the residuals obtained from generalized regression models, which can be easily derived from the second component of equation (8).

Examining the diagnostic graphs obtained for Gibbs sampler simulations, one can easily notice that in the case where ordinary HB scheme is used normality for the model estimates (denoted mu – see Figure 4), no autocorrelation (see Figure 5) and relative stability of simulations run are observed. Similar results were also presented in the authors' previous works (see: Kubacki (2012)). It can be noted that autocorrelation plots show the correlation between the values coming from the simulation obtained for the iteration k and the iteration $k+t$, where t indicates the lag between k and $k+t$ value. The absence of autocorrelation on the plot indicates that for $t=0$ the correlation is equal to 1, and for further lags it is close to zero.

However, for the spatial version of HB estimator, some autocorrelation (see Figure 9), and sometimes lack of normality (see Figure 8) in model estimates is observed. In the case considered here, this is partially due to serious direct estimation errors (as it was observed for Mikołowski, Pszczyński and Bieruńsko-Lędziński counties). The consistency between Spatial EBLUP and HB SAR-based estimates (see Figure 6), between the estimation error and the obtained random effects (see Figure 7), is relatively weaker, but it is achieved for the considered case. Random effects, obtained for Spatial EBLUP estimator, are presented on the map below (see Figure 10). Here, some regularity between the absolute values of random effects and the geographic location of the county is observed. This regularity is connected with their central or peripheral location, which means that the central part of the considered region dominates over the rest of the region, and no isolated counties (islands) in the considered region are observed.

The comparison of relative estimation error (REE) distribution and the distribution of reduction of this error shows that all the considered model-based techniques are significantly more efficient than the corresponding direct ones (Figure 11). In fact, all the considered techniques present similar efficiency and have similar REE reduction structure (Figure 12) - only HB-SAR performs slightly better, as compared to the other model-based techniques. This regularity can also be observed, when a comparison between a spatial and a respective non-spatial model is made (Figure 13).

Table 1. Estimation results for per capita available income by counties in the Silesian region obtained using direct estimation method, EBLUP method (REML technique) and HB method (Gibbs sampler).

County (NUTS-4 unit)	Available income							
	Direct estimation		EBLUP estimation – REML			HB estimation		
	Parameter estimate	REE %	Parameter estimate	REE %	Random effects	Parameter estimate	REE %	Random effects
będziński	821.59	5.82	784.71	4.28	23.81	789.52	4.40	27.52
bielski	781.94	1.22	778.63	1.20	53.92	779.40	1.19	54.18
cieszyński	762.78	4.99	734.21	3.92	29.12	738.61	3.97	33.71
częstochowski	570.65	6.54	590.44	5.00	-21.00	586.37	5.08	-21.73
gliwicki	693.20	11.14	706.85	5.19	-3.38	705.97	5.61	-3.95
kłobucki	539.04	8.00	574.29	5.59	-28.02	568.43	5.77	-30.27
lubliniecki	596.73	4.14	610.82	3.63	-34.12	608.03	3.63	-34.85
mikołowski	796.64	15.83	793.96	5.01	0.25	796.39	5.41	-0.06
myszkowski	613.46	7.74	622.49	5.35	-5.92	620.01	5.42	-5.43
pszczyński	629.75	12.14	724.15	5.41	-23.85	719.27	5.79	-30.21
raciborski	758.34	4.50	716.66	3.88	52.76	722.32	3.94	59.80
rybnicki	783.98	8.18	764.15	4.69	7.12	766.10	5.08	7.97
tarnogórski	671.46	5.07	686.73	3.92	-19.45	684.87	3.99	-21.04
bieruńsko-lędziński	625.85	11.69	764.96	4.93	-38.35	757.22	5.32	-49.19
wodzisławski	855.72	4.24	812.73	3.53	48.34	819.80	3.66	54.13
zawierciański	671.04	7.36	674.02	4.89	-1.80	672.20	5.06	-2.18
żywiecki	730.75	1.80	725.43	1.75	45.63	726.56	1.75	47.68
Bielsko-Biała city	792.14	7.38	771.68	4.42	8.84	774.27	4.66	9.78
Bytom city	705.56	1.29	705.28	1.27	4.93	705.23	1.26	5.59
Chorzów city	656.28	2.50	666.95	2.34	-58.78	664.70	2.37	-61.08
Częstochowa city	771.36	10.35	715.50	5.12	12.94	719.17	5.57	16.93
Dąbrowa Górnicza city	777.52	4.38	782.61	3.47	-6.48	783.03	3.47	-8.58
Gliwice city	745.60	5.50	761.17	3.91	-13.66	760.25	3.96	-16.57
Jastrzębie-Zdrój city	748.66	5.52	774.86	3.90	-22.65	771.66	4.01	-28.39
Jaworzno city	748.49	6.85	780.11	4.22	-17.74	778.39	4.42	-22.11
Katowice city	859.53	1.13	859.17	1.12	5.66	859.38	1.12	1.51
Mysłowice city	813.19	2.41	810.39	2.23	10.79	811.09	2.20	8.73
Piekary Śląskie city	744.14	13.14	741.89	5.31	0.35	741.68	5.88	-0.41
Ruda Śląska city	671.92	15.46	772.90	5.09	-13.82	768.99	5.75	-19.80
Rybnik city	763.03	1.45	764.52	1.41	-17.86	764.20	1.41	-20.39
Siemianowice Śląskie city	915.69	9.15	774.56	4.93	29.68	783.96	5.63	38.39
Sosnowiec city	818.88	7.44	803.93	4.39	5.95	806.06	4.65	5.59
Świętochłowice city	686.82	9.03	708.28	4.90	-8.24	706.42	5.28	-10.05
Tychy city	832.03	0.07	832.03	0.07	-4.80	832.02	0.07	-9.04
Zabrze city	703.72	0.21	703.74	0.21	-15.66	703.73	0.21	-15.83
Żory city	830.68	8.20	781.97	4.53	15.50	786.79	4.88	18.58

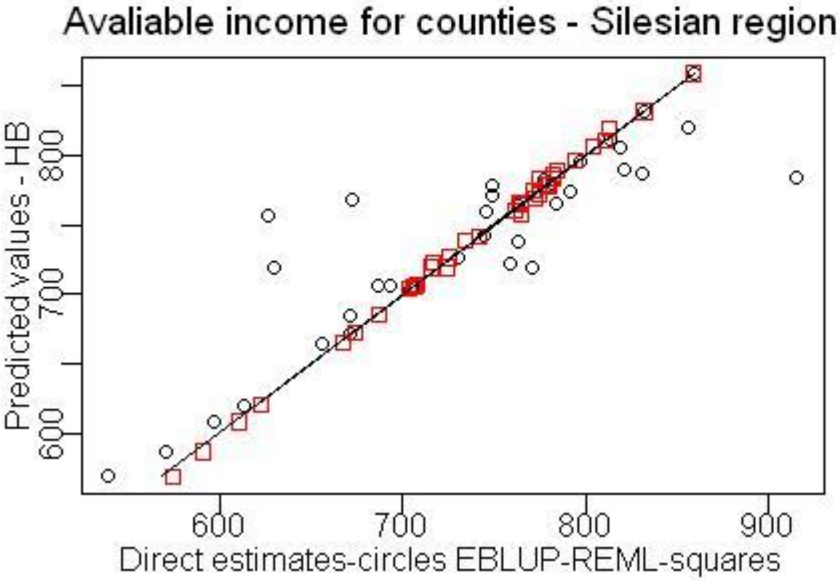


Figure 2. Observed per capita available income (direct estimates - black circles, EBLUP estimates - red squares) vs. predicted values estimated under hierarchical Bayes model for counties in the Silesian region.

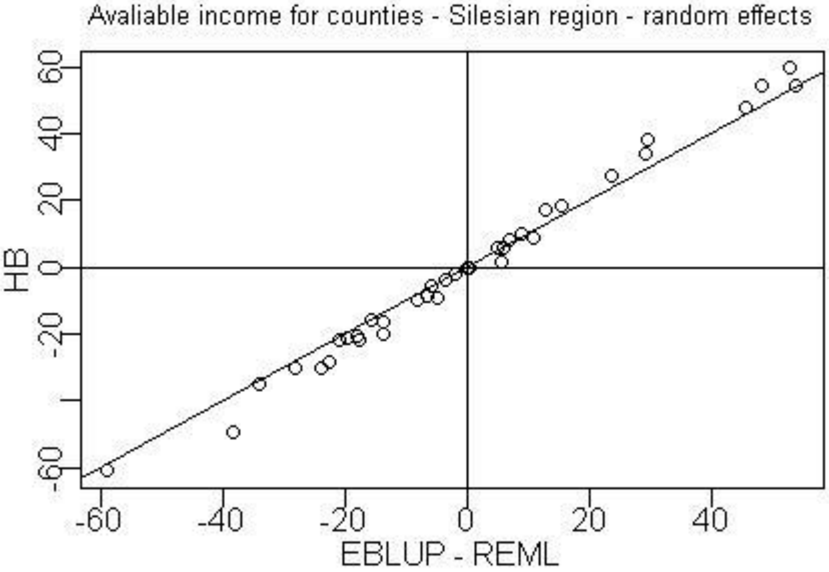


Figure 3. EBLUP vs. HB estimates of random effects for small area models of per capita available income, obtained for counties in the Silesian region.

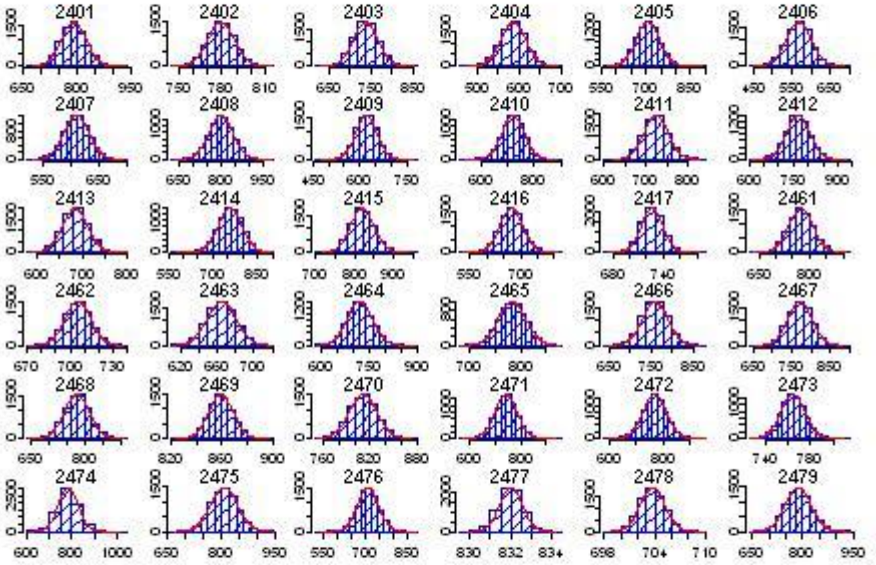


Figure 4. A posteriori distributions of per capita available income for counties in the Silesian region obtained by MCMC simulation (Gibbs sampler) under conventional HB model.

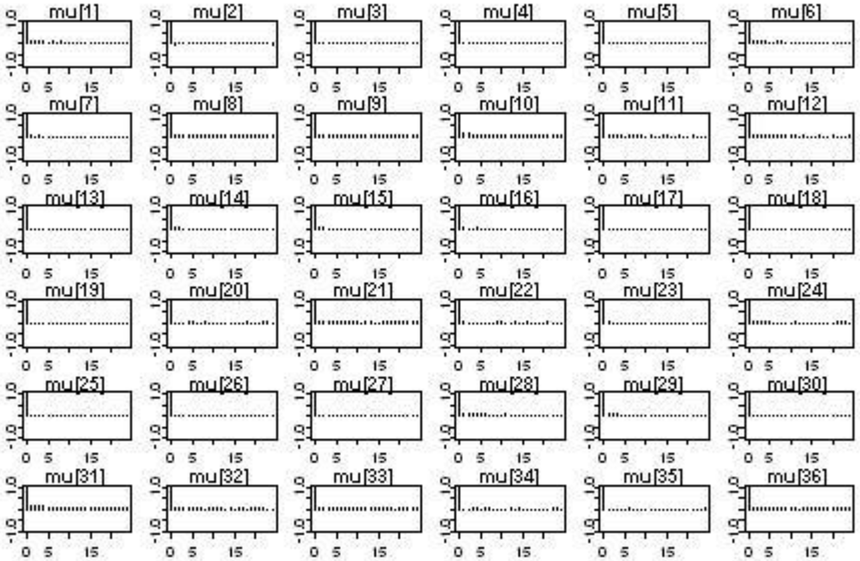


Figure 5. Autocorrelations of model estimates for per capita available income obtained for counties in the Silesian region by MCMC simulation using Gibbs sampler.

Table 2. Estimation results for per capita available income by counties in the Silesian region obtained using direct estimation method and Spatial EBLUP method (REML technique) and HB-SAR method (Gibbs sampler).

County (NUTS-4 unit)	Available income							
	Direct estimation		EBLUP estimation – Spatial REML			HB-SAR estimation		
	Parameter estimate	REE %	Parameter estimate	REE %	Random effects	Parameter estimate	REE %	Random effects
będziński	821.59	5.82	781.77	3.95	11.06	785.61	3.87	22.85
bielski	781.94	1.22	779.45	1.20	59.60	780.12	1.16	67.81
cieszyński	762.78	4.99	746.23	3.51	42.66	748.93	3.51	52.75
częstochowski	570.65	6.54	584.70	4.74	-36.01	582.12	4.68	-31.99
gliwicki	693.20	11.14	711.72	4.55	-5.22	714.97	3.88	5.50
kłobucki	539.04	8.00	569.85	5.37	-41.62	566.06	5.40	-38.91
lubliniecki	596.73	4.14	603.87	3.58	-41.71	601.72	3.54	-37.00
mikołowski	796.64	15.83	788.05	4.24	4.11	798.48	5.15	22.67
myszkowski	613.46	7.74	618.68	4.95	-22.87	617.30	4.89	-17.47
pszczyński	629.75	12.14	734.98	4.50	-0.23	747.92	4.78	20.42
raciborski	758.34	4.50	710.69	4.00	49.21	718.45	3.90	63.98
rybnicki	783.98	8.18	780.23	4.20	17.16	782.12	4.34	26.94
tarnogórski	671.46	5.07	683.65	3.66	-24.50	684.51	3.56	-16.23
bieruńsko-lędziński	625.85	11.69	770.56	4.07	-20.52	776.08	4.08	-6.79
wodzisławski	855.72	4.24	813.85	3.47	42.01	819.97	3.56	56.09
zawierciański	671.04	7.36	673.66	4.56	-12.07	671.67	4.64	-6.88
żywiecki	730.75	1.80	729.35	1.75	51.79	729.74	1.69	59.34
Bielsko-Biała city	792.14	7.38	799.78	3.96	41.89	797.44	4.03	47.44
Bytom city	705.56	1.29	703.46	1.27	-6.94	703.80	1.26	0.81
Chorzów city	656.28	2.50	672.64	2.31	-62.83	668.96	2.28	-58.88
Częstochowa city	771.36	10.35	682.88	5.31	-19.93	683.66	5.42	-11.78
Dąbrowa Górnicza city	777.52	4.38	781.40	3.48	-0.41	780.42	3.35	6.72
Gliwice city	745.60	5.50	753.29	3.78	-17.26	753.28	3.57	-9.28
Jastrzębie-Zdrój city	748.66	5.52	795.11	3.67	-1.68	789.51	3.61	0.94
Jaworzno city	748.49	6.85	784.13	3.88	-10.23	782.63	3.92	-3.53
Katowice city	859.53	1.13	857.99	1.12	9.63	859.12	1.07	19.47
Mysłowice city	813.19	2.41	806.30	2.22	11.49	806.86	2.15	20.27
Piekary Śląskie city	744.14	13.14	738.50	4.50	-12.01	738.26	4.86	-4.48
Ruda Śląska city	671.92	15.46	765.34	4.29	-25.23	758.58	4.60	-23.85
Rybnik city	763.03	1.45	768.11	1.41	-10.11	767.02	1.39	-3.12
Siemianowice Śląskie city	915.69	9.15	760.25	4.35	7.50	768.19	5.00	23.23
Sosnowiec city	818.88	7.44	807.21	3.85	8.00	805.53	4.00	14.55
Świętochłowice city	686.82	9.03	692.04	4.51	-29.30	689.37	4.76	-24.46
Tychy city	832.03	0.07	832.03	0.07	10.70	832.02	0.07	19.18
Zabrze city	703.72	0.21	703.73	0.21	-16.69	703.71	0.21	-9.18
Żory city	830.68	8.20	770.40	4.19	8.08	773.90	4.45	19.50

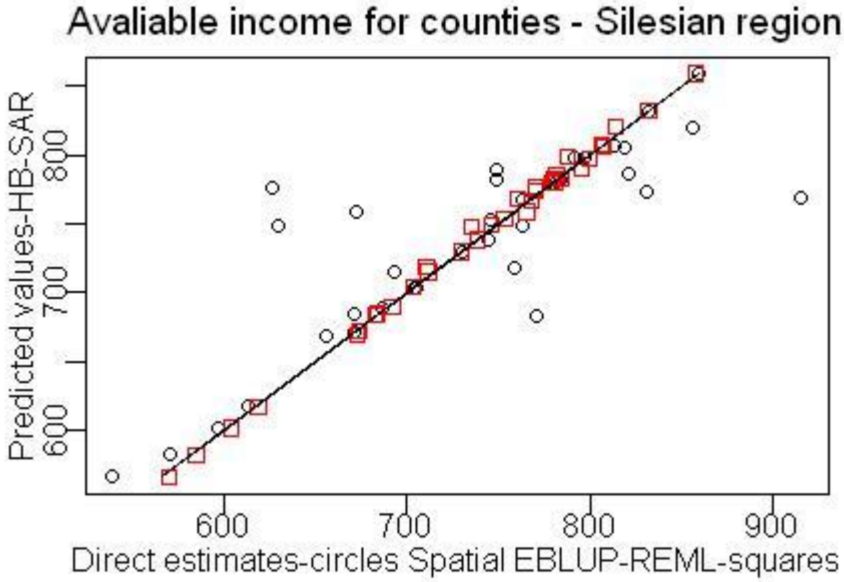


Figure 6. Observed per capita available income (direct estimates - black circles, EBLUP estimates - red squares) vs. predicted values estimated under hierarchical Bayes with SAR relationships between areas, for counties in the Silesian region.

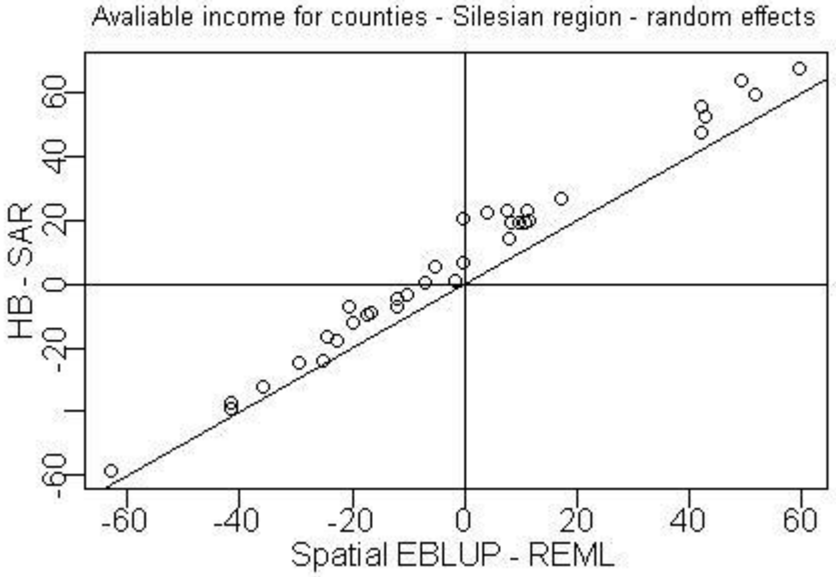


Figure 7. EBLUP vs. HB-SAR estimates of random effects for small area models of per capita available income, obtained for counties in the Silesian region.

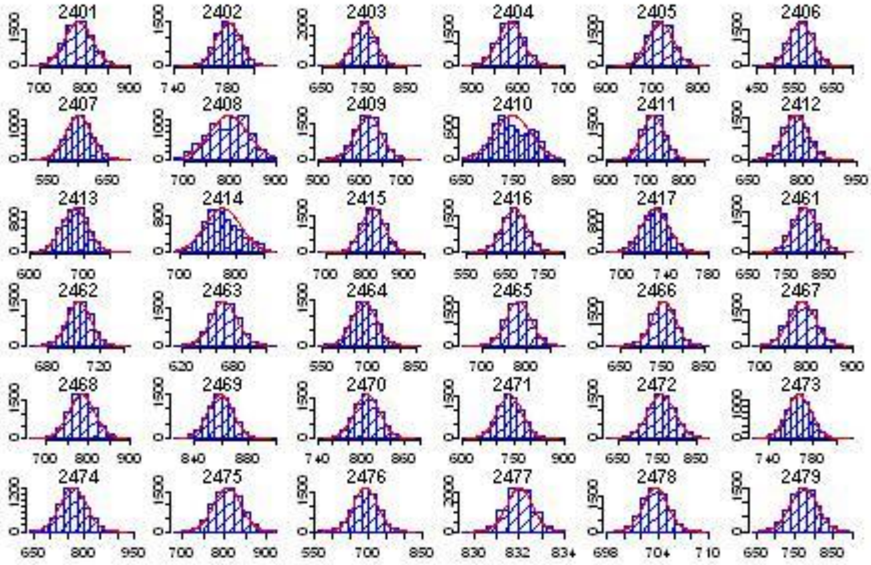


Figure 8. A posteriori distributions of per capita available income for counties in the Silesian region obtained by MCMC simulation (Gibbs sampler) under HB-SAR model.

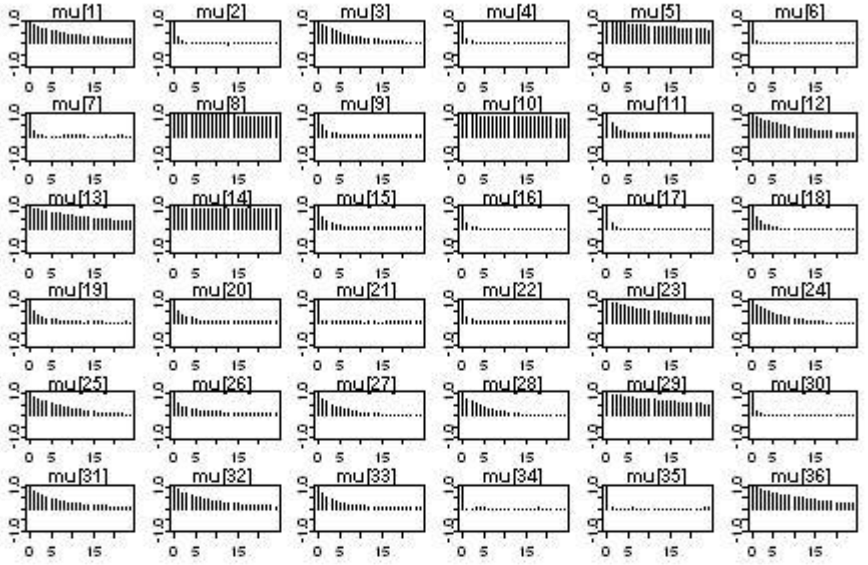


Figure 9. Autocorrelations of model estimates for per capita available income obtained for counties in the Silesian region by MCMC with HB-SAR simulation using Gibbs sampler.

Table 3. Relative estimation error reduction and spatial gain for estimation errors calculated for available income estimates using EBLUP (both ordinary and spatial) method using REML technique and HB (both ordinary and SAR version) using Gibbs sampler.

County (NUTS-4 unit)	Relative estimation error reduction				Spatial estimation error reduction	
	EBLUP	HB	Spatial EBLUP	HB-SAR	Spatial EBLUP	HB-SAR
będziński	1.361	1.322	1.476	1.504	1.084	1.137
bielski	1.013	1.026	1.015	1.049	1.002	1.023
cieszyński	1.273	1.257	1.422	1.420	1.117	1.130
częstochowski	1.307	1.286	1.379	1.396	1.055	1.086
gliwicki	2.145	1.987	2.448	2.869	1.141	1.444
kłobucki	1.430	1.386	1.488	1.482	1.041	1.069
lubliniecki	1.140	1.138	1.156	1.167	1.013	1.025
mikołowski	3.159	2.924	3.737	3.074	1.183	1.051
myszkowski	1.446	1.429	1.563	1.583	1.081	1.108
pszczyński	2.242	2.097	2.698	2.541	1.203	1.211
raciborski	1.161	1.144	1.125	1.155	0.969	1.010
rybnicki	1.746	1.610	1.950	1.886	1.116	1.172
tarnogórski	1.294	1.272	1.385	1.423	1.071	1.118
bieruńsko-lędziński	2.374	2.199	2.871	2.867	1.209	1.303
wodzisławski	1.199	1.156	1.221	1.190	1.019	1.029
zawierciański	1.506	1.454	1.614	1.588	1.072	1.092
żywiecki	1.024	1.024	1.028	1.065	1.004	1.040
Bielsko-Biała city	1.669	1.582	1.864	1.829	1.117	1.156
Bytom city	1.014	1.021	1.011	1.019	0.997	0.998
Chorzów city	1.066	1.051	1.081	1.093	1.014	1.040
Częstochowa city	2.022	1.858	1.949	1.908	0.964	1.027
Dąbrowa Górnicza city	1.263	1.262	1.261	1.306	0.999	1.035
Gliwice city	1.407	1.388	1.456	1.543	1.035	1.111
Jastrzębie-Zdrój city	1.417	1.375	1.504	1.531	1.062	1.113
Jaworzno city	1.624	1.550	1.767	1.749	1.088	1.128
Katowice city	1.014	1.014	1.010	1.053	0.996	1.039
Mysłowice city	1.077	1.095	1.085	1.117	1.007	1.020
Piekary Śląskie city	2.473	2.234	2.921	2.706	1.181	1.211
RudaŚląska city	3.035	2.689	3.602	3.359	1.187	1.250
Rybnik city	1.027	1.029	1.032	1.046	1.005	1.017
Siemianowice Śląskie city	1.857	1.626	2.103	1.831	1.132	1.126
Sosnowiec city	1.693	1.598	1.932	1.858	1.141	1.162
Świętochłowice city	1.844	1.710	2.000	1.897	1.085	1.109
Tychy city	1.000	1.012	1.000	0.994	1.000	0.983
Zabrze city	1.000	0.991	1.000	0.991	1.000	1.001
Żory city	1.811	1.682	1.959	1.845	1.082	1.097

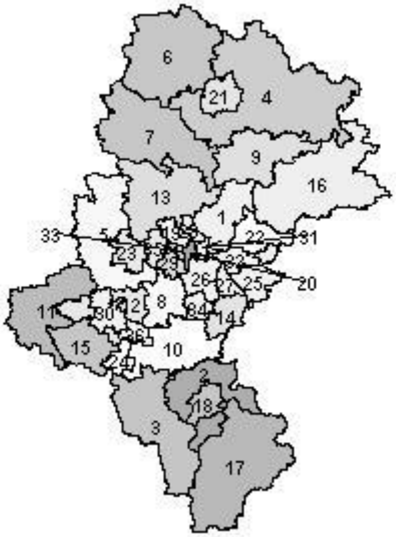


Figure 10. Choropleth map of counties in the Silesian region presenting the absolute values of random effects obtained for per capita available income estimated by Spatial EBLUP estimator (more intense colour means higher absolute random effect).

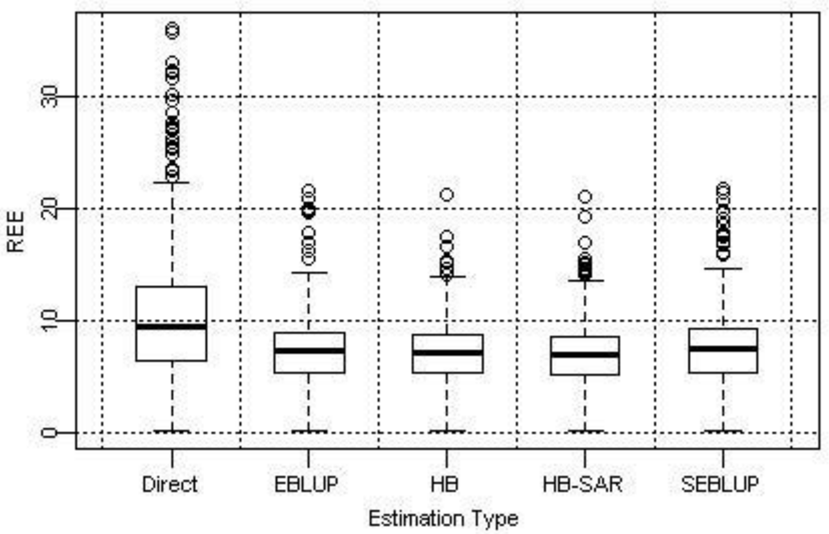


Figure 11. Distribution of relative estimation error for direct estimator, EBLUP (both ordinary and spatial) and HB estimator (ordinary and using SAR relationships) for counties in Poland.

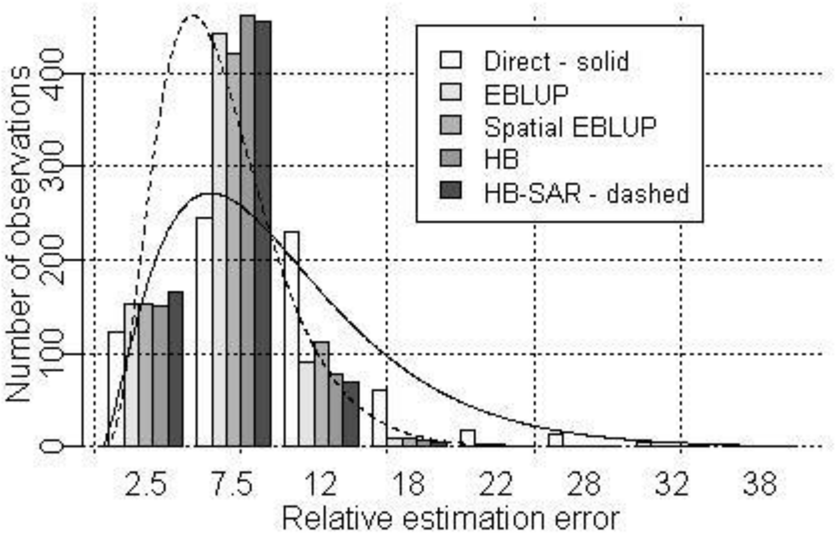


Figure 12. Distribution of relative estimation error for direct estimator, EBLUP (both ordinary and spatial) and for HB estimator (ordinary and using SAR relationships) of per capita available income by NUTS4 in Poland.

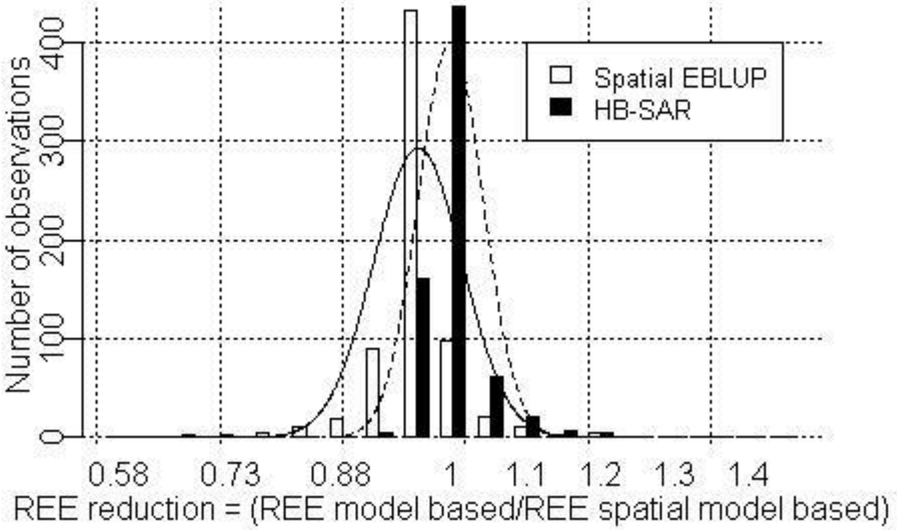


Figure 13. Distribution of relative estimation error reduction due to spatial relationships for EBLUP, Spatial EBLUP and HB (ordinary and using SAR relationships) obtained for per capita available income by NUTS4 in Poland.

This comparison reveals that HB-SAR technique has slightly better performance than its corresponding Spatial EBLUP estimator. However, it should be noted here that only some of the considered models have large enough parameter of spatial autoregression ρ. For most of the regions this measure is below 0.5 and sometimes the ρ coefficient is negative, which may mean that the REE reduction due to the spatial relationships may be not very significant. In order to verify this assumption, we conduct the simulation study, which has to resolve how the ρ value affects the efficiency of spatial estimation. The simulation was prepared in such a way that only MSE values were changed according to the a priori ρ value. The rest of the estimation process, i.e. point estimation procedure for Spatial EBLUP, spatial weight matrix, direct estimates and explanatory variables, remains unchanged. This is in contrast with the experiment conducted in SAMPLE project, where some arbitrary assumptions concerning direct estimates (with fixed values of direct estimation precision), the fixed value of σ_u^2 and explanatory variables were made. In our opinion this may slightly change the real-world conditions and may affect such simulation results.

It is assumed in our experiment that ρ value has four a priori values equal to 0.95, 0.75, 0.25 and -0.50. In our case we also observe that setting the ρ parameter can improve the performance of the estimates. However, that was clear only for SAR-based HB estimator. For spatial EBLUP technique the influence of ρ value is ambiguous (see Figure 14). This is evident when the analysis of REE reduction due to the spatial relationships is made (see Figure 15). Here, for higher ρ values some reduction of REE values obtained for spatial version of HB estimator is observed. For Spatial EBLUP this reduction is rather not the rule (the average spatial REE reduction is slightly below 1). Similar results were also obtained in the work published recently by Gharde, Rai, Jaggi (2013). The authors reach the conclusion that “there is % gain in efficiency in Spatial HB (SHB) approach with respect to SEBLUP approach”. It should also be noted that for lower ρ values this gain obtained in our simulation is not significant, even for HB-SAR method.

The simulation results conducted for HB-SAR method reveals also some interesting properties of the obtained stochastic processes generated by Gibbs sampler. It is related to the level of ρ values. When the ρ value is high, the obtaining process for random effect v reveals a characteristic trace, which reveals the simultaneous nature of this process for all v random effects. Their nature has a typical autoregressive run, which becomes evident when Hurst exponent is determined for such a process (using aggvarFit function from ‘fArma’ package for R-project environment). When ρ value is equal to 0.95, this is practically the rule that the process of v has an autoregressive nature (with Hurst exponent higher than 0.9), which is in contrast to the trace of the process for random effects u in ordinary HB simulations (where Hurst exponents for most cases are considerably lower – see Fig. 16).

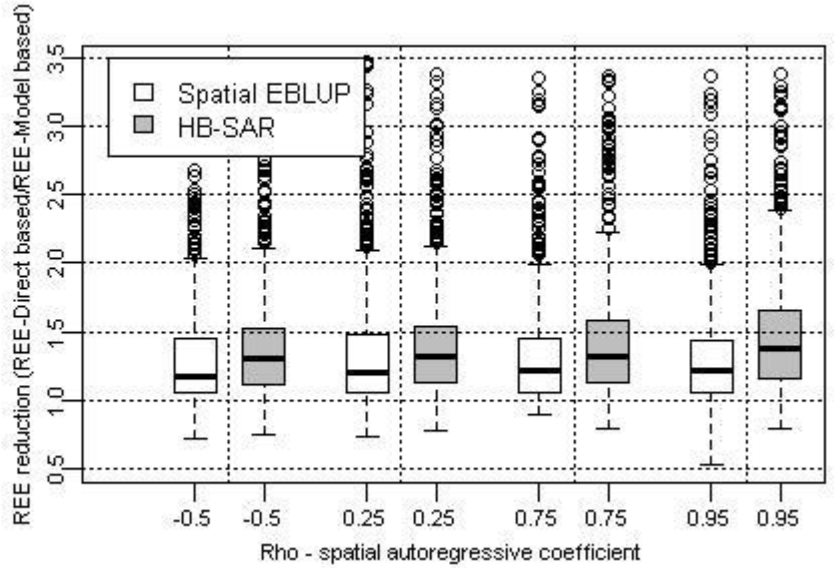


Figure 14. Relative estimation error reduction for Spatial EBLUP and HB-SAR estimators of per capita available income in Poland by counties, for different a priori spatial autoregressive coefficients.

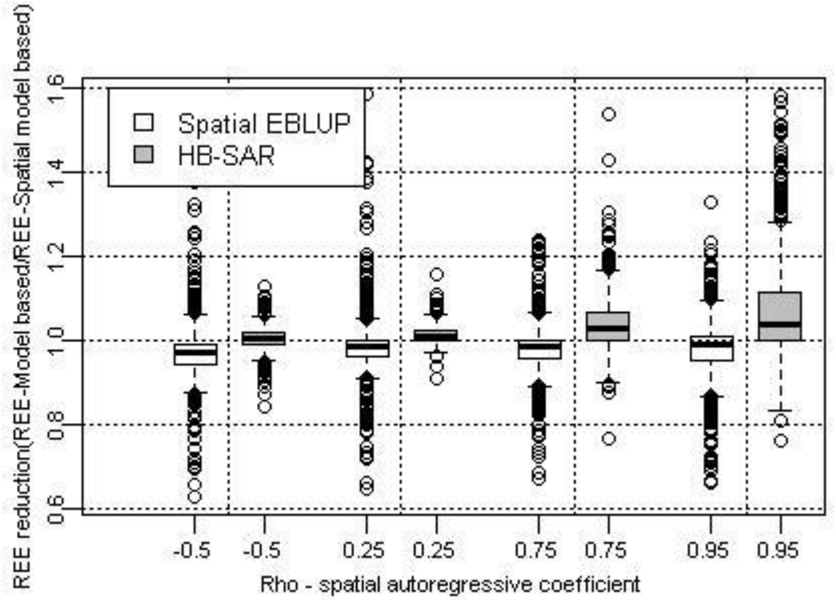


Figure 15. Relative estimation error reduction due to spatial relationships for Spatial EBLUP and HB-SAR estimators of per capita available income in counties, for different a priori spatial autoregressive coefficients.

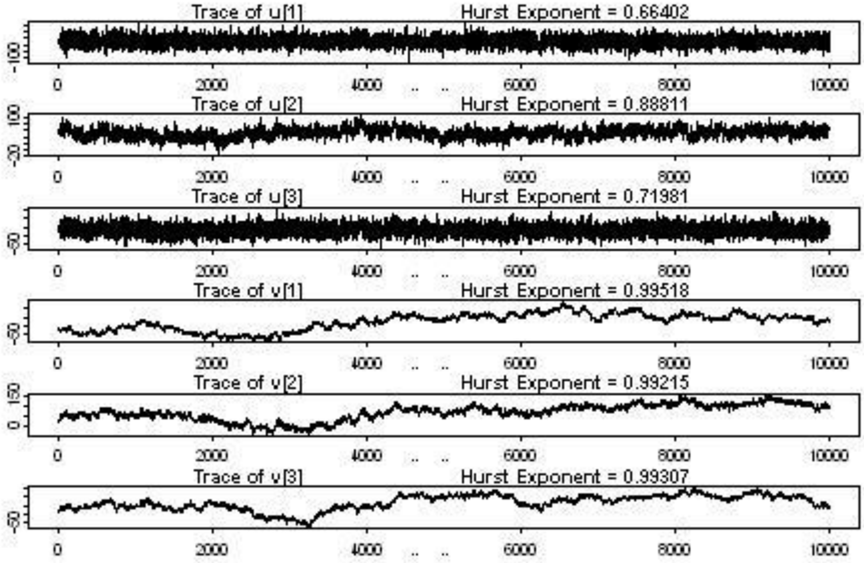


Figure 16. Trace of results for Gibbs sampler simulation obtained for first 3 values of random effects (u description) obtained for ordinary HB method and first 3 values of random effects (v description) obtained for HB method (using SAR relationships) assuming (for v values) that the spatial autoregressive coefficient is equal to 0.95.

From these results, it can be concluded that the Gibbs sampler reproduces the simultaneous nature of the SAR process in a proper way. However, it is still questionable whether the Markov Chain inference can be done in such situations (mainly because of autocorrelation and lack of stability). To overcome this difficulty, some other simulation techniques, like that shown in De Oliveira, V., Jin Song, J., (2008), would be advisable. A non-iterative Monte Carlo algorithm based on factoring the posterior distribution and the adaptive rejection Metropolis sampling (ARMS) proposed by Gilks, Best and Tan (1995) can also be a reasonable choice. The comparison of these two techniques may reveal whether the MCMC approach is valid for the SAR-based simulation conditions. It seems that the Gibbs sampler can be a good starting point for obtaining such simulations. Moreover, for lower ρ values, the autoregressive nature of the process is rather small, and because of this it can be useful in practice for moderate ρ values, as it was shown for the Silesian region in our paper.

5. Conclusions

The paper shows a procedure of efficient estimation for small areas based on the application of the hierarchical Bayes approach to the general linear mixed model with spatially correlated random effects. In particular, the spatial Simultaneous Autoregressive Process, using spatial neighbourhood as auxiliary information, was incorporated into the estimation process. The efficiency of the proposed method was proven on the basis of real-world examples prepared for the Polish data coming from the Household Budget Survey and the tax register. The comparison of relative estimation error distribution and REE reduction shows that all the considered model-based techniques are significantly more efficient than the direct estimation one, however HB-SAR technique shows slightly more REE reduction than the other model techniques. The simulation-based calculations, where some additional assumptions on the spatial autoregressive coefficient were made, also confirm efficiency gains for spatial-based estimators, especially for higher values of this coefficient. However, such a correspondence does not always occur for all the regions, so one should be conscious that for lower ρ values the benefit of using the spatial method may be ambiguous. However, this effect is more evident for Spatial HB method than for Spatial EBLUP technique.

REFERENCES

BIVAND, R., LEWIN-KOH, N., (2013). maptools: Tools for reading and handling spatial objects. R package version 0.8-25, http://CRAN.R-project.org/package=maptools.

BIVAND, R., PIRAS, G., (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics, Journal of Statistical Software, 63, No. 18, pp. 1–36, http://www.jstatsoft.org/v63/i18/.

CRESSIE, N. A .C., (1991). Small-area prediction of undercount using the general linear model, Proceedings of Statistic Symposium 90: Measurement and Improvement of Data Quality, Ottawa, Statistics Canada, pp. 93–105.

GHARDE, Y., RAI, A., JAGGI, S., (2013). Bayesian Prediction in Spatial Small Area Models, Journal of the Indian Society of Agricultural Statistics, 67, pp. 355–362.

GILKS, W. R., BEST, N. G., TAN, K. K. C., (1995). Adaptive Rejection Metropolis Sampling within Gibbs Sampling, Applied Statistics, 4, pp. 455–472.

GOMEZ-RUBIO, V., (2008). Tutorial: Small Area Estimation with R The R User Conference 2008, August 12-14, Technische Universitat Dortmund, Germany.

GRIFFITH, D. A., PAELINCK, J. H. P., (2011). Non-standard Spatial Statistics and Spatial Econometrics, Advances in Geographic Information Science, Springer Berlin Heidelberg.

KUBACKI, J. (2012). Estimation of parameters for small areas using hierarchical Bayes method in the case of known model hyperparameters, Statistics in Transition-new series, 13, No. 2, pp. 261–278.

KUBACKI, J., JĘDRZEJCZAK, A., (2012). The Comparison of Generalized Variance Function with Other Methods of Precision Estimation for Polish Household Budget Survey, Studia Ekonomiczne, 120, pp. 58–69.

MOLINA, I., MARHUENDA, Y., (2013). SAE: Small Area Estimation, R package version 1.0-2 http://CRAN.R-project.org/package=sae.

MOLINA, I., MARHUENDA, Y., (2015). R package sae: Methodology - sae package vignette: https://cran.r-project.org/web/packages/sae/vignettes/sae_methodology.pdf

DE OLIVEIRA, V., JOON, JIN SONG, (2008). Bayesian Analysis of Simultaneous Autoregressive Models, Sankhya, 70-B, No. 2, pp. 323–350.

PETRUCCI, A., SALVATI, N., (2006). Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment, Journal of Agricultural, Biological & Environmental Statistics, 11, No. 2, pp. 169–182, http://dx.doi.org/10.1198/108571106x110531.

PRATESI, M., SALVATI, N., (2004). Spatial EBLUP in agricultural survey. An application based on census data, Working paper no. 256, Universitá di Pisa, Dipartimento di statistica e matematica applicata all'economia.

PRATESI, M., SALVATI, N., (2005). Small Area Estimation: The EBLUP Estimator with Autoregressive Random Area Effects, Working paper n. 261 Pubblicazioni del Dipartimento di statistica e matematica applicata all'economia.

PRATESI, M., SALVATI, N., (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects, Statistical Methods and Applications, 17, No. 1, pp. 113–141, http://dx.doi.org/10.1007/s10260-007-0061-9.

RAO, J. N. K., (2003). Small Area Estimation, John Wiley & Sons, http://books.google.pl/books?id=f8NY6M-5EEwC.

SAEI, A., CHAMBERS, R., (2003). Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects, M03/15, Southampton Statistical Sciences Research Institute, http://eprints.soton.ac.uk/8165/.

SALVATI, N., GIUSTI, C., MARCHETTI, S., PRATESI, M., TZAVIDIS, N., MOLINA, I., MORALES, D., ESTEBAN, M. D., SANTAMARIA, L., MARHUENDA, Y., PEREZ, A., PAGLIARELLA, M., CHAMBERS, R., RAO, J. N. K., FERRETTI, C., (2011). Software on small area estimation, Deliverable 22, http://sample-project.eu/images/stories/deliverables/d22.pdf.

SINGH, B. B., SHUKLA, G. K., KUNDU, D., (2005). Spatio-Temporal Models in Small Area Estimation, Survey Methodology, 31, No. 2, pp. 183–195.

SPIEGELHALTER, D. J., THOMAS, A., BEST, N., LUNN, D., (2003). WinBUGS User Manual, Version 1.4, http://www.mrc-bsu.cam.ac.uk/wp-content/uploads/manual14.pdf.

STURTZ, S., LIGGES, U., GELMAN, A., (2005). R2WinBUGS: A Package for Running WinBUGS from R, Journal of Statistical Software, 12, No. 3.

VOGT, M., (2010). Bayesian Spatial Modeling: Propriety and Applications to Small Area Estimation with Focus on the German Census 2011, Universitat Trier.

APPENDIX

Implementations of computations for EBLUP and hierarchical models

At the computation stage, the WinBUGS (Spiegelhalter et.al. (2003)) and R-project software were used, including the modules ‘sae’ (Molina and Marhuenda (2013)), ‘R2WinBUGS’ (Sturtz, Ligges and Gelman (2005)), ‘coda’, ‘maptools’ (Bivand and Lewin-Koh (2013)), ‘spdep’ (Bivand and Piras (2015)) and ‘MASS’.

The computation scheme applied to obtain the normal and Spatial HB estimates for counties in Poland is the following:

```
model
{for(p in 1 : N)
   {Y[p] ~ dnorm(mu[p], tau[p])
    mu[p] <- alpha[1] + alpha[2] * A[p] + alpha[3] * B[p] + u[p]
    u[p] ~ dnorm(0, precu)   }
  precu ~ dgamma (a0,b0)
  alpha[1] ~ dflat()
  alpha[2] ~ dflat()
  alpha[3] ~ dflat()
  sigmau<-1/precu }
```

For SAR version of hierarchical models the following scheme is used:

```
model
{for(p in 1 : N)
   {Y[p] ~ dnorm(mu[p], tau[p])
    mu[p] <- alpha[1] + alpha[2] * A[p] + alpha[3] * B[p] + v[p]
    u[p] ~ dnorm(0, precu)
    v[p] <- inprod(rho_w[p,1:N],u[1:N])   }
  precu ~ dgamma (a0,b0)
  alpha[1] ~ dflat()
  alpha[2] ~ dflat()
  alpha[3] ~ dflat()
  sigmau<-1/precu }
```

In the notation presented above, the symbol Y[p] stands for the direct estimates while tau[p] is their estimation error; the values of A[p] to B[p] are assumed as observed explanatory variables specified for the regression model. The parameters a0 and b0 come from the empirical distribution of model errors for EBLUP (ordinary or spatial), while alphas denote the linear regression coefficients. It should be stressed that the random effects v[p] are linked with the spatial weight matrix W and the values of ρ by the inprod WinBUGS function. It uses rho_w matrix passed to the WinBUGS by a special macro prepared in R-project environment. Moreover, a special authors' macro for R-project was prepared, which was used as a connector with data input, performing necessary computations and automatic visualization of results (by means of 'coda' module). This macro has a (simplified) form given by the code presented below:

```
# fitting the Spatial EBLUP model
resultSREML <- eblupSFH(Y ~ 1 + A + B, desvar, W, method="REML",
data=d, MAXITER=1500)
mseSREML     <- mseSFH(Y ~ 1 + A + B, desvar, W, method="REML",
data=d, MAXITER=1500)
sigmas_reml <- resultSREML$fit$refvar
rho_REML <- resultSREML$fit$spatialcorr
# determining the model parameters
I <-diag(1,N)
for (j in 1:N) {
 W_row <- W[j,]
 for (k in 1:lpow) {
  W_mat[j,k] <- W_row[k]
 }}
rho_W <- solve(I-rho_REML*W_mat)
a0 <- dochg_shape_Sp
b0 <- dochg_rate_Sp
infile <- "coda1.txt"
indfile <- "codaindex.txt"
data <- list(N=N, Y=Y, tau=tau, A=A, B=B, a0=a0, b0=b0, rho_w=rho_W)
model <- lm(Y ~ 1 + A + B)
mod_smry <- summary(model)
alpha <- as.vector(mod_smry$coefficients[,1])
sigma_2 <- (mod_smry$sigma)*(mod_smry$sigma)
precu <- 1/sigma_2
v <- vector(mode = "numeric", length = N)
u <- vector(mode = "numeric", length = N)
inits <- list(list(alpha=alpha, precu=precu, u=u))
parameters <- c("mu", "alpha", "precu", "v", "u")
working.directory <- getwd()
# simulations - WinBUGS call and collecting the data
sim_HB <- bugs(data, inits, parameters, model_HB,n.chains=1, n.burnin = 1,
n.iter=10000, n.thin = 1, codaPkg=TRUE, working.directory =
working.directory)
 results <- read.coda(infile, indfile, 2, 10000, 1)
```

The code includes (for clarity of expression) only the sections that present how the model parameters were determined and how the simulations were run - with WinBUGS call. The rest of the code has a more ordered character and includes the processes of loading the necessary packages (RODBC, sae, R2WinBUGS, maptools, spdep and MASS), setting the gamma parameters for σ_u^2 (fitdistr function is called here), reading the input data for particular region (functions from RODBC package were used and functions from ‘maptools’ and ‘spdep’ for digital maps were applied here), fitting the EBLUP model (ordinary and spatial version) using ‘sae’ package (eblupFH, mseFH, eblupSFH and mseSFH functions are used here) and – after completing the simulations in WinBUGS – arranging the results and estimating the mean and variance (previously using read.coda function) as well as saving the results to the file (standard cat and format functions were used here).

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 391–410

SUJATHA DISTRIBUTION AND ITS APPLICATIONS

Rama Shanker [1]

ABSTRACT

In this paper a new one-parameter lifetime distribution named “Sujatha Distribution” with an increasing hazard rate for modelling lifetime data has been suggested. Its first four moments about origin and moments about mean have been obtained and expressions for coefficient of variation, skewness, kurtosis and index of dispersion have been given. Various mathematical and statistical properties of the proposed distribution including its hazard rate function, mean residual life function, stochastic ordering, mean deviations, Bonferroni and Lorenz curves, and stress-strength reliability have been discussed. Estimation of its parameter has been discussed using the method of maximum likelihood and the method of moments. The applications and goodness of fit of the distribution have been discussed with three real lifetime data sets and the fit has been compared with one-parameter lifetime distributions including Akash, Shanker, Lindley and exponential distributions.

Key words: lifetime distributions, Akash distribution, Shanker distribution, Lindley distribution, mathematical and statistical properties, estimation of parameter, goodness of fit.

1. Introduction

The analyses and modelling of lifetime data are crucial in all applied sciences including engineering, medical science, insurance and finance, among other things. Although a number of lifetime distributions have evolved in statistical literature, including exponential, Lindley, Akash, Shanker, gamma, lognormal and Weibull distributions, amongst other things, each has its own advantages and disadvantages in modelling lifetime data. The exponential, Lindley, Akash, Shanker and Weibull distributions are more popular than the gamma and the lognormal distributions because the survival functions of the gamma and the lognormal distributions cannot be expressed in closed forms and both require numerical integration. Although Akash, Shanker, Lindley and exponential distributions are of one parameter, Akash, Shanker and Lindley distributions have

[1] Department of Statistics, Eritrea Institute of Technology, Asmara, Eritrea.
E-mail: shankerrama2009@gmail.com.

an advantage over the exponential distribution that the exponential distribution has a constant hazard rate, whereas Akash, Shanker, and Lindley distributions have a monotonically increasing hazard rate. Further, Akash, Shanker, and Lindley distributions have many interesting mathematical and statistical properties in terms of shape, moments, skewness, kurtosis, hazard rate function: mean residual life function, stochastic ordering, mean deviations, order statistics, Bonferroni and Lorenz curves, entropy measure, stress-strength reliability and index of dispersion.

The probability density function (p.d.f.) and the cumulative distribution function (c.d.f.) of Lindley (1958) distribution are given by

$$f_1\left(x;\theta\right)=\frac{\theta^2}{\theta+1}\left(1+x\right)e^{-\theta x}\quad ;x>0,\ \theta>0 \tag{1.1}$$

$$F_1\left(x;\theta\right)=1-\left[1+\frac{\theta x}{\theta+1}\right]e^{-\theta x}\ ;x>0,\theta>0 \tag{1.2}$$

The density (1.1) is a two-component mixture of an exponential distribution with scale parameter θ and a gamma distribution with shape parameter 2 and scale parameter θ with their mixing proportions $\frac{\theta}{\theta+1}$ and $\frac{1}{\theta+1}$ respectively. A detailed study on its various mathematical properties, estimation of parameter and application showing the superiority of Lindley distribution over exponential distribution for the waiting times before service of the bank customers has been done by Ghitany *et al.* (2008). The Lindley distribution has been generalized, extended and modified along with its applications in modelling lifetime data from different fields of knowledge by different researchers including Zakerzadeh and Dolati (2009), Nadarajah *et al.* (2011), Deniz and Ojeda (2011), Bakouch *et al.* (2012), Shanker and Mishra (2013 a, 2013 b), Shanker and Amanuel (2013), Shanker *et al*. (2013), Elbatal *et al.* (2013), Ghitany *et al.* (2013), Merovci (2013), Liyanage and Pararai (2014), Ashour and Eltehiwy (2014), Oluyede and Yang (2014), Singh *et al.* (2014), Shanker *et al* (2015, 2016 a, 2016 b), Sharma *et al.* (2015 a, 2015 b), Alkarni (2015), Pararai *et al.* (2015), Abouammoh *et al.* (2015), among others.

Shanker (2015 a) has introduced one-parameter Akash distribution for modelling lifetime data defined by its p.d.f. and c.d.f.

$$f_2\left(x;\theta\right)=\frac{\theta^3}{\theta^2+2}\left(1+x^2\right)e^{-\theta x}\quad ;x>0,\ \theta>0 \tag{1.3}$$

$$F_2\left(x;\theta\right)=1-\left[1+\frac{\theta x\left(\theta x+2\right)}{\theta^2+2}\right]e^{-\theta x}\ ;x>0,\theta>0 \tag{1.4}$$

Shanker (2015 a) has shown that density (1.3) is a two-component mixture of an exponential distribution with scale parameter θ and a gamma distribution with shape parameter 3 and a scale parameter θ with their mixing proportions $\frac{\theta^2}{\theta^2+2}$ and $\frac{2}{\theta^2+2}$ respectively. Shanker (2015 a) has discussed its various mathematical and statistical properties including its shape, moment generating function, moments, skewness, kurtosis, hazard rate function, mean residual life function, stochastic orderings, mean deviations, distribution of order statistics, Bonferroni and Lorenz curves, Renyi entropy measure, stress-strength reliability, among other things. Shanker *et al* (2015 c) has a detailed study on modelling of various lifetime data from different fields using Akash, Lindley and exponential distributions and concluded that Akash distribution has some advantage over Lindley and exponential distributions. Further, Shanker (2015 c) has obtained Poisson mixture of Akash distribution named Poisson-Akash distribution (PAD) and discussed its various mathematical and statistical properties, estimation of its parameter and applications for various count data sets.

The probability density function and the cumulative distribution function of Shanker distribution introduced by Shanker (2015 b) are given by

$$f_3(x;\theta)=\frac{\theta^2}{\theta^2+1}(\theta+x)e^{-\theta x} \quad ;x>0,\ \theta>0 \tag{1.5}$$

$$F_3(x,\theta)=1-\frac{(\theta^2+1)+\theta x}{\theta^2+1}e^{-\theta x} \quad ;x>0,\theta>0 \tag{1.6}$$

Shanker (2015 b) has shown that density (1.5) is a two-component mixture of an exponential distribution with scale parameter θ and a gamma distribution with shape parameter 2 and a scale parameter θ with their mixing proportions $\frac{\theta^2}{\theta^2+1}$ and $\frac{1}{\theta^2+1}$ respectively. Shanker (2015 b) has discussed its various mathematical and statistical properties including its shape, moment generating function, moments, skewness, kurtosis, hazard rate function, mean residual life function, stochastic orderings, mean deviations, distribution of order statistics, Bonferroni and Lorenz curves, Renyi entropy measure, stress-strength reliability, among other things. Further, Shanker (2015 d) has obtained Poisson mixture of Shanker distribution named Poisson-Shanker distribution (PSD) and discussed its various mathematical and statistical properties, estimation of its parameter and applications for various count data sets.

In this paper we have proposed a new continuous distribution, which is better than Akash, Shanker, Lindley and exponential distributions for modelling lifetime data by considering a three-component mixture of an exponential distribution with scale parameter θ, a gamma distribution with shape parameter 2 and scale parameter θ, and a gamma distribution with shape parameter 3 and scale parameter θ with their mixing proportions $\dfrac{\theta^2}{\theta^2+\theta+2}$, $\dfrac{\theta}{\theta^2+\theta+2}$ and $\dfrac{2}{\theta^2+\theta+2}$, respectively. The probability density function (p.d.f.) of a new one-parameter lifetime distribution can be introduced as

$$f_4(x;\theta)=\frac{\theta^3}{\theta^2+\theta+2}\left(1+x+x^2\right)e^{-\theta x}\quad ;x>0,\ \theta>0 \tag{1.7}$$

We would call this new one-parameter continuous lifetime distribution "Sujatha distribution (S.D)". The corresponding cumulative distribution function (c.d.f.) of Sujatha distribution (1.7) is obtained as

$$F_4(x,\theta)=1-\left[1+\frac{\theta x(\theta x+\theta+2)}{\theta^2+\theta+2}\right]e^{-\theta x};x>0,\theta>0 \tag{1.8}$$

The graphs of the p.d.f. and the c.d.f. of Sujatha distribution (1.7) for different values of θ are shown in Figures 1(a) and 1(b).

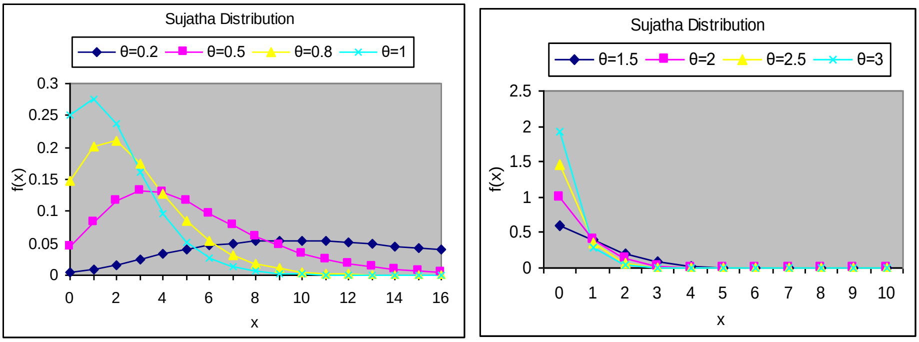


Figure 1(a). Graphs of p.d.f. of Sujatha distribution for selected values of parameter

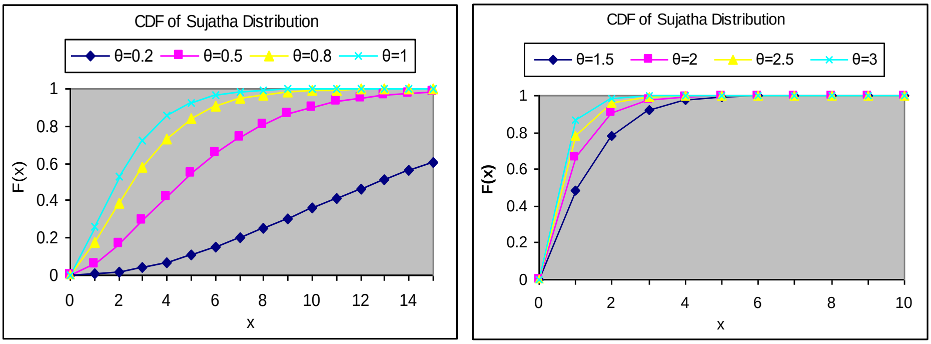


Figure 2(a). Graphs of c.d.f. of Sujatha distribution for selected values of parameter

2. Moment generating function, moments and associated measures

The moment generating function of Sujatha distribution (1.7) can be obtained as

$$M_X(t)=\frac{\theta^3}{\theta^2+\theta+2}\int_0^\infty e^{-(\theta-t)x}\left(1+x+x^2\right)dx$$

$$=\frac{\theta^3}{\theta^2+\theta+2}\left[\frac{1}{\theta-t}+\frac{1}{(\theta-t)^2}+\frac{2}{(\theta-t)^3}\right]$$

$$=\frac{\theta^3}{\theta^2+\theta+2}\left[\frac{1}{\theta}\sum_{k=0}^{\infty}\left(\frac{t}{\theta}\right)^k+\frac{1}{\theta^2}\sum_{k=0}^{\infty}\binom{k+1}{k}\left(\frac{t}{\theta}\right)^k+\frac{2}{\theta^3}\sum_{k=0}^{\infty}\binom{k+2}{k}\left(\frac{t}{\theta}\right)^k\right]$$

$$=\sum_{k=0}^{\infty}\frac{\theta^2+(k+1)\theta+(k+1)(k+2)}{\left(\theta^2+\theta+2\right)}\left(\frac{t}{\theta}\right)^k$$

The r moment about origin μ_r' obtained as the coefficient of $\frac{t^r}{r!}$ in $M_X(t)$, of Sujatha distribution (1.7) has been obtained as

$$\mu_r'=\frac{r!\left[\theta^2+(r+1)\theta+(r+1)(r+2)\right]}{\theta^r\left(\theta^2+\theta+2\right)}\quad ;r=1,2,3,4,...$$

The first four moments about origin of Sujatha distribution (1.7) are thus obtained as

$$\mu_1'=\frac{\theta^2+2\theta+6}{\theta\left(\theta^2+\theta+2\right)}\ ,\qquad \mu_2'=\frac{2\left(\theta^2+3\theta+12\right)}{\theta^2\left(\theta^2+\theta+2\right)},$$

$$\mu_3'=\frac{6\left(\theta^2+4\theta+20\right)}{\theta^3\left(\theta^2+\theta+2\right)},\qquad \mu_4'=\frac{24\left(\theta^2+5\theta+30\right)}{\theta^4\left(\theta^2+\theta+2\right)}$$

Using the relationship between moments about mean and the moments about origin, the moments about mean of Sujatha distribution (1.7) are obtained as

$$\mu_2 = \frac{\theta^4 + 4\theta^3 + 18\theta^2 + 12\theta + 12}{\theta^2\left(\theta^2 + \theta + 2\right)^2}$$

$$\mu_3 = \frac{2\left(\theta^6 + 6\theta^5 + 36\theta^4 + 44\theta^3 + 54\theta^2 + 36\theta + 24\right)}{\theta^3\left(\theta^2 + \theta + 2\right)^3}$$

$$\mu_4 = \frac{3\left(3\theta^8 + 24\theta^7 + 172\theta^6 + 376\theta^5 + 736\theta^4 + 864\theta^3 + 912\theta^2 + 480\theta + 240\right)}{\theta^4\left(\theta^2 + \theta + 2\right)^4}$$

The coefficient of variation $\left(C.V\right)$, coefficient of skewness $\left(\sqrt{\beta_1}\right)$, coefficient of kurtosis $\left(\beta_2\right)$, index of dispersion $\left(\gamma\right)$ of Sujatha distribution (1.7) are thus obtained as

$$C.V = \frac{\sigma}{\mu_1'} = \frac{\sqrt{\theta^4 + 4\theta^3 + 18\theta^2 + 12\theta + 12}}{\theta^2 + 2\theta + 6}$$

$$\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{2\left(\theta^6 + 6\theta^5 + 36\theta^4 + 44\theta^3 + 54\theta^2 + 36\theta + 24\right)}{\left(\theta^4 + 4\theta^3 + 18\theta^2 + 12\theta + 12\right)^{3/2}}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{3\left(3\theta^8 + 24\theta^7 + 172\theta^6 + 376\theta^5 + 736\theta^4 + 864\theta^3 + 912\theta^2 + 480\theta + 240\right)}{\left(\theta^4 + 4\theta^3 + 18\theta^2 + 12\theta + 12\right)^2}$$

$$\gamma = \frac{\sigma^2}{\mu_1'} = \frac{\theta^4 + 4\theta^3 + 18\theta^2 + 12\theta + 12}{\theta\left(\theta^2 + \theta + 2\right)\left(\theta^2 + 2\theta + 6\right)}$$

The over-dispersion, equi-dispersion, under-dispersion of Sujatha, Akash, Shanker, Lindley and exponential distributions for varying values of their parameter θ are presented in Table 1.

Table 1. Over-dispersion, equi-dispersion and under-dispersion of Sujatha, Akash, Shanker, Lindley and exponential distributions for varying values of their parameter θ

Distribution	Over-dispersion $\left(\mu < \sigma^2\right)$	Equi-dispersion $\left(\mu = \sigma^2\right)$	Under-dispersion $\left(\mu > \sigma^2\right)$
Sujatha	$\theta < 1.364271174$	$\theta = 1.364271174$	$\theta > 1.364271174$
Akash	$\theta < 1.515400063$	$\theta = 1.515400063$	$\theta > 1.515400063$
Shanker	$\theta < 1.171535555$	$\theta = 1.171535555$	$\theta > 1.171535555$
Lindley	$\theta < 1.170086487$	$\theta = 1.170086487$	$\theta > 1.170086487$
Exponential	$\theta < 1$	$\theta = 1$	$\theta > 1$

3. Hazard rate function and mean residual life function

Let X be a continuous random variable with p.d.f. $f\left(x\right)$ and c.d.f. $F\left(x\right)$. The hazard rate function (also known as the failure rate function) and the mean residual life function of X are respectively defined as

$$h\left(x\right) = \lim_{\Delta x \to 0} \frac{P\left(X < x + \Delta x \middle| X > x\right)}{\Delta x} = \frac{f\left(x\right)}{1 - F\left(x\right)} \tag{3.1}$$

and $$m\left(x\right) = E\left[X - x \middle| X > x\right] = \frac{1}{1 - F\left(x\right)} \int_x^\infty \left[1 - F\left(t\right)\right] dt \tag{3.2}$$

The corresponding hazard rate function, $h\left(x\right)$ and the mean residual life function, $m\left(x\right)$ of Sujatha distribution are thus obtained as

$$h\left(x\right) = \frac{\theta^3\left(1 + x + x^2\right)}{\theta^2\left(1 + x + x^2\right) + 2\theta x + \theta + 2} \tag{3.3}$$

and $$m\left(x\right) = \frac{\theta^2 + \theta + 2}{\left[\left(\theta^2 + \theta + 2\right) + \theta x\left(\theta x + \theta + 2\right)\right] e^{-\theta x}} \int_x^\infty \left[1 + \frac{\theta t\left(\theta t + \theta + 2\right)}{\theta^2 + \theta + 2}\right] e^{-\theta t} dt$$

$$=\frac{\theta^2\left(x^2+x+1\right)+2\theta\left(2x+1\right)+6}{\theta\left[\left(\theta^2+\theta+2\right)+\theta x\left(\theta x+\theta+2\right)\right]} \tag{3.4}$$

It can be easily verified that $h(0)=\dfrac{\theta^3}{\theta^2+\theta+2}=f(0)$ and $m(0)=\dfrac{\theta^2+2\theta+6}{\theta\left(\theta^2+\theta+2\right)}=\mu_1'$. The graphs of $h(x)$ and $m(x)$ of Sujatha distribution (1.7) for different values of its parameter are shown in Figures 4(a) and 4(b).

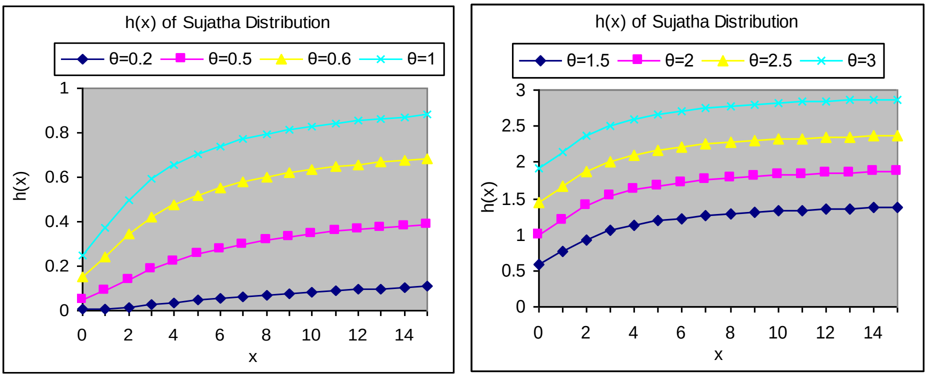


Figure 4(a). Graphs of $h(x)$ of Sujatha distribution for selected values of parameter

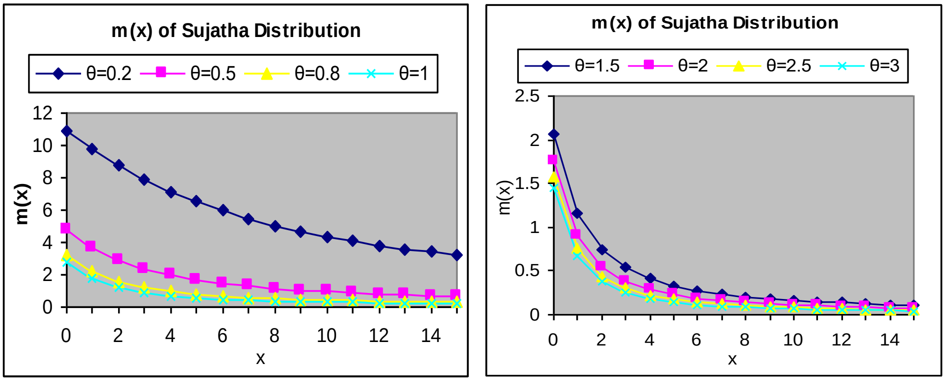


Figure 4(b). Graphs of $m(x)$ of Sujatha distribution for selected values of parameter

It is also obvious from the graphs of $h(x)$ and $m(x)$ that $h(x)$ is a monotonically increasing function of x and θ, whereas $m(x)$ is a monotonically decreasing function of x, and θ.

4. Stochastic orderings

Stochastic ordering of positive continuous random variables is an important tool for judging the comparative behaviour of continuous distributions. A random variable X is said to be smaller than a random variable Y in the

(i) stochastic order $\left(X \le_{st} Y\right)$ if $F_X(x) \ge F_Y(x)$ for all x

(ii) hazard rate order $\left(X \le_{hr} Y\right)$ if $h_X(x) \ge h_Y(x)$ for all x

(iii) mean residual life order $\left(X \le_{mrl} Y\right)$ if $m_X(x) \le m_Y(x)$ for all x

(iv) likelihood ratio order $\left(X \le_{lr} Y\right)$ if $\dfrac{f_X(x)}{f_Y(x)}$ decreases in x.

The following results due to Shaked and Shanthikumar (1994) are well known for establishing stochastic ordering of distributions

$$\begin{tikzcd} X \le_{lr} Y \arrow[r, Rightarrow] & X \le_{hr} Y \arrow[r, Rightarrow] \arrow[d, Rightarrow] & X \le_{mrl} Y \\ & X \le_{st} Y & \end{tikzcd}$$

Sujatha distribution is ordered with respect to the strongest ‘likelihood ratio’ ordering as shown in the following theorem:

Theorem: Let $X \sim$ Sujatha distribution$\left(\theta_1\right)$ and $Y \sim$ Sujatha distribution$\left(\theta_2\right)$. If $\theta_1 > \theta_2$, then $X \le_{lr} Y$ and hence $X \le_{hr} Y$, $X \le_{mrl} Y$ and $X \le_{st} Y$.

Proof: We have

$$\frac{f_X(x)}{f_Y(x)} = \frac{\theta_1^3\left(\theta_2^2+\theta_2+2\right)}{\theta_2^3\left(\theta_1^2+\theta_1+2\right)} e^{-(\theta_1-\theta_2)x} \quad ;\ x>0$$

Now

$$\log\frac{f_X(x)}{f_Y(x)} = \log\left[\frac{\theta_1^3\left(\theta_2^2+\theta_2+2\right)}{\theta_2^3\left(\theta_1^2+\theta_1+2\right)}\right] - \left(\theta_1-\theta_2\right)x$$

This gives $\frac{d}{dx}\log\frac{f_X(x)}{f_Y(x)}=-\left(\theta_1-\theta_2\right)$

Thus, for $\theta_1>\theta_2$, $\frac{d}{dx}\log\frac{f_X(x)}{f_Y(x)}<0$. This means that $X\leq_{lr}Y$ and hence $X\leq_{hr}Y$, $X\leq_{mrl}Y$ and $X\leq_{st}Y$.

5. Deviations from mean and median

The amount of scatter in a population is evidently measured to some extent by the totality of deviations from the mean and the median. These are known as the mean deviation about the mean and the mean deviation about the median and are defined by

$$\delta_1\left(X\right)=\int_0^\infty\left|x-\mu\right|f\left(x\right)dx \text{ and } \delta_2\left(X\right)=\int_0^\infty\left|x-M\right|f\left(x\right)dx, \text{ respectively,}$$

where $\mu=E\left(X\right)$ and $M=\text{Median }\left(X\right)$.

The measures $\delta_1\left(X\right)$ and $\delta_2\left(X\right)$ can be calculated using the following relationships

$$\delta_1\left(X\right)=\int_0^\mu\left(\mu-x\right)f\left(x\right)dx+\int_\mu^\infty\left(x-\mu\right)f\left(x\right)dx$$

$$=\mu F\left(\mu\right)-\int_0^\mu x\,f\left(x\right)dx-\mu\left[1-F\left(\mu\right)\right]+\int_\mu^\infty x\,f\left(x\right)dx$$

$$=2\mu F\left(\mu\right)-2\mu+2\int_\mu^\infty x\,f\left(x\right)dx$$

$$=2\mu F\left(\mu\right)-2\int_0^\mu x\,f\left(x\right)dx \tag{5.1}$$

and

$$\delta_2\left(X\right)=\int_0^M\left(M-x\right)f\left(x\right)dx+\int_M^\infty\left(x-M\right)f\left(x\right)dx$$

$$=M\,F\left(M\right)-\int_0^M x\,f\left(x\right)dx-M\left[1-F\left(M\right)\right]+\int_M^\infty x\,f\left(x\right)dx$$

$$=-\mu+2\int_M^\infty x\,f\left(x\right)dx$$

$$= \mu - 2\int_0^M x\, f\left(x\right) dx \tag{5.2}$$

Using p.d.f. (1.7) and expression for the mean of Sujatha distribution (1.7), we get

$$\int_0^\mu x\, f_4\left(x\right) dx = \mu - \frac{\left\{\theta^3\left(\mu^3+\mu^2+\mu\right)+\theta^2\left(3\mu^2+2\mu+1\right)+2\theta\left(3\mu+1\right)+6\right\}e^{-\theta\mu}}{\theta\left(\theta^2+\theta+2\right)} \tag{5.3}$$

$$\int_0^M x\, f_4\left(x\right) dx = \mu - \frac{\left\{\theta^3\left(M^3+M^2+M\right)+\theta^2\left(3M^2+2M+1\right)+2\theta\left(3M+1\right)+6\right\}e^{-\theta M}}{\theta\left(\theta^2+\theta+2\right)} \tag{5.4}$$

Using expressions from (5.1), (5.2), (5.3) and (5.4), and after some mathematical simplifications, the mean deviation about the mean, $\delta_1\left(X\right)$ and the mean deviation about the median, $\delta_2\left(X\right)$ of Sujatha distribution are obtained as

$$\delta_1\left(X\right) = \frac{2\left[\theta^2\left(\mu^2+\mu+1\right)+2\theta\left(2\mu+1\right)+6\right]e^{-\theta\mu}}{\theta\left(\theta^2+\theta+2\right)} \tag{5.5}$$

and

$$\delta_2\left(X\right) = \frac{2\left[\theta^3\left(M^3+M^2+M\right)+\theta^2\left(3M^2+2M+1\right)+2\theta\left(3M+1\right)+6\right]e^{-\theta M}}{\theta\left(\theta^2+\theta+2\right)} - \mu \tag{5.6}$$

6. Bonferroni and Lorenz curves and indices

The Bonferroni and Lorenz curves (Bonferroni, 1930) and Bonferroni and Gini indices have applications not only in economics to study income and poverty, but also in other fields like reliability, demography, insurance and medicine. The Bonferroni and Lorenz curves are defined as

$$B(p)=\frac{1}{p\mu}\int_0^q x\,f(x)\,dx=\frac{1}{p\mu}\left[\int_0^\infty x\,f(x)\,dx-\int_q^\infty x\,f(x)\,dx\right]=\frac{1}{p\mu}\left[\mu-\int_q^\infty x\,f(x)\,dx\right]$$

(6.1)

and

$$L(p)=\frac{1}{\mu}\int_0^q x\,f(x)\,dx=\frac{1}{\mu}\left[\int_0^\infty x\,f(x)\,dx-\int_q^\infty x\,f(x)\,dx\right]=\frac{1}{\mu}\left[\mu-\int_q^\infty x\,f(x)\,dx\right]$$

(6.2)

respectively, or equivalently as

$$B(p)=\frac{1}{p\mu}\int_0^p F^{-1}(x)\,dx \tag{6.3}$$

and
$$L(p)=\frac{1}{\mu}\int_0^p F^{-1}(x)\,dx \tag{6.4}$$

respectively, where $\mu=E(X)$ and $q=F^{-1}(p)$.

The Bonferroni and Gini indices are thus defined as

$$B=1-\int_0^1 B(p)\,dp \tag{6.5}$$

and
$$G=1-2\int_0^1 L(p)\,dp \tag{6.6}$$

respectively.

Using p.d.f. of Sujatha distribution (1.7), we get

$$\int_q^\infty x\,f_4(x)\,dx=\frac{\left\{\theta^3\left(q^3+q^2+q\right)+\theta^2\left(3q^2+2q+1\right)+2\theta\left(3q+1\right)+6\right\}e^{-\theta q}}{\theta\left(\theta^2+\theta+2\right)}$$

(6.7)

Now using equation (6.7) in (6.1) and (6.2), we get

$$B(p)=\frac{1}{p}\left[1-\frac{\left\{\theta^3\left(q^3+q^2+q\right)+\theta^2\left(3q^2+2q+1\right)+2\theta\left(3q+1\right)+6\right\}e^{-\theta q}}{\theta^2+2\theta+6}\right] \tag{6.8}$$

and $$L(p)=1-\frac{\left\{\theta^3\left(q^3+q^2+q\right)+\theta^2\left(3q^2+2q+1\right)+2\theta\left(3q+1\right)+6\right\}e^{-\theta q}}{\theta^2+2\theta+6} \tag{6.9}$$

Now using equations (6.8) and (6.9) in (6.5) and (6.6), the Bonferroni and Gini indices of Sujatha distribution are obtained as

$$B=1-\frac{\left\{\theta^3\left(q^3+q^2+q\right)+\theta^2\left(3q^2+2q+1\right)+2\theta\left(3q+1\right)+6\right\}e^{-\theta q}}{\theta^2+2\theta+6} \qquad (6.10)$$

$$G=-1+\frac{2\left\{\theta^3\left(q^3+q^2+q\right)+\theta^2\left(3q^2+2q+1\right)+2\theta\left(3q+1\right)+6\right\}e^{-\theta q}}{\theta^2+2\theta+6} \qquad 6.11)$$

7. Stress-strength reliability

The stress-strength reliability of a component illustrates the life of the component which has random strength X that is subjected to a random stress Y. When the stress of the component Y applied to it exceeds the strength of the component X, the component fails instantly, and the component will function satisfactorily until $X > Y$. Therefore, $R = P(Y < X)$ is a measure of the component reliability and is known as stress-strength reliability in statistical literature. It has extensive applications in almost all areas of knowledge especially in engineering such as structures, deterioration of rocket motors, static fatigue of ceramic components, aging of concrete pressure vessels, etc.

Let X and Y be independent strength and stress random variables with Sujatha distribution (1.7) with parameter θ_1 and θ_2 respectively. Then the stress-strength reliability R of Sujatha distribution can be obtained as

$$R = P\left(Y < X\right) = \int_0^\infty P\left(Y < X \mid X = x\right) f_X\left(x\right) dx$$

$$= \int_0^\infty f_4\left(x;\theta_1\right)\, F_4\left(x;\theta_2\right) dx$$

$$= 1 - \frac{\theta_1^3 \begin{bmatrix} \theta_2^{\,6} + \left(4\theta_1 + 3\right)\theta_2^{\,5} + \left(6\theta_1^{\,2} + 10\theta_1 + 13\right)\theta_2^{\,4} + \left(4\theta_1^{\,3} + 12\theta_1^{\,2} + 33\theta_1 + 18\right)\theta_2^{\,3} \\ + \left(\theta_1^{\,4} + 6\theta_1^{\,3} + 29\theta_1^{\,2} + 26\theta_1 + 40\right)\theta_2^{\,2} + \left(\theta_1^{\,3} + 11\theta_1^{\,2} + 10\theta_1 + 20\right)\theta_1\theta_2 \\ + 2\left(\theta_1^{\,2} + \theta_1 + 2\right)\theta_1^{\,2} \end{bmatrix}}{\left(\theta_1^{\,2} + \theta_1 + 2\right)\left(\theta_2^{\,2} + \theta_2 + 2\right)\left(\theta_1 + \theta_2\right)^5}$$

8. Estimation of the parameter

8.1. Maximum Likelihood Estimation of the Parameter

Let $\left(x_1, x_2, x_3, \ldots, x_n\right)$ be random sample from Sujatha distribution (1.7). The likelihood function L is given by

$$L = \left(\frac{\theta^3}{\theta^2 + \theta + 2}\right)^n \prod_{i=1}^{n}\left(1 + x_i + x_i^{\,2}\right) e^{-n\theta\bar{x}}$$

The natural log likelihood function is thus obtained as

$$\ln L = n\ln\left(\frac{\theta^3}{\theta^2 + \theta + 2}\right) + \sum_{i=1}^{n}\ln\left(1 + x_i + x_i^{\,2}\right) - n\theta\bar{x}$$

Now $\dfrac{d\ln L}{d\theta} = \dfrac{3n}{\theta} - \dfrac{n\left(2\theta + 1\right)}{\theta^2 + \theta + 2} - n\bar{x}$, where $\bar{x}$ is the sample mean.

The maximum likelihood estimate, $\hat{\theta}$ of θ of Sujatha distribution (1.7) is the solution of the equation $\dfrac{d\ln L}{d\theta} = 0$ and is given by solution of the following cubic equation

$$\bar{x}\theta^3 + \left(\bar{x} - 1\right)\theta^2 + 2\left(\bar{x} - 1\right)\theta - 6 = 0 \qquad (8.1.1)$$

8.2. Method of Moment Estimation (MOME) of the Parameter

Equating the population mean of Sujatha distribution to the corresponding sample mean, the method of moment (MOM) estimate, $\tilde{\theta}$, of θ is the same as given by equation (8.1.1).

9. Applications and goodness of fit

Since Sujatha, Akash, Shanker, and Lindley distributions have an increasing hazard rate and exponential distribution has a constant hazard rate, Sujatha distribution has been fitted to some data sets to test its goodness of fit over Akash, Shanker, Lindley and exponential distributions. In this section, we present the fitting of Sujatha distribution using maximum likelihood estimate to three real lifetime data sets and compare its goodness of fit with Akash, Shanker, Lindley and exponential distributions. The following three real lifetime data sets have been considered for goodness of fit of distributions.

Data set 1: This data set represents the lifetime data relating to relief times (in minutes) of 20 patients receiving an analgesic and reported by Gross and Clark (1975, P. 105).

1.1 1.4 1.3 1.7 1.9 1.8 1.6 2.2 1.7 2.7 4.1 1.8
1.5 1.2 1.4 3 1.7 2.3 1.6 2

Data Set 2: This data set is the strength data of glass of the aircraft window reported by Fuller *et al.* (1994):

18.83 20.80 21.657 23.03 23.23 24.05 24.321 25.5 25.52 25.80
26.69 26.77 26.78 27.05 27.67 29.90 31.11 33.2 33.73 33.76
33.89 34.76 35.75 35.91 36.98 37.08 37.09 39.58 44.045
45.29 45.381

Data Set 3: The following data represent the tensile strength, measured in GPa, of 69 carbon fibres tested under tension at gauge lengths of 20 mm (Bader and Priest, 1982):

1.312 1.314 1.479 1.552 1.700 1.803 1.861 1.865 1.944 1.958
1.966 1.997 2.006 2.021 2.027 2.055 2.063 2.098 2.140 2.179
2.224 2.240 2.253 2.270 2.272 2.274 2.301 2.301 2.359 2.382
2.382 2.426 2.434 2.435 2.478 2.490 2.511 2.514 2.535 2.554
2.566 2.570 2.586 2.629 2.633 2.642 2.648 2.684 2.697 2.726
2.770 2.773 2.800 2.809 2.818 2.821 2.848 2.880 2.954 3.012
3.067 3.084 3.090 3.096 3.128 3.233 3.433 3.585 3.585

In order to compare the goodness of fit of Sujatha, Akash, Shanker, Lindley and exponential distributions, $-2\ln L$, AIC (Akaike Information Criterion), AICC (Akaike Information Criterion Corrected), BIC (Bayesian Information Criterion), and K-S Statistics (Kolmogorov-Smirnov Statistics) of distributions for three real lifetime data sets have been computed and presented in Table 2. The formulae for computing AIC, AICC, BIC, and K-S Statistics are as follows:

$AIC = -2\ln L + 2k$, $AICC = AIC + \dfrac{2k(k+1)}{(n-k-1)}$, $BIC = -2\ln L + k\ln n$ and

$D = \underset{x}{\text{Sup}}\left|F_n(x) - F_0(x)\right|$, where k = the number of parameters, $n =$ the sample size, and $F_n(x)$ = the empirical distribution function.

Table 2. MLE's, $-2\ln L$, AIC, AICC, BIC, and K-S Statistics of the fitted distributions of data sets 1, 2 and 3

	Model	Parameter estimate	$-2\ln L$	AIC	AICC	BIC	K-S statistic
Data 1	Sujatha	1.136745	57.50	59.50	59.72	60.49	0.309
	Akash	1.156923	59.52	61.52	61.74	62.51	0.320
	Shanker	0.803867	59.78	61.78	61.22	62.77	0.315
	Lindley	0.816118	60.50	62.50	62.72	63.49	0.341
	Exponential	0.526316	65.67	67.67	67.90	68.67	0.389
Data 2	Sujatha	0.09561	241.50	243.50	243.64	244.94	0.270
	Akash	0.097062	240.68	242.68	242.82	244.11	0.266
	Shanker	0.064712	252.35	254.35	254.49	255.78	0.326
	Lindley	0.062988	253.99	255.99	256.13	257.42	0.333
	Exponential	0.032455	274.53	276.53	276.67	277.96	0.426
Data 3	Sujatha	0.936119	221.61	223.61	223.67	225.84	0.319
	Akash	0.964726	224.28	226.28	226.34	228.51	0.348
	Shanker	0.658029	233.01	235.01	235.06	237.24	0.355
	Lindley	0.659000	238.38	240.38	240.44	242.61	0.390
	Exponential	0.407941	261.74	263.74	263.80	265.97	0.434

The best fit of the distribution is the distribution which corresponds to the lower values of $-2\ln L$, AIC, AICC, BIC, and K-S statistics. It is obvious from the fitting of distributions for three data sets in the Table 2 that Sujatha distribution provides better fit than Akash, Shanker, Lindley and exponential distributions for modelling lifetime data in data sets 1 and 3, whereas Akash distribution provides slightly better fit than Sujatha distribution in data set 2.

10. Concluding remarks

A new lifetime distribution named "Sujatha distribution" with an increasing hazard rate has been introduced to model lifetime data. Its moment generating function, moments about origin, moments about mean and expressions for skewness and kurtosis have been given. Various interesting mathematical and statistical properties of Sujatha distribution such as its hazard rate function, mean residual life function, stochastic ordering, mean deviations, Bonferroni and Lorenz curves, and stress-strength reliability have been discussed. The method of maximum likelihood and the method of moments for estimating its parameter have been discussed. Three examples of real lifetime data sets have been presented to show the applications and goodness of fit of Sujatha distribution with Akash, Shanker, Lindley and exponential distributions.

NOTE: The paper is dedicated to my inspirational friend and colleague, Dr. Sujatha Selvaraj, former faculty, Department of Banking and Finance, College of Business and Economics, Halhale, Eritrea.

Acknowledgment

The author is grateful for the editor-in-chief and the anonymous reviewer for constructive comments.

REFERENCES

ABOUAMMOH, A. M., ALSHANGITI, A. M., RAGAB, I. E., (2015). A new generalized Lindley distribution, Journal of Statistical Computation and Simulation, preprint http://dx.doi.org/10.1080/ 00949655.2014.995101.

ALKARNI, S., (2015). Extended power Lindley distribution - a new statistical model for non-monotone survival data, European journal of statistics and probability, 3(3), pp. 19–34.

ASHOUR, S., ELTEHIWY, M., (2014). Exponentiated Power Lindley distribution, Journal of Advanced Research, preprint http://dx.doi.org/10.1016/ j.jare. 2014.08.005.

BAKOUCH, H. S., AL-ZAHARANI, B., AL-SHOMRANI, A., MARCHI, V., LOUZAD, F., (2012). An extended Lindley distribution, Journal of the Korean Statistical Society, 41, pp. 75–85.

BADER, M. G., PRIEST, A. M., (1982). Statistical aspects of fiber and bundle strength in hybrid composites, In: Hayashi, T., Kawata, K., Umekawa, S. (Eds), Progress in Science and Engineering composites, ICCM-IV, Tokyo, 1129–1136.

BONFERRONI, C. E., (1930). Elementi di Statistca generale, Seeber, Firenze.

DENIZ, E., OJEDA, E., (2011). The discrete Lindley distribution - properties and applications, Journal of Statistical Computation and Simulation, 81, pp. 1405–1416.

ELBATAL, I., MEROVI, F., ELGARHY, M., (2013). A new generalized Lindley distribution, Mathematical Theory and Modeling, 3 (13), pp. 30–47.

FULLER, E. J., FRIEMAN, S., QUINN, J., QUINN, G., CARTER, W., (1994). Fracture mechanics approach to the design of glass aircraft windows: a case study, SPIE Proc 2286, pp. 419–430.

GHITANY, M. E., ATIEH, B., NADARAJAH, S., (2008). Lindley distribution and its application, Mathematics Computing and Simulation, 78, pp. 493–506.

GHITANY, M., AL-MUTAIRI, D., BALAKRISHNAN, N., AL-ENEZI, I., (2013). Power Lindley distribution and associated inference, Computational Statistics and Data Analysis, 64, pp. 20–33.

GROSS, A. J., CLARK, V. A., (1975). Survival Distributions: Reliability Applications in the Biometrical Sciences, John Wiley, New York.

LINDLEY, D. V., (1958). Fiducial distributions and Bayes' theorem, Journal of the Royal Statistical Society, Series B, 20, pp. 102–107.

LIYANAGE, G. W., PARARAI, M., (2014). A generalized Power Lindley distribution with applications, Asian Journal of Mathematics and Applications, pp. 1–23.

MEROVCI, F., (2013). Transmuted Lindley distribution, International Journal of Open Problems in Computer Science and Mathematics, 6, pp. 63–72.

NADARAJAH, S., BAKOUCH, H. S., TAHMASBI, R., (2011). A generalized Lindley distribution, Sankhya Series B, 73, pp. 331–359.

OLUYEDE, B. O., YANG, T., (2014). A new class of generalized Lindley distribution with applications, Journal of Statistical Computation and Simulation, 85 (10), pp. 2072–2100.

PARARAI, M., LIYANAGE, G. W., OLUYEDE, B. O., (2015). A new class of generalized Power Lindley distribution with applications to lifetime data, Theoretical Mathematics & Applications, 5 (1), pp. 53–96.

RENYI, A., (1961). On measures of entropy and information, in proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, 1, pp. 547–561, Berkeley, University of California Press.

SHAKED, M., SHANTHIKUMAR, J. G., (1994). Stochastic Orders and Their Applications, Academic Press, New York.

SHANKER, R., (2015 a). Akash distribution and Its Applications, International Journal of Probability and Statistics, 4 (3), pp. 65–75.

SHANKER, R., (2015 b). Shanker distribution and Its Applications, International Journal of Statistics and Applications, 5 (6), pp. 338–348.

SHANKER, R., (2015 c). The discrete Poisson-Akash distribution, communicated.

SHANKER, R., (2015 d). The discrete Poisson-Shanker distribution, Accepted for publication in Jacobs Journal of Biostatistics.

SHANKER, R., HAGOS, F., SUJATHA, S., (2015). On modeling of lifetimes data using exponential and Lindley distributions, Biometrics & Biostatistics International Journal, 2 (5), pp. 1–9.

SHANKER, R., HAGOS, F., SHARMA, S., (2016 a): On two parameter Lindley distribution and Its Applications to model Lifetime data, Biometrics & Biostatistics International Journal, 3 (1), pp. 1–8.

SHANKER, R., HAGOS, F., SUJATHA, S., (2016 b): On modeling of Lifetimes data using one parameter Akash, Lindley and exponential distributions, Biometrics & Biostatistics International Journal, 3 (2), pp. 1–10.

SHANKER, R., MISHRA, A., (2013 a). A two-parameter Lindley distribution, Statistics in Transition-new series, 14 (1), pp. 45–56.

SHANKER, R., MISHRA, A., (2013 b). A quasi Lindley distribution, African Journal of Mathematics and Computer Science Research, 6(4), pp. 64–71.

SHANKER, R., AMANUEL, A. G., (2013). A new quasi Lindley distribution, International Journal of Statistics and Systems, 8 (2), pp. 143–156.

SHANKER, R., SHARMA, S., SHANKER, R., (2013). A two-parameter Lindley distribution for modeling waiting and survival times data, Applied Mathematics, 4, pp. 363–368.

SHARMA, V., SINGH, S., SINGH, U., AGIWAL, V., (2015). The inverse Lindley distribution - a stress-strength reliability model with applications to head and neck cancer data, Journal of Industrial & Production Engineering, 32 (3), pp. 162–173.

SINGH, S. K., SINGH, U., SHARMA, V. K., (2014). The Truncated Lindley distribution-inference and Application, Journal of Statistics Applications & Probability, 3 (2), pp. 219–228.

SMITH, R. L, NAYLOR, J. C., (1987). A comparison of Maximum likelihood and Bayesian estimators for the three parameter Weibull distribution, Applied Statistics, 36, pp. 358–369.

ZAKERZADEH, H., DOLATI, A., (2009). Generalized Lindley distribution, Journal of Mathematical extension, 3 (2), pp. 13–25.

STATISTICS IN TRANSITION NEW SERIES, September 2016
Vol. 17, No. 3, pp. 411–428

ESTIMATION OF MEAN ON THE BASIS OF CONDITIONAL SIMPLE RANDOM SAMPLE

Janusz Wywiał[1]

ABSTRACT

Estimation of the population mean in a finite and fixed population on the basis of the conditional simple random sampling design dependent on order statistics (quantiles) of an auxiliary variable is considered. Properties of the well-known Horvitz-Thompson and ratio type estimators as well as the sample mean are taken into account under the conditional simple random sampling designs. The considered examples of empirical analysis lead to the conclusion that under some additional conditions the proposed estimation strategies based on the conditional simple random sample are usually more accurate than the mean from the simple random sample drawn without replacement.

Key words: conditional sampling design, order statistic, concomitant, sample quantile, auxiliary variable, Horvitz-Thompson statistic, inclusion probabilities, sampling scheme, ratio estimator.

1. Introduction

Sampling designs dependent on an auxiliary variable are constructed in order to improve accuracy of population parameters estimation. Application of auxiliary information to construction of conditional versions of sampling designs are considered, e.g. by Royall and Cumberland (1981), Tillé (1998, 2006) and Wywiał (2003).

The fixed population of size N denoted by U will be taken into account. The observation of a variable under study and an auxiliary variable are identifiable and denoted by y_i and $x_i, i=1,\ldots,N$, respectively. We assume that $x_i \leq x_{i+1}, i = 1,\ldots,N-1$. Our general purpose is estimation of the population average: $\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$ where y_i, $i=1,...,N$, are values of the variable under study.

The well-known simple random sampling design is defined as follows: $P_0(s) = \binom{N}{n}^{-1}$ for all $s \in \mathbf{S}$ where $\mathbf{S}$ is the sample space of the samples s with fixed effective size $1 < n < N$.

Let $s = \{s_1, i, s_2\}$ where $s_1 = \{i_1,...,i_{r-1}\}$, $s_2 = \{i_{r+1},...,i_n\}$, $i_j < i$ for $j = 1,...,r$, $i_r = i$ and $i_j > i$ for $j = r+1,...,n$. Thus, x_i is one of the possible observations of order statistic $X_{(r)}$ of rank r $(r = 1,...,n)$ from sample s. Let $\mathbf{S}(r,i) =$

[1] Department of Statistics, Faculty of Management, University of Economics in Katowice. E-mail: janusz.wywial@ue.katowice.pl

$\{s : X_{(r)} = x_i\}$ be the set of all samples whose r -th order statistic of the auxiliary variable is equal to x_i where $r \leq i \leq N-n+r$. Hence, $\bigcup_{i=r}^{N-n+r} \mathbf{S}(r,i) = \mathbf{S}$.

The size of the set $\mathbf{S}(r,i)$ is denoted by $g(r,i) = Card(\mathbf{S}(r,i))$ and

$$g(r,i) = \binom{i-1}{r-1}\binom{N-i}{n-r}. \tag{1}$$

The conditional version of the order statistic distribution is as follows:

$$P\left(X_{(r)} = x_i | x_u \leq X_{(r)} \leq x_w\right) = \frac{P\left(X_{(r)} = x_i\right)}{P\left(x_u \leq X_{(r)} \leq x_w\right)} = \frac{g(r,i)}{z(r,u,w)} \tag{2}$$

where

$$P\left(X_{(r)} = x_i\right) = \frac{g(r,i)}{\binom{N}{n}}, \quad i = r, ..., N-n+r. \tag{3}$$

$$P\left(x_u \leq X_{(r)} \leq x_w\right) = \frac{z(r,u,w)}{\binom{N}{n}}, \tag{4}$$

$$z(r,u,w) = \sum_{t=u}^{w} g(r,t). \tag{5}$$

Wywiał (2014) proposed the following conditional version of the simple random sampling design:

$$P_0\left(s|r,u,w\right) = P_0\left(s|x_u \leq X_{(r)} \leq x_w\right) = \frac{1}{z(r,u,w)}. \tag{6}$$

$P_0\left(s|r,u,w\right)$ provide such the simple random samples that r-th order $X_{(r)}$ takes a value from interval $[x_u;\ x_w]$ where $u \leq r \leq w$. Let us note that in the particular case when $u = r$ and $w = N-n+r$ sampling design $P_0(s|r,u,w)$ becomes ordinary simple random sample design $P_0(s)$.

Wywiał (2014) derived the first and second order inclusion probabilities for the sampling design. Moreover, Wywiał proposed the following sampling scheme implementing $P_0(s|r,u,w)$. Firstly, population elements are ordered according to the increasing values of the auxiliary variable. Next, the i-th element of the population where $i = u, u+1, ..., w$ and $r = [n\alpha]+1$, is drawn with probability:

$$P(X_{(r)} = x_i | x_u \leq X_{(r)} \leq x_w) = \frac{g(r,i)}{\sum_{j=u}^{w} g(r,j)}. \tag{7}$$

Finally, two simple samples $s_1(i)$ and $s_2(i)$ are drawn without replacement from subpopulations $U_1 = \{1, ..., i-1\}$ and $U_2 = \{i+1, i+2, ..., N\}$, respectively. Sample $s_1(i)$ is of size $r-1$ and sample $s_2(i)$ is of size $n-r$. The sampling designs of these samples are independent and

$$P_0(s_1(i)) = \binom{i-1}{r-1}^{-1}, \quad P_0(s_2(i)) = \binom{N-i}{n-r}^{-1}. \tag{8}$$

2. Strategies dependent on conditional simple random sample

2.1. The Horvitz-Thompson estimator

The well known Horvitz-Thompson (1952) estimator is given by:

$$\bar{y}_{HT,s} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k} \tag{9}$$

Estimation strategy $(\bar{y}_{HT,s}, P(s))$ is unbiased for $\bar{y}$ if $\pi_k > 0$ for $k = 1,...,N$, where π_k is the inclusion probability of sampling design $P(s)$. The variance of the strategy is:

$$V_0\left(\bar{y}_{HT,s}, P(s)\right) = \frac{1}{N^2}\left(\sum_{k\in U}\sum_{l\in U}\Delta_{k,l}\frac{y_k y_l}{\pi_k\pi_l}\right), \qquad \Delta_{k,l} = \pi_{k,l} - \pi_k\pi_l. \tag{10}$$

Particularly, under simple random sampling design $P_0(s)$ the strategy $(t_{HT,s}, P(s))$ reduces to the simple random sample mean denoted by $(\bar{y}_s, P_0(s))$, where

$$\bar{y}_s = \frac{1}{n}\sum_{k\in s} y_k. \tag{11}$$

It is the unbiased estimator of the population mean and its variance is:

$$V_0(\bar{y}_s) = \frac{N-n}{Nn}v_*(y), \qquad v_*(y) = \frac{1}{N-1}\sum_{k\in U}(y_k - \bar{y})^2. \tag{12}$$

Moreover, let us note that in the case of the unconditional simple random scheme the Horvitz-Thompson strategy reduces to the simple random sample mean.

Example 2.1. *In the book by Särndal C. E., B. Swensson, J. Wretman (1992) the data about Sweden municipalities are presented. The size of the population of municipalities is $N = 284$. We take into account two variables. The first is revenues from the 1985 municipal taxation (in millions of kronor) and it is treated as the variable under study denoted by y. The second one is the 1975 population of municipalities (in thousands) and it is treated as the auxiliary variable denoted by x. Our purpose is the estimation of population mean $\bar{y}$. The mean of the auxiliary variable is $\bar{x} = 28.810$. The population mean of the variable under study is estimated by means of strategy $(\bar{y}_{HT,s}, P_0(|r,u,w))$. The relative efficiency is denoted by:*

$$deff(r,u,w|n) = V(\bar{y}_{HT,s}, P_0(s|r,u,w))/V_0(\bar{y}_s)$$

Particulary, we have $deff(3,260,270|3) = 0.045$*,* $deff(11,270,280|15) = 0.14$ *and* $deff(22,267,277|29) = 0.147$*. Thus, in all the considered cases the mean from the conditional simple random sample is several times more accurate than the simple random sample mean.*

2.2. Conditional simple random sample mean

Let H and T be statistics dependent on observations of the variable under study and the auxiliary variable observed in the sample s drawn according to sampling design $P_0(s|r,u,w)$. The basic moments of statistics H and T are as follows:

$$E_0(H|r,u,w) = \sum_{s\in\mathbf{S}(r,u,w)} hP_0(s|r,u,w),$$

$$E_0(HT|r,u,w) = \sum_{s\in\mathbf{S}(r,u,w)} htP_0(s|r,u,w).$$

$$V_0(H,T|r,u,w) = E_0(HT|r,u,w) - E_0(H|r,u,w)E_0(T|r,u,w).$$

Now, let H and T be statistics dependent on order statistic $X_{(r)}$ or its concomitant $Y_{[r]}$. The basic moments of the statistics H and T are denoted as follows:

$$E(H|r,u,w) = \sum_{i=u}^{w} h_i P(X_{(r)} = x_i|x_u \leq X_{(r)} \leq x_w),$$

$$E(HT|r,u,w) = \sum_{i=u}^{w} h_i t_i P(X_{(r)} = x_i|x_u \leq X_{(r)} \leq x_w),$$

$$V(H,T|r,u,w)) = E(HT|r,u,w) - E(H|r,u,w)E(H|r,u,w),$$

$$V(H|r,u,w)) = V(H,H|r,u,w)).$$

Let random variable I_r have the following probability function:

$$P(I_r = i|u \leq r \leq w) = P\left(X_{(r)} = x_i|x_u \leq X_{(r)} \leq x_w\right) = \frac{z(r,u,w)}{\binom{N}{n}} \tag{13}$$

where $i = u,...,w$ and $z(r,u,w)$ explains equation (5) and $r \leq u < w \leq N-n+r$. Let $\bar{x}(1,i-1)$, $i = u,...,w$, be the following population mean of the left-truncated (in the point x_i) distribution of the auxiliary variable:

$$\bar{x}(1,i-1) = \frac{1}{i-1}\sum_{k=1}^{i-1} x_k, \quad 1 < r \leq n. \tag{14}$$

Let us note that $\bar{x}(1,i-1)$ is a value of the random variable denoted by $\bar{X}(1,I_r-1)$ and

$$P(I_r=i|u\leq r\leq w)=P\left(X_{(r)}=x_i|x_u\leq X_{(r)}\leq x_w\right). \tag{15}$$

Similarly, we define the following random variables: $\bar{X}(I_r+1,N)$, $\bar{Y}[1,I_r-1]$, $\bar{Y}[I_r+1,N]$, $\overline{XY}[1,I_r-1]$ and $\overline{XY}[I_r+1,N]$, $V_{x,y}[1,I_r-1]$ and $V_{x,y}[I_r+1,N]$ which take values equal to the following moments, respectively:

$$\bar{x}(i+1,N)=\frac{1}{N-i}\sum_{k=i+1}^{N}x_k, \quad 1\leq r<n, i<N \tag{16}$$

$$\bar{y}[1,i-1]=\frac{1}{i-1}\sum_{k=1}^{i-1}y_k, \quad 1<r\leq n, \tag{17}$$

$$\bar{y}[i+1,N]=\frac{1}{N-i}\sum_{k=i+1}^{N}y_k \quad 1\leq r<n, i<N, \tag{18}$$

$$\overline{xy}[1,i-1]=\frac{1}{i-1}\sum_{k=1}^{i-1}x_ky_k, \quad 1<r\leq n, \tag{19}$$

$$\overline{xy}[i+1,N]=\frac{1}{N-i}\sum_{k=i+1}^{N}x_ky_k, \quad 1\leq r<n, i<N, \tag{20}$$

$$v_{x,y}[1,i-1]=\overline{xy}[1,i-1]-\bar{x}(1,i-1)\bar{y}[1,i-1], \quad 1<r\leq n, \tag{21}$$

$$v_{x,y}[i+1,N]=\overline{xy}[i+1,N]-\bar{x}(i+1,N)\bar{y}[i+1,N], \quad 1\leq r<n, i<N. \tag{22}$$

Particularly, $v_x[1,i-1]=v_{x,x}[1,i-1]$ and $v_y[i+1,N]=v_{y,y}[i+1,N]$. Parameters of sample means $\bar{x}_s$, $\bar{y}_s$ under the conditional simple random sample design are considered in the following theorem.

Lemma 2.1. *Under the sampling design defined by expression (6) the basic parameters of $\bar{x}_s$, $\bar{y}_s$ are as follows:*

$$E_0(\bar{x}_s|r,u,w))=\frac{r-1}{n}E\left(\bar{X}(1,I_r-1)|r,u,w\right)+\frac{1}{n}E\left(X_{(r)}|r,u,w\right)+ \\ +\frac{n-r}{n}E\left(\bar{X}(I_r+1,N)|r,u,w\right) \tag{23}$$

where

$$E\left(\bar{X}(1,I_r-1)|r,u,w\right)=\sum_{i=u}^{w}\bar{x}(1,i-1)P\left(X_{(r)}=x_i|r,u,w\right), \tag{24}$$

$$E\left(X_{(r)}|r,u,w\right)=\sum_{i=u}^{w}x_iP\left(X_{(r)}=x_i|r,u,w\right), \tag{25}$$

$$E\left(\bar{X}(I_r+1,N)|r,u,w\right)=\sum_{i=u}^{w}\bar{x}(i+1,N)P\left(X_{(r)}=x_i|r,u,w\right). \tag{26}$$

$$E_0(\bar{y}_s|r,u,w)=\frac{r-1}{n}E\left(\bar{Y}[1,I_r-1]|r,u,w\right)+\\+\frac{1}{n}E\left(Y_{[r]}|r,u,w\right)+\frac{n-r}{n}E\left(\bar{Y}[I_r+1,N]|r,u,w\right) \tag{27}$$

where

$$E\left(\bar{Y}[1,I_r-1]|r,u,w\right)=\sum_{i=u}^{w}\bar{y}[1,i-1]P\left(X_{(r)}=x_i|r,u,w\right), \tag{28}$$

$$E\left(Y_{[r]}|r,u,w\right)=\sum_{i=u}^{w}y_iP\left(X_{(r)}=x_i|r,u,w\right), \tag{29}$$

$$E\left(\bar{Y}[I_r+1,N]|r,u,w\right)=\sum_{i=u}^{w}\bar{y}[i+1,N]P\left(X_{(r)}=x_i|r,u,w\right). \tag{30}$$

$$V_0(\bar{x}_s,\bar{y}_s|r,u,w)=\\=\frac{(r-1)^2}{n^2}V_0(\bar{x}_{s_1},\bar{y}_{s_1}|r,u,w)+\frac{r-1}{n^2}V_0(\bar{x}_{s_1},Y_{[r]}|r,u,w)+\\+\frac{(r-1)(n-r)}{n^2}V_0(\bar{x}_{s_1},\bar{y}_{s_2}|r,u,w)+\frac{r-1}{n^2}V_0(X_{(r)},\bar{y}_{s_1}|r,u,w)+\\+\frac{1}{n^2}V_0(X_{(r)},Y_{[r]}|r,u,w)+\frac{n-r}{n^2}V_0(X_{(r)},\bar{y}_{s_2}|r,u,w)+\\+\frac{(r-1)(n-r)}{n^2}V_0(\bar{x}_{s_2},\bar{y}_{s_1}|r,u,w)+\frac{n-r}{n^2}V_0(\bar{x}_{s_2},Y_{[r]}|r,u,w)+\\+\frac{(n-r)^2}{n^2}V_0(\bar{x}_{s_2},\bar{y}_{s_2}|r,u,w) \tag{31}$$

where

$$V_0(\bar{x}_{s_1},\bar{y}_{s_1}|r,u,w)=\frac{1}{r-1}E\left(\frac{I_r-r}{I_r-1}V_{xy}[1,I_r-1]|r,u,w\right)+\\+V(\bar{X}(1,I_r-1),\bar{Y}[1,I_r-1]|r,u,w), \tag{32}$$

$$V_0(\bar{x}_{s_1}, Y_{[r]}|r,u,w) = V(Y_{[r]}, \bar{X}(1, I_r - 1)|r,u,w), \tag{33}$$

$$\begin{aligned} V_0(X_{(r)}, \bar{y}_{s_1}|r,u,w) &= V(X_{(r)}, \bar{Y}[1, I_r - 1]|r,u,w) = \\ &= E(X_{(r)}\bar{Y}[1, I_r - 1]|r,u,w) - E(X_{(r)}|r,u,w)E(\bar{Y}[1, I_r - 1]|r,u,w), \end{aligned} \tag{34}$$

$$\begin{aligned} V_0(\bar{x}_{s_1}, \bar{y}_{s_2}|r,u,w) &= V\left(\bar{X}(1, I_r - 1), \bar{Y}[I_r + 1, N]|r,u,w\right) = \\ &= E\left(\bar{X}(1, I_r - 1)\bar{Y}[I_r + 1, N]|r,u,w\right) + \\ &\quad - E\left(\bar{X}(1, I_r - 1)|r,u,w\right)E\left(\bar{Y}[I_r + 1, N]|r,u,w\right), \end{aligned} \tag{35}$$

$$\begin{aligned} V_0(\bar{x}_{s_2}, \bar{y}_{s_1}|r,u,w) &= V\left(\bar{X}(I_r + 1, N), \bar{Y}[1, I_r - 1]|r,u,w\right) = \\ &= E\left(\bar{X}(I_r + 1, N)\bar{Y}[1, I_r - 1]|r,u,w\right) + \\ &\quad - E\left(\bar{X}(I_r + 1, N)|r,u,w\right)E\left(\bar{Y}[1, I_r - 1]|r,u,w\right), \end{aligned} \tag{36}$$

$$V(X_{(r)}, Y_{[r]}|r,u,w) = E(X_{(r)}, Y_{[r]}|r,u,w) - E(X_{(r)}|r,u,w)E(_{[r]}|r,u,w), \tag{37}$$

$$E(X_{(r)}, Y_{[r]}|r,u,w) = \sum_{i=u}^{w} x_i y_i P\left(X_{(r)} = x_i|r,u,w\right),$$

$$V_0(\bar{x}_{s_2}, Y_{[r]}|r,u,w) = V(Y_{[r]}, \bar{X}(I_r + 1, N)|r,u,w), \tag{38}$$

$$V_0(X_{(r)}, \bar{y}_{s_2}|r,u,w) = V(X_{(r)}, \bar{Y}[I_r + 1, N]|r,u,w), \tag{39}$$

$$\begin{aligned} V_0(\bar{x}_{s_2}, \bar{y}_{s_2}|r,u,w) = \frac{1}{n-r}E\left(\frac{N-n+r-I_r}{N-I_r}V_{xy}[I_r+1,N]|r,u,w\right) + \\ + V(\bar{X}(I_r+1,N), \bar{Y}(I_r+1,N)|r,u,w)). \end{aligned} \tag{40}$$

The proof is presented in the Appendix.Let $v_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$ and $v_x = v_{xx}$, $v_y = v_{yy}$.

Theorem 2.1. *When $y_k \approx \bar{y} + a(x_k - \bar{x})$ for all $k = 1, ..., N$ where $a = \frac{v_{xy}}{v_x}$ and $E_0(\bar{x}_s|r,u,w) = \bar{x}$ where $E_0(\bar{x}_s|r,u,w)$ is expressed by (23), then strategy $(\bar{y}_s, P_0(|r,u,w))$ is approximately unbiased for $\bar{y}$.*

The expressions (31)-(41) of Lemma 4.1 determine approximate variance:

$$\begin{aligned}
&V(\bar{y}_s|r,u,w) = \\
&= \frac{1}{n^2}\Bigg\{(r-1)E\left(\frac{I_r-r}{I_r-1}V_y[1,I_r-1]|r,u,w\right)+(r-1)^2V(\bar{Y}[1,I_r-1]|r,u,w)+ \\
&\qquad +2(r-1)V(Y_{[r]},\bar{Y}(1,I_r-1)|r,u,w)+ \\
&\qquad +2(r-1)(n-r)V\left(\bar{Y}(1,I_r-1),\bar{Y}[I_r+1,N]|r,u,w\right)+V(Y_{[r]}|r,u,w)+ \\
&\qquad +2(n-r)V(Y_{[r]},\bar{Y}(I_r+1,N)|r,u,w)+(n-r)^2V(\bar{Y}(I_r+1,N)|r,u,w))+ \\
&\qquad +(n-r)E\left(\frac{N-n+r-I}{N-I}V_y[I_r+1,N]|r,u,w\right)\Bigg\}. \qquad (41)
\end{aligned}$$

The proof is presented in the Appendix. Hence, if sample size n and parameters u and w are fixed then parameter r has to be determined in such a way that $|E_0(\bar{x}_s|r,u,w)-\bar{x}| = minimum$.

Example 2.2. *Let us consider the same data as those taken into account in Example 2.1. Our purpose is estimation population mean $\bar{y}$ by means of $(\bar{y}_s, P_0(s|r,u,w))$. The relative bias of the strategy is denoted by:*

$$\delta(r,u,w|n) = b(\bar{y}_s, P_0(s|r,u,w))/\sqrt{V(\bar{y}_s, P_0(s|r,u,w))}$$

where $b(r,u,w) = \bar{y}_s - \bar{y}$ is the bias of $(\bar{y}_s, P_0(s|r,u,w))$. The relative efficiency is denoted by

$$deff(r,u,w|n) = MSE(\bar{y}_s, P_0(s|r,u,w))/V_0(\bar{y}_s)$$

where

$$MSE(\bar{y}_s, P_0(s|r,u,w)) = V(\bar{y}_s, P_0(s|r,u,w)) + b^2(r,u,w)$$

and $V_0(\bar{y}_s)$ is the variance of the mean from the simple random sample drawn without replacement. After some computations we have:
$\delta(3,195,205|3) = -0.915$, $deff(3,195,205|3) = 0.372$,
$\delta(11,170,230|15) = -0.022$, $deff(11,170,230|15) = 3.458$,
$\delta(22,203,212|29) = -0.124$, $deff(22,203,212|29) = 5.74$.
Hence, only in the case of the small sample size $n = 3$ the mean from the conditional simple sample is more accurate than the simple sample mean.

David and Nagaraja (2003), p. 145 show that

$$E(Y_{[r]}) = \bar{y} + \frac{v_{xy}}{v_x}(E(X_{(r)}) - \bar{x}), \qquad r = 1,...,n \qquad (42)$$

This let us consider concomitant $Y_{[r]}$ as the estimator of the population mean. Equation (42) straightforwardly leads to the following theorem.

Theorem 2.2. *Under the sampling design defined by expression (6), concomitant* $Y_{[r]}$ *is unbiased estimator of the population mean when*

$$E(X_{(r)}) = \bar{x} \quad and \quad V(Y_{[r]}) = \sum_{i=r}^{N-n+r} (y_i - \bar{y})^2 P(X_{(r)} = x_i). \tag{43}$$

Example 2.3. *Let us consider the same data as those taken into account in Example 2.1. Our purpose is estimation* $\bar{y}$. *The relationship between the variable under study and the auxiliary one is strict because their correlation coefficient is equal to* 0.967. *The population mean of the variable under study we estimate by means of* $(Y_{[r]}, P_0(s))$. *The range of* $Y_{[r]}$ *is the same as the range of* $X_{(r)}$ *where range r minimizes quantity* $|E(X_(r)) - \bar{x}|$.

The relative efficiency of the strategy is determined according the following expression: $deff(r|n) = V(Y_{[r]}, P_0(s))/V_0(\bar{y}_s)$ *After appropriate calculation, we have* $deff(2|3) = 0.235$, $deff(11|15) = 0.370$, $deff(22|29) = 0.430$. *Thus, in all the considered cases the mean from the conditional simple sample is more accurate than the simple sample mean.*

Example 2.4. *We still consider the problem formulated in Example 2.3. Now the population mean of municipal taxation is estimated on the basis of strategy* $(Y_{[r]}, P_0(s|r,u,w))$. *The relative bias of the strategy is denoted by* $\delta(r,u,w|n) = b(\bar{y}_s, P_0(s|r,u,w))/\sqrt{V(\bar{y}_s, P_0(s|r,u,w))}$ *where* $b(r,u,w)$ *is the bias of* $(\bar{y}_s, P_0(s|r,u,w))$. *The relative efficiency is defined as* $deff(r,u,w|n) = MSE(Y_{[r]}, P_0(s|r,u,w))/V_0(\bar{y}_s)$. *After some calculations we have:*
$\delta(3,213,222|3) = -0.796$, $deff(3,213,222|3) = 0.009$,
$\delta(11,213,222|15) = -0.777$, $deff(11,213,222|15) = 0.05$,
$\delta(22,200,210|29) = -1.604$, $deff(22,200,210|29) = 0.092$.
Thus, in all the considered cases the mean from the conditional simple sample is more accurate than the simple random sample mean.

2.3. Conditional ratio strategy

Let us consider the following ratio-type estimator:

$$\hat{y}_{r,u,w,s} = \bar{y}_s \frac{E_0(\bar{x}_s|r,u,w)}{\bar{x}_s} \tag{44}$$

where $E_0(\bar{x}_s|r,u,w)$ is explained by (23)-(26).

Lemma 2.2. *Under the sampling design defined by (6):*

$$E_0(\hat{y}_{r,u,w,s}|r,u,w) \approx E_0(\bar{y}_s|r,u,w), \tag{45}$$

and

$$V_0(\hat{y}_{r,u,w,s}|r,u,w) \approx V_0(\bar{y}_s|r,u,w) - 2h(r,u,w)V_0(\bar{x}_s,\bar{y}_s|r,u,w)+ \\ + h^2(r,u,w)V_0(\bar{x}_s|r,u,w) \quad (46)$$

where

$$h(r,u,w) = \frac{E_0(\bar{y}_s|r,u,w)}{E_0(\bar{x}_s|r,u,w)}, \quad (47)$$

Expected values $E_0(\bar{y}_s|r,u,w)$ and $E_0(\bar{x}_s|r,u,w)$ are explained by (23)-(30). Expressions (31)-(41) of Lemma 2.1. let approximate variances $V_0(\bar{x}_s|r,u,w)$, $V_0(\bar{y}_s|r,u,w)$ and covariance $V_0(\bar{x}_s,\bar{y}_s|r,u,w)$. The proof is in the Appendix.

Theorem 2.3. *If $E_0(\bar{x}_s|r,u,w) = \bar{x}$, then $(\hat{y}_{r,u,w,s}, P_0(s|r,u,w))$ is approximately unbiased for m_y. Hence: $E_0(\hat{y}_{r,u,w,s}|r,u,w) \approx m_y$.* The proof is similar to the proof of Theorem 2.1 presented in Appendix.

Example 2.5. *We continue the problem formulated in the previous examples. Now the population mean of municipal taxation is estimated on the basis of ratio strategy $(\hat{y}_{r,u,w,s}, P_0(s|r,u,w))$. Some calculations lead to*
$\delta(3,243,252|3) = -1.354$, $deff(3,243,252|3) = 0.010$,
$\delta(11,203,212|15) = 0.119$, $deff(11,203,212|15) = 0.111$,
$\delta(22,203,212|29) = -0.338$, $deff(22,203,212|29) = 0.115$.
In the considered cases the ratio estimator from the conditional simple sample is more accurate than the simple random sample mean. The simpler version of $\hat{y}_{r,u,w,s}$ is as follows:

$$\tilde{y}_{r,u,w,s} = Y_{[r]} \frac{E(X_{(r)}|r,u,w))}{X_{(r)}}, \quad (48)$$

where $E(X_{(r)}|r,u,w)$ is given by (25).

Corollary 2.1. *Under the sampling design defined by expression (6) strategy $(\tilde{y}_{r,u,w,s}, P_0(s|r,u,w))$ is approximately unbiased for m_y and*

$$V_0(\tilde{y}_{r,u,w,s}|r,u,w) \approx V(Y_{[r]}|r,u,w) - 2hV(X_{(r)},Y_{[r]}|r,u,w)+ \\ + h^2V(X_{(r)}|r,u,w) \quad (49)$$

where

$$h = h(r,u,w) = \frac{E(Y_{[r]}|r,u,w)}{E(X_{(r)}|r,u,w)}$$

and $V(X_{(r)},Y_{[r]}|r,u,w)$ are explained by (25), (29) (37). The proof is almost the same as the proof of Theorem 2.1. Strategy $(\tilde{y}_{r,u,w,s}, P_0(s|r,u,w))$ does not depend on the shortest or largest values of the auxiliary variable. Hence, the strategy is e.g. useful when there are right or left censored observations of the auxiliary variable.

Example 2.6. *Now the population mean of municipal taxation is estimated on the basis of the strategy* $(\tilde{y}_{r,u,w,s}, P_0(s|r,u,w))$. *After appropriate calculations, we have:*
$\delta(3,200,210|3) = -1.560$, $deff(3,200,210|3) = 0.009$,
$\delta(11,213,223|15) = -0.628$, $deff(11,213,223|15) = 0.047$,
$\delta(22,220,230|29) = 0.209$, $deff(22,220,230|29) = 0.086$.
In the considered cases ratio estimator $\tilde{y}_{r,u,w,s}$ *from the conditional simple sample is more accurate than the simple random sample mean.*

3. Conclusions

Let M_s be the sample median of the auxiliary variable. Thus, when we assume that the distribution of the auxiliary variable is symmetric then $\bar{x} = Me$, where Me is the population median of the auxiliary variable. When we assume that the distribution of the sample median is an approximation of the distribution of the sample mean $\bar{x}_s$ then $P_0(s|x_u \leq M_s \leq x_w)$ can be treated as an approximation of the conditional simple random sampling design denoted by $P_0(s|x_u \leq \bar{x}_s \leq x_w)$, considered by Royall and Cumberland (1981). This consideration can be generalized to the case when the distribution of the auxiliary variable is not necessary symmetric. It is possible to find such rank r that $|E(X_{(r)}) - \bar{x}| = minimum$. Thus, when we assume that the distribution of $\bar{x}_s$ is sufficiently approximated by the distribution of $X_{(r)}$ then $P_0(s|x_u \leq \bar{x}_s \leq x_w)$ can be approximated by $P_0(s|x_u \leq X_{(r)} \leq x_w)$. We can expect that the sampling design can be useful in the case when there are censored observations of the auxiliary variable as well as when outliers exist.

Acknowledgement

The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education.

REFERENCES

DAVID, H. A., NAGARAJA, H. N., (2003). Order statistics. John Wiley & Sons.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of the sampling without replacement from finite universe. Journal of the American Statistical Association, Vol. 47, pp. 663–685.

ROYALL, R. M., CUMBERLAND, W. G., (1981). An empirical study of the ratio estimator and estimators of its variance. Journal of the American Statistical Association, Vol. 76.

SÄRNDAL, C. E., SWENSSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling. Springer Verlag, New York-Berlin-Heidelberg-London-Paris-Tokyo-Hong Kong-Barcelona-Budapest.

TILLé, Y., (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. International Statistical Review, 66, pp. 303–322.

TILLé, Y., (2006). Sampling Algorithms. Springer.

WYWIAŁ, J. L., (2003). On conditional sampling strategies. Statistical Papers, Vol. 44, 3, pp. 397–419.

WYWIAŁ, J. L., (2014). On conditional simple random sample. Statistics in Transition new series, Vol. 15, 4, pp. 525–534.

APPENDICES

Let $\mathbf{S}_1(r,i)=\mathbf{S}(U_1(i),s_1(i))$ and $\mathbf{S}_2(r,i)=\mathbf{S}(U_2(i),s_2(i))$ be the sample spaces of the samples $s_1(i)$ and $s_2(i)$ selected from the sets $U_1(i)=\{1,...,i-1\}$ and $U_2(i)=\{i+1,...,N\}$, respectively. The samples $s(i)$, $s_1(i)$ and $s_2(i)$, are of size n, $r-1$ and $n-r$, respectively, where $s(i)=s_1(i)\cup\{i\}\cup s_2(i)$ and the index i is fixed, $i=r,...,N-n+r$. Sample $s=s_1\cup\{i\}\cup s_2$ where index i is not fixed although $i=r,...,N-n+r$.
Thus, $\mathbf{S}(r,i)=\mathbf{S}(\{1,...,i-1\},s_1(i))\times\{i\}\times\mathbf{S}(\{i+1,...,N\},s_2(i))$
or $\mathbf{S}(r,i)=\mathbf{S}_1(r,i))\times\{i\}\times\mathbf{S}_2(r,i))$
and $\mathbf{S}(r;u,w)=\mathbf{S}(r,u)\times\mathbf{S}(r,u+1)\times...\times\mathbf{S}(r,i)\times...\times\mathbf{S}(r,w)$
where $\mathbf{S}(r,i)$ was defined in Introduction.

Proof of Lemma 2.1

Let us make the following derivation

$$E_0(\bar{x}_s|r,u,w)=E_0\left(\frac{r-1}{n}\bar{x}_{s_1}+\frac{1}{n}X_{(r)}+\frac{n-r}{n}\bar{x}_{s_2}|r,u,w\right)=$$
$$=\frac{r-1}{n}E_0\left(\bar{x}_{s_1}|r,u,w\right)+\frac{1}{n}E_0\left(X_{(r)}|r,u,w\right)+\frac{n-r}{n}E_0\left(\bar{x}_{s_2}|r,u,w\right),$$

On the basis of Definition 2.1 we have:

$$E_0\left(\bar{x}_{s_1}|r,u,w\right)=\sum_{s\in\mathbf{S}(r;u,w)}\bar{x}_{s_1}P_0(s|r,u,w)=\frac{1}{z(r,u,w)}\sum_{i=u}^{w}\sum_{s\in\mathbf{S}(r,i)}\bar{x}_{s_1(i)}=$$
$$=\frac{1}{(r-1)z(r,u,w)}\sum_{i=u}^{w}\sum_{s(i)\in\mathbf{S}_1(r,i)\times\{i\}\times\mathbf{S}_2(r,i)}\sum_{k\in s_1(i)}x_k=$$
$$=\frac{1}{(r-1)z(r,u,w)}\sum_{i=u}^{w}\sum_{s_1(i)\in\mathbf{S}_1(r,i)}\sum_{k\in s_1(i)}x_k=$$

$$
\begin{aligned}
&= \frac{1}{(r-1)z(r,u,w)} \sum_{i=u}^{w} \sum_{k=1}^{i-1} \binom{i-2}{r-2} \binom{N-i}{n-r} x_k = \\
&\qquad = \frac{1}{(r-1)z(r,u,w)} \sum_{i=u}^{w} \binom{i-1}{r-1} \binom{N-i}{n-r} \frac{r-1}{i-1} \sum_{k=1}^{i-1} x_k = \\
&\qquad = \frac{1}{z(r,u,w)} \sum_{i=u}^{w} \binom{i-1}{r-1} \binom{N-i}{n-r} \bar{x}(1,i-1) = \\
&\qquad = \sum_{i=u}^{w} \bar{x}(1,i-1) P\left(X_{(r)} = x_i | r,u,w\right) = E\left(\bar{X}(1,I_r-1)|r,u,w\right). \qquad (50)
\end{aligned}
$$

The next derivation is:

$$
\begin{aligned}
E_0\left(X_{(r)}|r,u,w\right) &= \frac{1}{z(r,u,w)} \sum_{i=u}^{w} \sum_{s \in \mathbf{S}(r,i)} x_i = \\
&= \frac{1}{z(r,u,w)} \sum_{i=u}^{w} \sum_{s(i) \in \mathbf{S}_1(r,i) \times \{i\} \times \mathbf{S}_2(r,i)} x_i = \\
&= \frac{1}{z(r,u,w)} \sum_{i=u}^{w} \binom{i-1}{r-1} \binom{N-i}{n-r} x_i = \\
&= \sum_{i=u}^{w} x_i P\left(X_{(r)} = x_i | r,u,w\right) = E\left(X_{(r)}|r,u,w\right) \qquad (51)
\end{aligned}
$$

Similar derivation of the parameter $E_0\left(\bar{x}_{s_2}|r,u,w\right)$ and expressions (50), (51) lead to (23).

$$
\begin{aligned}
&V_0(\bar{x}_s, \bar{y}_s | r,u,w) = \\
&\quad = E_0\Bigg(\Bigg(\frac{r-1}{n}\left(\bar{x}_{s_1} - E_0(\bar{x}_{s_1}|r,u,w)\right) + \frac{1}{n}(X_{(r)} - E_0(X_{(r)}|r,u,w)) + \\
&\quad + \frac{n-r}{n}(\bar{x}_{s_2} - E_0(\bar{x}_{s_2}|r,u,w))\Bigg)\Bigg(\frac{r-1}{n}\left(\bar{y}_{s_1} - E_0(\bar{y}_{s_1}|r,u,w)\right) + \\
&\quad + \frac{1}{n}(Y_{[r]} - E_0(Y_{[r]}|r,u,w)) + \frac{n-r}{n}(\bar{y}_{s_2} - E_0(\bar{y}_{s_2}|r,u,w))\Bigg)|r,u,w\Bigg) = \\
&\qquad = \frac{(r-1)^2}{n^2} V_0(\bar{x}_{s_1}, \bar{y}_{s_1}|r,u,w) + \frac{r-1}{n^2} V_0(\bar{x}_{s_1}, Y_{[r]}|r,u,w) +
\end{aligned}
$$

$$+\frac{(r-1)(n-r)}{n^2}V_0(\bar{x}_{s_1},\bar{y}_{s_2}|r,u,w)+\frac{r-1}{n^2}V_0(X_{(r)},\bar{y}_{s_1}|r,u,w)+$$
$$+\frac{1}{n^2}V_0(X_{(r)},Y_{[r]}|P_0(r,u,w))+\frac{n-r}{n^2}V_0(X_{(r)},\bar{y}_{s_2}|r,u,w)+$$
$$+\frac{(r-1)(n-r)}{n^2}V_0(\bar{x}_{s_2},\bar{y}_{s_1}|r,u,w)+\frac{n-r}{n^2}V_0(\bar{x}_{s_2},Y_{[r]}|r,u,w)+$$
$$+\frac{(n-r)^2}{n^2}V_0(\bar{x}_{s_2},\bar{y}_{s_2}|r,u,w) \quad (52)$$

$$V_0(\bar{x}_{s_1},\bar{y}_{s_1}|r,u,w)=$$
$$=\frac{1}{z(r,u,w)}\sum_{s\in\mathbf{S}(r;u,w)}(\bar{x}_{s_1}-E_0(\bar{x}_{s_1}|r,u,w))(\bar{y}_{s_1}-E_0(\bar{y}_{s_1}|r,u,w))=$$
$$=\frac{1}{z(r,u,w)}\sum_{i=u}^{w}\sum_{s\in\mathbf{S}(r,i)}(\bar{x}_{s_1(i)}-E_0(\bar{x}_{s_1}|r,u,w))(\bar{y}_{s_1(i)}-E_0(\bar{y}_{s_1}|r,u,w)).$$

In order to simplify the notation let

$$E(H|r,u,w)=E(H),\qquad V(H,T|r,u,w)=V(H,T),$$

$$p_i=P(X_{(i)}=x_i|r,u,w)=\frac{1}{z(r,u,w)}\binom{i-1}{r-1}\binom{N-i}{n-r}.$$

Let

$$e_k=x_k-E_0(\bar{x}_{s_1}|r,u,w)=x_k-E(\bar{X}(1,I_r-1)|r,u,w)=x_k-E(\bar{X}(1,I_r-1)),\quad (53)$$

$$d_k=y_k-E_0(\bar{y}_{s_1})=y_k-E(\bar{Y}[1,I_r-1]),\quad (54)$$

$$\bar{e}_{s_1(i)}=\frac{1}{r-1}\sum_{k\in s_1(i)}e_k=\bar{x}_{s_1(i)}-E(\bar{X}(1,I_r-1)),\quad (55)$$

$$\bar{d}_{s_1(i)}=\frac{1}{r-1}\sum_{k\in s_1(i)}d_k=\bar{y}_{s_1(i)}-E(\bar{Y}[1,I_r-1]).\quad (56)$$

Thus,

$$V_0(\bar{x}_{s_1}, \bar{y}_{s_1}|r,u,w) = \frac{1}{z(r,u,w)} \sum_{i=u}^{w} \sum_{s_1(i)\in \mathbf{S}_1(r,i)} \bar{e}_{s_1(i)} \bar{d}_{s_1(i)} =$$

$$= \frac{1}{(r-1)^2 z(r,u,w)} \sum_{i=u}^{w} \sum_{s_1(i)\in \mathbf{S}_1(r,i)} \sum_{k\in s_1(i)} e_k \sum_{h\in s_1(i)} d_h =$$

$$= \frac{1}{(r-1)^2 z(r,u,w)} \sum_{i=u}^{w} \sum_{s_1(i)\in \mathbf{S}_1(r,i)} \left(\sum_{k\in s_1(i)} e_k d_k + \sum_{k\in s_1(i)} \sum_{h\in s_1(i), h\neq k} e_k d_h \right) =$$

$$= \frac{1}{(r-1)^2 z(r,u,w)} \sum_{i=u}^{w} \sum_{s_1(i)\in \mathbf{S}_1(r,i)} \sum_{k\in s_1(i)} e_k d_k +$$

$$+ \frac{1}{(r-1)^2 z(r,u,w)} \sum_{i=u}^{w} \sum_{s_1(i)\in \mathbf{S}_1(r,i)} \sum_{k\in s_1(i)} \sum_{h\in s_1(i), h\neq k} e_k d_h =$$

$$= \frac{1}{(r-1)^2 z(r,u,w)} \sum_{i=u}^{w} \binom{i-2}{r-2} \binom{N-i}{n-r} \sum_{k\in U_1(i)} e_k d_k +$$

$$+ \frac{1}{(r-1)^2 z(r,u,w)} \sum_{i=u}^{w} \binom{i-3}{r-3} \binom{N-i}{n-r} \sum_{k\in U_1(i)} \sum_{h\in U_1(i), h\neq k} e_k d_h =$$

$$= \frac{r-1}{(r-1)^2 z(r,u,w)} \sum_{i=u}^{w} \binom{i-1}{r-1} \binom{N-i}{n-r} \frac{1}{i-1} \sum_{k\in U_1(i)} e_k d_k +$$

$$+ \frac{(r-1)(r-2)}{(r-1)^2 z(r,u,w)} \sum_{i=u}^{w} \binom{i-1}{r-1} \binom{N-i}{n-r} \frac{1}{(i-1)(i-2)} \sum_{k\in U_1(i)} \sum_{h\in U_1(i), h\neq k} e_k d_k. \quad (57)$$

Let

$$a_k(i) = x_k - \bar{x}(1, i-1), \qquad b_k(i) = y_k - \bar{y}[1, i-1]. \quad (58)$$

Thus, $\sum_{k\in U_1} a_k(i) = \sum_{k\in U_1} b_k(i) = 0$ and

$$e_k = a_k(i) + \bar{x}(1, i-1) - E(\bar{X}(1, I_r - 1)), \qquad d_k = b_k(i) + \bar{y}[1, i-1] - E(\bar{Y}[1, I_r - 1]). \quad (59)$$

$$
\begin{aligned}
&V_0(\bar{x}_{s_1},\bar{y}_{s_1}|r,u,w) = \\
&\quad = \frac{1}{r-1}\sum_{i=u}^{w} p_i \frac{1}{i-1}\sum_{k\in U_1(i)} (a_k(i)+\bar{x}(1,i-1)-E(\bar{X}(1,I_r-1))) \\
&\qquad (b_k(i)+\bar{y}[1,i-1]-E(\bar{Y}[1,I_r-1]|r,u,w))+ \\
&\qquad + \frac{r-2}{r-1}\sum_{i=u}^{w}\frac{p_i}{(i-1)(i-2)}\sum_{k\in U_1(i)}\sum_{h\in U_1(i),h\neq k} \\
&\quad (a_k(i)+\bar{x}(1,i-1)-E(\bar{X}(1,I_r-1)))(b_h(i)+\bar{y}[1,i-1]-E(\bar{Y}[1,I_r-1]) =
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{r-1}\sum_{i=u}^{w}\frac{p_i}{i-1}\sum_{k\in U_1(i)} a_k(i)b_k(i) + \frac{1}{r-1}\sum_{i=u}^{w}\frac{p_i}{i-1} \\
&\qquad \sum_{k\in U_1(i)} (\bar{x}(1,i-1)-E(\bar{X}(1,I_r-1)))(\bar{y}[1,i-1]-E(\bar{Y}[1,I_r-1]))+
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{r-2}{r-1}\sum_{i=u}^{w}\frac{p_i}{(i-1)(i-2)}\sum_{k\in U_1(i)}\sum_{h\in U_1(i),h\neq k} a_k(i)b_h(i)+ \\
&\quad +\frac{r-2}{r-1}\sum_{i=u}^{w}\frac{(\bar{y}[1,i-1]-E(\bar{Y}[1,I_r-1]))p_i}{(i-1)(i-2)}\sum_{k\in U_1(i)}\sum_{h\in U_1(i),h\neq k} a_k(i)+ \\
&\quad +\frac{r-2}{r-1}\sum_{i=u}^{w}\frac{(\bar{x}(1,i-1)-E(\bar{X}(1,I_r-1))p_i}{(i-1)(i-2)}\sum_{k\in U_1(i)}\sum_{h\in U_1(i),h\neq k} b_h(i)+ \\
&\quad +\frac{r-2}{r-1}\sum_{i=u}^{w}\frac{(\bar{x}(1,i-1)-E(\bar{X}(1,I_r-1)))(\bar{y}[1,i-1]-E(\bar{Y}[1,I_r-1]))p_i}{(i-1)(i-2)} \\
&\qquad\qquad \sum_{k\in U_1(i)}\sum_{h\in U_1(i),h\neq k} 1 =
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{r-1}\sum_{i=u}^{w}\frac{i-2}{i-1}v_{xy}(1,i-1)p_i + \frac{1}{r-1}\sum_{i=u}^{w} p_i(\bar{x}(1,i-1)-E(\bar{X}(1,I_r-1)) \\
&\qquad (\bar{y}[1,i-1]-E(\bar{Y}[1,I_r-1]))\frac{1}{i-1}\sum_{k\in U_1(i)} 1+
\end{aligned}
$$

$$+\frac{r-2}{r-1}\sum_{i=u}^{w}\frac{p_i}{(i-1)(i-2)}\left(\sum_{k\in U_1(i)}a_k(i)\sum_{h\in U_1(i)}b_h(i)-\sum_{k\in U_1(i)}a_k(i)b_k(i)\right)+$$

$$+\frac{r-2}{r-1}\sum_{i=u}^{w}\frac{\bar{y}[1,i-1]-E(\bar{Y}[1,I_r-1]|r,u,w)}{(i-1)(i-2)}p_i\sum_{k\in U_1(i)}a_k(i)\sum_{h\in U_1(i),h\neq k}1+$$

$$+\frac{r-2}{r-1}\sum_{i=u}^{w}\frac{\bar{x}(1,i-1)-E(\bar{X}(1,I_r-1)|r,u,w)}{(i-1)(i-2)}p_i\sum_{h\in U_1(i)}b_h(i)\sum_{k\in U_1(i),k\neq h}1+$$

$$+\frac{r-2}{r-1}\sum_{i=u}^{w}(\bar{x}(1,i-1)-E(\bar{X}(1,I_r-1)))(\bar{y}[1,i-1]-E(\bar{Y}[1,I_r-1]))p_i=$$

$$=\frac{1}{r-1}\sum_{i=u}^{w}\frac{i-2}{i-1}v_{xy}(1,i-1)p_i+$$

$$+\frac{1}{r-1}\sum_{i=u}^{w}(\bar{x}(1,i-1)-E(\bar{X}(1,I_r-1)))(\bar{y}[1,i-1]-E(\bar{Y}[1,I_r-1]))p_i+$$

$$+\frac{r-2}{r-1}\left(V\left(\bar{X}(1,I_r-1),\bar{Y}[1,I_r-1]\right)-\sum_{i=u}^{w}\frac{\sum_{k\in U_1(i)}a_k(i)b_k(i)}{(i-1)(i-2)}p_i\right)=$$

$$=\frac{1}{r-1}\sum_{i=u}^{w}\frac{i-r}{i-1}v_{xy}(1,i-1)p_i+V(\bar{X}(1,I_r-1),\bar{Y}(1,I_r-1)|r,u,w)=$$

$$=\frac{1}{r-1}E\left(\frac{I-r}{I-1}V_{xy}(1,I_r-1)\right)+V(\bar{X}(1,I_r-1),\bar{Y}(1,I_r-1)|r,u,w).$$

Derivations of other expression of Lemma 4.1 are similar to the above ones.

Proof of Theorem 2.1

$$E_0(\bar{y}_s|r,u,w)=$$

$$=\sum_{i=u}^{w}\left(\frac{r-1}{n}\bar{y}[1,i-1]+\frac{y_i}{n}+\frac{n-r}{n}\bar{y}[i+1,N]\right)P\left(X_{(r)}=x_i|r,u,w\right)=$$

$$=\sum_{i=u}^{w}\left(\frac{r-1}{n(i-1)}\sum_{k=1}^{i-1}y_k+\frac{y_i}{n}+\frac{n-r}{n(N-i)}\sum_{k=i+1}^{N}y_k\right)P\left(X_{(r)}=x_i|r,u,w\right)\approx$$

$$\approx\sum_{i=u}^{w}\left(\frac{r-1}{n(i-1)}\sum_{k=1}^{i-1}(\bar{y}+a(x_k-\bar{x}))+\frac{\bar{y}+a(x_i-\bar{x})}{n}+\right.$$

$$+\frac{n-r}{n(N-i)}\sum_{k=i+1}^{N}(\bar{y}+a(x_k-\bar{x}))\Big)P\left(X_{(r)}=x_i|r,u,w\right)=$$

$$=\bar{y}-a\bar{x}+a\sum_{i=u}^{w}\left(\frac{r-1}{n(i-1)}\sum_{k=1}^{i-1}x_k+\frac{x_i}{n}+\frac{n-r}{n(N-i)}\sum_{k=i+1}^{N}x_k\right)P\left(X_{(r)}=x_i|r,u,w\right)=$$

$$=\bar{y}-a\bar{x}+a\sum_{i=u}^{w}\left(\frac{r-1}{n}\bar{x}(1,i-1)+\frac{x_i}{n}+\frac{n-r}{n}\bar{x}(i+1,N)\right)P\left(X_{(r)}=x_i|r,u,w\right)=$$

$$=\bar{y}+a\left(E_0(\bar{x}_s|r,u,w)-\bar{x}\right).$$

Thus, the proof is completed.

Proof of Lemma 2.2

Estimator $\hat{y}_{r,u,w,s}=\bar{y}_s\frac{E_0(\bar{x}_s|r,u,w)}{\bar{x}_s}$ can be treated as the function of statistics $\bar{x}_s$ and $\bar{y}_s$ denoted by $f(\bar{x}_s,\bar{y}_s)$. The first derivative of $f(\bar{x}_s,\bar{y}_s)$ in points $\bar{x}_s=E_0(\bar{x}_s|r,u,w)$ and $\bar{y}_s=E_0(\bar{y}_s|r,u,w)$ are as follows: $f_x=\frac{\partial f}{\partial\bar{x}_s}=-h$ where $h=\frac{E_0(\bar{y}_s|r,u,w)}{E_0(\bar{x}_s|r,u,w)}$ and $f_y=\frac{\partial f}{\partial\bar{y}_s}=1$, respectively. This let us write the following Taylor's linearisation of $\hat{y}_{r,u,w,s}$:

$$\hat{y}_{r,u,w,s}-E_0(\bar{y}_s|r,u,w)\approx(\bar{y}_s-E_0(\bar{y}_s|r,u,w))-h(\bar{x}_s-E_0(\bar{x}_s|r,u,w))$$

This leads to the derivation of expressions (45) - (47).

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 429–447

PREDICTION OF A FUNCTION OF MISCLASSIFIED BINARY DATA

Noriah M. Al-Kandari[1], **Partha Lahiri**[2]

ABSTRACT

We consider the problem of predicting a function of misclassified binary variables. We make an interesting observation that the naive predictor, which ignores the misclassification errors, is unbiased even if the total misclassification error is high as long as the probabilities of false positives and false negatives are identical. Other than this case, the bias of the naive predictor depends on the misclassification distribution and the magnitude of the bias can be high in certain cases. We correct the bias of the naive predictor using a double sampling idea where both inaccurate and accurate measurements are taken on the binary variable for all the units of a sample drawn from the original data using a probability sampling scheme. Using this additional information and design-based sample survey theory, we derive a bias-corrected predictor. We examine the cases where the new bias-corrected predictors can also improve over the naive predictor in terms of mean square error (MSE).

Key words: binary classification, double sampling, finite population sampling, misclassification, linkage error, sampling design.

1. Introduction

In many disciplines, misclassified binary data are frequently encountered. For example, in device testing, Zhong (2002) studied the specificity and sensitivity of an inaccurate diagnostic test along with a gold standard. Stamey *et al.* (2007) proposed a Bayesian estimation of an intervention effect with pre and post misclassified binomial data. Lyles *et al.* (2004) discussed single-armed studies with misclassification of a repeated binary outcome. In epidemiology and medical studies, there are plenty of examples of misclassified binary data. For example, in studying the relationship between low level radiation exposure and cancer death rate using the Cox proportional hazard model, Krewski *et al.* (2005) noted that misclassified binary data arise in form of imperfect linkages caused by the computerized record linkage method.

Bross (1954) was probably the first to observe that classical estimators of the odds ratio can be heavily biased if the misclassification error in binary data is ignored; see Goldberg (1975) for a follow-up study. Neter *et al.* (1965) noticed that

[1] Department of Statistics and Operations Research, Kuwait University.
E-mail: noriah@stat.kuniv.edu.

[2] Joint Program in Survey Methodology, University of Maryland. E-mail: plahiri@umd.edu.

the matching errors pose an obstacle to the usefulness and correct interpretation of record checks.

There are mainly two different approaches available to correct for the bias in statistical procedures that arise from misclassified binary data. The key ingredient in both the approaches is to use additional data to deal with the identifiability problem. The first approach, pioneered by Tenenbein (1970), employs a double sampling scheme in which a training data set is collected and the binary responses are measured by an accurate instrument (in the case of a random subsample from the original data) or by both an accurate instrument and the same inaccurate instrument used to collect the original data (in the case of an independent new sample). An accurate instrument results in error-free but expensive binary data. On the other hand, an inaccurate instrument results in misclassified but relatively less expensive binary data. Tenenbein's idea is intuitive and uses both accurate and inaccurate procedures to yield not only model identifiability but also economical viability.

For the single proportion problem, when a training data is obtained using a double sampling scheme, Tenenbein (1970) proposed a maximum likelihood estimator and derived its asymptotic variance. Boese *et al.* (2006) constructed several likelihood-based confidence intervals for a proportion using data subject to only false positive misclassification. Rahardja and Zhou (2013) proposed a modification of the Wald test in presence of misclassified binary data and applied their test to traffic data. Rahardja and Yang (2015) constructed two likelihood-based confidence intervals for a binomial proportion parameter using a double-sampling scheme with misclassified binary data.

When an accurate instrument is unavailable or prohibitively expensive but certain data related to the cause of misclassification are available, one can develop an identifiable model in an attempt to correct for the misclassification bias in the estimators and predictors. For the single proportion problem using misclassified data with no training data, Gaba and Winkler (1992) and Viana *et al.* (1993) developed Bayesian approaches with highly informative priors. Bayesian inferences with informative priors were also developed for two-sample problems for two proportions. For example, see Evans *et al.* (1996) for risk difference (the difference of two proportions) and Gustafson *et al.* (2001) for odds ratios. Lahiri and Larsen (2005) used a mixture model to correct for the bias of the ordinary least square estimators of regression coefficients due to imperfect linkages.

In this paper, we assume the existence of a training sample such as the one proposed by Tenenbien (1970) and exploit a design-based sample survey approach to predict a function of misclassified binary data. Consider a set U of N units. For unit $i \in U$, we define a binary variable δ_i taking on values 0 and 1, and a $K \times 1$ vector of measurements $\mathbf{y}_i = (y_{i1}, \cdots, y_{iK})\prime$. We consider a situation when we do not observe δ_i, but instead observe a predictor $\hat{\delta}_i$ subject to a misclassification error $e_i = \hat{\delta}_i - \delta_i$ $(i \in U)$. In this paper, we are interested in the prediction of $\mathbf{Y}_\delta =$

$\sum_{i\in U}\delta_i\mathbf{y}_i$ or a non-linear function of the components of $\mathbf{Y}_\delta$, say $f(Y_{\delta 1},\cdots,Y_{\delta K})$, where $Y_{\delta k}=\sum_{i=1}^N \delta_i y_{ik}$, $(k=1,\cdots,K)$, based on data $\{(\hat{\delta}_i,\mathbf{y}_i),\ i\in S\subseteq U\}$.

A natural predictor of $\mathbf{Y}_\delta$ is given by $\mathbf{Y}_{\hat{\delta}}(S)=\sum_{i\in S}\hat{\delta}_i\mathbf{y}_i$. If additional data that explain the mechanism for misclassification errors e_i are available, it is possible to correct $\mathbf{Y}_{\hat{\delta}}(S)$ for bias due to the misclassification errors. The misclassification errors could arise due to a variety of reasons. For example, the data set may be obtained by merging two or more data sets using a computerized record linkage method, which may introduce misclassification errors due to incorrect linkages. There are a large number of papers available in the literature that provide valid inferences under linkage errors when data on linkage error mechanism through matching weights are available; for a review of record linkage methodology, see Felligi and Sunter (1969) and Herzog *et al.* (2007). However, in this paper, we assume that we do not have any data that explain the misclassification error for all records in $i\in S$. Thus, even for the special case when the misclassification errors is due to incorrect linkages, in this paper we deal with a situation that cannot be handled by a regular record linkage methodology such as the ones given in Scheuren and Winkler (1993) and Lahiri and Larsen (2005).

In Section 2.1, we first obtain the bias and mean squared error of the natural predictor for the scaler case $K=1$ and then generalize to $K\geq 1$. The extent of the bias clearly depends on the misclassification error distribution. When the positive and negative misclassification errors are equally likely, $Y_{\hat{\delta}}(S)\equiv Y_{\hat{\delta}}$ turns out to be an unbiased predictor of Y_δ. This is an interesting observation. We propose a method for correcting the bias in the general case by drawing a probability sample and then obtaining the misclassification errors e_i for all the units in the sample. Using this extra information, we propose a bias-corrected predictor of Y_δ. We obtain an exact expression for the MSE of the proposed bias-corrected predictor that incorporates both the sampling and misclassification errors. We also propose an estimator of the MSE of the new predictor. In Section 2.2, we discuss the estimation of relative risk estimation as an illustration of the methodology proposed in Section 2.1. In Section 2.3, we evaluate the method proposed in Section 2.1 using a numerical example. In Section 3, we consider the case $S\subseteq U$. Finally, some conclusions are presented in Section 4.

2. Prediction of $\mathbf{Y}_\delta$ when $S=U$

2.1. The Methodology

For the simplicity of exposition, we first consider the scaler case, i.e., $K=1$. We define the bias and mean squared error (MSE) of $\hat{Y}_\delta$ as follows:

$$\begin{aligned}\text{Bias}_M(\hat{Y}_\delta) &= \mathrm{E}_M(\hat{Y}_\delta-Y_\delta),\\ \text{MSE}_M(\hat{Y}_\delta) &= \mathtt{Var}_M(\hat{Y}_\delta-Y_\delta),\end{aligned}$$

Table 1. The probability distribution of the misclassification error e_i (probabilities are given within parenthesis)

	δ_i	
$\hat{\delta}_i$	0	1
0	0 (p_{i00})	-1 (p_{i01})
1	1 (p_{i10})	0 (p_{i11})

where E_M and Var_M denote the expectation and variance with respect to a misclassification model described by the two-way table given in Table 1.

For unit $i \in U$, the table displays the misclassification error distribution, where $p_{i11}+p_{i10}+p_{i01}+p_{i00}=1$. We call p_{i10} and p_{i01} false positive and false negative probabilities, respectively. We say that we have *high*, *moderate* and *low* linkage errors if $p_{i10}+p_{i01}=p_{i;T}$ (say) is close to 1, 0.5 and 0, respectively.

Theorem 1. Under the misclassification model given in Table 1, we have

$$\begin{aligned}\text{(i) } \mathrm{Bias}_M(\hat{Y}_\delta) &= \sum_{i\in U} y_i p_{i;D} = Y_{p_D} \text{ (say)},\\ \text{(ii) } \mathrm{MSE}_M(\hat{Y}_\delta) &= \sum_{i\in U}\left[p_{i;T}-p_{i;D}^2\right]y_i^2,\end{aligned}$$

where $p_{i;D}=p_{i10}-p_{i01}$ and $p_{i;T}=p_{i10}+p_{i01}$ $(i\in U)$.

Proof: First note that $Y_{\hat{\delta}}-Y_\delta=\sum_{i\in U} y_i e_i = Y_e$, (say). Under the misclassification model, we have $\mathrm{E}_M(e_i)=p_{i;D}$ and $\mathrm{E}_M(e_i^2)=p_{i;T}$. Thus, part (i) follows immediately. To prove part (ii), using $\mathrm{Cov}_M(e_i,e_j)=0$ $(i\neq j\in U)$, we have

$$\begin{aligned}\mathrm{MSE}_M(Y_{\hat{\delta}}) &= \mathrm{E}_M\left(\sum_{i\in U} y_i e_i\right)^2 - \left[\mathrm{E}_M(\sum_{i\in U} y_i e_i)\right]^2\\ &= \mathrm{E}_M\left(\sum_{i\in U} y_i^2 e_i^2 + \sum_{i\neq j} y_i y_j e_i e_j\right) - \left(\sum_{i\in U} y_i p_{i;D}\right)^2\\ &= \sum_{i=1}^N y_i^2 p_{i;T} + \sum_{i\neq j} y_i y_j p_{i;D} p_{j;D} - \left(\sum_{i=1}^N y_i p_{i;D}\right)^2.\end{aligned}$$

Part (ii) now follows using algebra.

Throughout the paper, we assume that we do not have any additional data that explain the misclassification errors e_i for all units in U. Thus, we propose to draw a sample s_1 of size n from U, using a probability sampling scheme. For each unit in the sample, we assume that we can obtain e_i with some extra effort. Let $\pi_i = \mathtt{Pr}(s_1 \ni i)$ denote the first-order inclusion probability of unit $i\in U$. We propose to estimate Y_δ by $\hat{Y} = Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e}$, where $\hat{Y}_{\pi^{-1}e} = \sum_{i\in s_1} \pi_i^{-1} e_i y_i$. The following theorem shows that $\hat{Y}$ is an unbiased predictor of Y_δ. Moreover, the theorem provides an expression for the total MSE of $\hat{Y}$, where total MSE incorporate errors due to both the misclassification and sampling errors.

Theorem 2. Under the sampling design and misclassification model, we have

$$\begin{aligned}&\text{(i) Bias}(\hat{Y}) = 0,\\ &\text{(ii) MSE}(\hat{Y}) = \sum_{i\in U}\sum_{j>i\in U} (\pi_i\pi_j - \pi_{ij})\psi_{ij},\\ &\text{(iii) E}\left[mse(\hat{Y})\right] = \text{MSE}(\hat{Y}),\end{aligned}$$

where $\pi_{ij} = \texttt{Pr}(s_1 \ni \{ij\})$, the second-order inclusion probability, $\psi_{ij} = \pi_i^{-2} y_i^2 p_{iT} + \pi_j^{-2} y_j^2 p_{jT} - 2(\pi_i\pi_j)^{-1} y_i y_j p_{i;D} p_{j;D}$, $\text{mse}(\hat{Y}) = \sum_{i\in s_1}\sum_{j>i\in s_1} \pi_{ij}^{-1}(\pi_i\pi_j - \pi_{ij})\left(\pi_i^{-1} y_i e_i - \pi_j^{-1} y_j e_j\right)^2$.

Proof: Let $\texttt{E}_d$ and $\texttt{Var}_d$ denote the expectation and variance with respect to the sample design. First note that $\texttt{E}_d\left(\hat{Y}_{\pi^{-1}e}\right) = Y_e = Y_{\hat{\delta}} - Y_\delta$, since $\hat{Y}_{\pi^{-1}e}$ is the Horvitz-Thompson estimator of Y_e. To prove part (i) of Theorem 2, note that

$$\begin{aligned}\text{Bias}(\hat{Y}) &= \texttt{E}(\hat{Y} - Y_\delta) = \texttt{E}(Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e} - Y_\delta)\\ &= \texttt{E}_M E_d(Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e} - Y_\delta)\\ &= \texttt{E}_M(Y_{\hat{\delta}} - Y_e - Y_\delta)\\ &= \texttt{E}_M(0)\\ &= 0,\end{aligned}$$

since $Y_{\hat{\delta}} - Y_\delta = 0$. To prove part (ii), we first apply the iterated formula for variance to obtain

$$\begin{aligned}\text{MSE}(\hat{Y}) &= \texttt{Var}(\hat{Y} - Y_\delta)\\ &= \texttt{E}_M\texttt{Var}_d(\hat{Y} - Y_\delta) + \texttt{Var}_M\texttt{E}_d(\hat{Y} - Y_\delta)\\ &= \texttt{E}_M\texttt{Var}_d(Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e} - Y_\delta) + \texttt{Var}_M\texttt{E}_d(Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e} - Y_\delta)\\ &= \texttt{E}_M\texttt{Var}_d(\hat{Y}_{\pi^{-1}e}) + \texttt{Var}_M(Y_{\hat{\delta}} - Y_e - Y_\delta)\\ &= \texttt{E}_M\left[\sum_{i\in U}\sum_{j>i\in U}(\pi_i\pi_j - \pi_{ij})\frac{y_i e_i y_j e_j}{\pi_i\pi_j}\right],\end{aligned}$$

since $Y_{\hat{\delta}} - Y_e - Y_\delta = 0$. Now, part (ii) follows using $\texttt{E}_M(e_i) = p_{i;D}$, $\texttt{E}_M(e_i^2) = p_{i;T}$, and $\text{Cov}_M(e_i, e_j) = 0$, $(i \neq j)$, and algebra. Part (iii) follows using the iterated formula for expectation and the design-unbiasedness of the well-known Yates-Grundy estimator, Yates and Grundy (1953).

We now turn our attention to the estimation of $f(\mathbf{Y}_\delta)$, where $\mathbf{Y}_\delta = (Y_{\delta 1}, \cdots, Y_{\delta K})'$. A natural estimator is given by $f(\mathbf{Y}_{\hat{\delta}})$. Using the Taylor's series argument, it can be shown that

$$\mathrm{E}_M\left[f(\mathbf{Y}_{\hat{\delta}}) - f(\mathbf{Y}_\delta)\right] \doteq [\nabla f(\mathbf{Y}_\delta)]' \mathrm{E}_M(\mathbf{Y}_{\hat{\delta}} - \mathbf{Y}_\delta) = [\nabla f(\mathbf{Y}_\delta)]' \mathbf{Y}_{p_D},$$

where $\nabla f(\mathbf{Y}_\delta)$ is the gradient of $f(\mathbf{Y}_\delta)$, $\mathbf{Y}_{p_D} = (Y_{p_D 1}, \cdots, Y_{p_D K})'$ and $Y_{p_D k} = \sum_{i \in U} p_{i;D} y_{ik}$ $(k = 1, \cdots, K)$. Thus, $f(\mathbf{Y}_{\hat{\delta}})$ is biased for $f(\mathbf{Y}_\delta)$. A bias-adjusted estimator is given by $f(\hat{\mathbf{Y}})$, where $\hat{\mathbf{Y}} = (Y_{\hat{\delta}1} - \hat{Y}_{\pi^{-1}e1}, \cdots, Y_{\hat{\delta}1} - \hat{Y}_{\pi^{-1}eK})'$.

Theorem 3 below is useful in obtaining the total MSE of $f(\hat{\mathbf{Y}})$. Define $\Sigma \equiv ((\sigma_{kl})) = \mathtt{Var}(\mathbf{Y}_{\hat{\delta}} - \mathbf{Y}_\delta)$, the $K \times K$ covariance matrix of $\mathbf{Y}_{\hat{\delta}} - \mathbf{Y}_\delta$, where $\sigma_{kl} = \mathrm{Cov}(\hat{Y}_k - Y_{\delta k},\ \hat{Y}_l - Y_{\delta l})$, $(k, l = 1, \cdots, K)$. Note that we can write $\sigma_{kl} = (\sigma_{k+l} - \sigma_{kk} - \sigma_{ll})/2$, where σ_{k+l} is obtained from $\mathtt{Var}(\hat{Y}_k - Y_{\delta k})$ when we replace y_{ik} by $y_{ik} + y_{il}$ $(k \neq l,\ k, l = 1, \cdots, K)$.

Theorem 3. Under the misclassification model, we have

$$\mathrm{MSE}\left[f(\hat{\mathbf{Y}})\right] \approx [\nabla f(\mathbf{Y}_\delta)]' \Sigma [\nabla f(\mathbf{Y}_\delta)],$$

where

$$\begin{aligned}\sigma_{kk} &= \mathrm{E}_M \left\{ \sum_{i \in U} \sum_{j > i \in U} (\pi_i \pi_j - \pi_{ij}) \left(\pi_i^{-1} y_{ik} e_i - \pi_j^{-1} y_{jk} e_j\right)^2 \right\} \\ &= \sum_{i \in U} \sum_{j > i \in U} (\pi_i \pi_j - \pi_{ij}) \psi_{ij;kk},\end{aligned}$$

with

$$\psi_{ij;kk} = \pi_i^{-2} p_{i;T} y_{ik}^2 + \pi_j^{-2} p_{j;T} y_{jk}^2 - 2(\pi_i \pi_j)^{-1} p_{i;D} p_{j;D} y_{ik} y_{jk}.$$

We propose to estimate σ_{kk} by $\hat{\sigma}_{kk} = \sum_{i \in s_1} \sum_{j > i \in s_1} \pi_{ij}^{-1} (\pi_i \pi_j - \pi_{ij}) \left(\pi_i^{-1} y_{ik} e_i - \pi_j^{-1} y_{jk} e_j\right)^2$, $(k = 1, \cdots, K)$. Using the property of the Yates-Grundy estimator, we have $\mathrm{E}_d(\hat{\sigma}_{kk}) = \sum_{i \in U} \sum_{j > i \in U} (\pi_i \pi_j - \pi_{ij}) \left(\pi_i^{-1} y_{ik} e_i - \pi_j^{-1} y_{jk} e_j\right)^2$ and hence $\mathrm{E}(\hat{\sigma}_{kk}) = \mathrm{E}_M \mathrm{E}_d(\hat{\sigma}_{kk}) = \sigma_{kk}$. Thus, $\hat{\sigma}_{kk}$ is an unbiased estimator of σ_{kk} $(k = 1, \cdots, K)$. Thus, an unbiased estimator of σ_{kl} is given by $\hat{\sigma}_{kl} = (\hat{\sigma}_{k+l} - \hat{\sigma}_{kk} - \hat{\sigma}_{ll})/2$, $(k \neq l,\ k, l = 1, \cdots, K)$. Thus, an approximately unbiased estimator of MSE $\left[f(\hat{\mathbf{Y}})\right]$ is given by

$$\mathrm{mse}\left[f(\hat{\mathbf{Y}})\right] = \left[\nabla f(\hat{\mathbf{Y}})\right]' \hat{\Sigma} \left[\nabla f(\hat{\mathbf{Y}})\right].$$

2.2. An illustrative example: bias adjusted SMR and relative regression coefficients in the presence of linkage errors

There has been an increasing use of computerized record linkage (CRL) method in various studies such as historical cohort mortality studies, cancer studies, political studies and crime studies, in several countries, Howe (1985, 1998), Bennell *et al.* (2012), Giraud-Carrier *et al.* (2015). With very little effort, the method enables us to collect a large amount of data by linking records of human exposure to environmental hazards with records on health status. Since CRL utilizes already existing databases, it saves a substantial amount of money to collect new data. Various government agencies have developed sophisticated software to implement CRL, usually attaching weights reflecting the likelihood of a match to pairs of records.

Fair (1989) listed a number of health studies where environmental exposure data were linked to the Canadian Mortality Data Base (CMDB). Krewski *et al.* (2005) provided an example where National Dose Registry (NDR) of Canada has been linked to CMDB in order to study the associations between excess mortality due to cancer and occupational exposure to low levels of ionizing radiation. In Beauchamp *et al.* (2011), a sample of 2000 participants from a cohort study was linked to a statewide hospitalisations dataset in Victoria, Australia using the national health insurance (Medicare) number and demographic data as identifying variables. Kabudula *et al.* (2014) applied deterministic and probabilistic record linkage approaches to mortality records from 2006 to 2009 from the Agincourt Health and Demographic Surveillance Systems (HDSS) to those in the national civil registration (CR) in South Africa.

In a cohort mortality study, CRL method introduces two types of linkage errors. The Type I linkage error (usually called a false positive) occurs when a member of the cohort who is alive is incorrectly identified as dead. The Type II (or a false negative) error occurs when a member of the cohort who is dead is incorrectly identified as alive. Krewski *et al.* (2005) investigated the impact of linkage errors on estimates of epidemiological indicators of risk such as standardized mortality ratio (SMR) and the parameters of relative risk regression model. Their analytical and simulation results indicate that these indicators are, in general, subject to biases and additional variabilities in the presence of linkage errors.

In this subsection, we use the notation used in Krewski *et al.* (2005). In the analysis of cohort studies, mortality is usually characterized by the hazard function which relates death rate as a function of time. Denoting T the time of death, the hazard function at time u is defined as

$$\lambda(u) = \lim_{\triangle u \downarrow 0} \frac{\mathtt{Pr}\{u \leq T < u + \triangle u \mid T \geq u\}}{\triangle u}.$$

The corresponding survival function and the probability density function are given by $S(u) = \mathtt{exp}\big(-\int_0^u \lambda(t)dt\big)$ and $f(u) = \lambda(u)S(u)$, respectively. Let $\lambda_i(u)$ and $\mathbf{Z}_i(u)$ be the hazard function for a specific cause of death and the value of the vector of covariates at time u for the ith member of the cohort, $i = 1, \cdots, N$. The relative risk regression is then described as

$$\lambda_i(u) = \lambda^*(u)\gamma\{\mathbf{z}_i(u)'\beta\},$$

where $\lambda^*(u)$, value of $\lambda_i(u)$ when $\beta = 0$, is known as the baseline hazard and γ is a positive function of the covariates and β.

Let t_i^0 and t_i^1 be the age at the time of entry into the study and time of loss to follow up for the ith member of the cohort $i = 1, \cdots, N$. Let $\delta_i = 1$ if the ith individual has died at the time of loss to follow-up and $\delta_i = 0$ otherwise. The likelihood based on the relative risk regression model is given by $L = \prod_{i=1}^N f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$. The corresponding log-likelihood is given by

$$\log L = \sum_{i=1}^N \left\{ \delta_i \mathtt{log}\left(\gamma\{\mathbf{z}_i(t_i^1)'\beta\}\right) - \int_{t_i^0}^{t_i^1} \gamma\{\mathbf{z}_i(u)'\beta\}\lambda^*(u)du \right\}.$$

The maximum likelihood estimate $\hat{\beta}$ of β is obtained as a solution of $\frac{\partial \log L}{\partial \beta} = 0$. Note that when $\mathbf{z}_i(u) = 1$, $\gamma\{\hat{\beta}\}$ reduces to the standardized mortality rate given by $SMR = OBS/EXP$, where $OBS = \sum_{i=1}^N \delta_i =$ observed number of death before time to follow-up and $EXP = \sum_{i=1}^N e_i =$ expected number of deaths, with $e_i = \int_{t_i^0}^{t_i^1} \lambda^*(u)du$.

Due to time-dependent covariates $\mathbf{z}_i(u)$, the integral must be re-evaluated at each iteration of the maximization process. Thus, it is very computer-intensive, specially when the cohort size is large. Breslow *et al.* (1983) simplified the likelihood by assuming $\mathbf{z}_i(u) = \mathbf{z}_j$ whenever the ith cohort member passes through the state S_j $(j = 1, \cdots, J)$. The states can be defined by cross-classification of the covariates of interest. In this case, the log-likelihood can be written as

$$\log L = \sum_{i=1}^N \left\{ d_j \mathtt{log}(\gamma\{\mathbf{z}_j'\beta\}) - \gamma\{\mathbf{z}_j'\beta\}e_j \right\},$$

where $e_j = \sum_{i=1}^N \int_{[\mathbf{z}_i(u)\in S_j]} \lambda^*(u)du$ is the contribution to the expected number of deaths from all person-years of observation in the state S_j and d_j is the total number of death in that state. The maximum likelihood estimate of β is then obtained as a solution to the following equation:

$$\sum_{i=1}^J \frac{\partial \Lambda_j(\beta)}{\partial \beta} \left\{ d_j - \mathtt{exp}\{\Lambda_j(\beta)\}e_j \right\} = 0,$$

where $\Lambda_j(\beta) = \log\left(\gamma\{\mathbf{z}_j{\prime}\beta\}\right)$.

Note that the results of Section 2.1 are valid for each state $S_j, j = 1, \cdots, J$. We introduce an additional suffix j to indicate that the parameter or estimator refers to the state $S_j, j = 1, \cdots, J$. For example, we shall have Y_j in place of Y. Note that with $Y_{1ij} = 1$ and $Y_{0ij} = 0$, Y_j reduces to d_j of Krewski *et al.* (2005). We have $Y_j = e_j$

if we choose $Y_{1ij} = \int_{t_{ij}^0}^{t_{ij}^1} \lambda^*(u)du$ and $Y_{0ij} = \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du$. Consider $f(X,Y) = X/Y$ with $X = \sum_{j=1}^J d_j = d$ and $Y = \sum_{j=1}^J e_j = e$. With these choices, $f(X,Y) = SMR$.

Suppose a sample of size n_j is selected using a probability sampling scheme from each state S_j $, j = 1, \cdots, J$. Let $\pi_{i(j)}$ and $\pi_{ik(j)}$ denote the first order and second order inclusion probabilities in state S_j $, j = 1, \cdots, J$. Using the results of Section 2.1, *SMR* can be adjusted for its bias due to linkage errors. The bias-corrected *SMR* is given by $\widehat{SMR} = (\hat{d}/\hat{e})$, where $\hat{d} = \sum_{j=1}^J \hat{d}_j, \hat{e} = \sum_{j=1}^J \hat{e}_j, \widehat{\Delta d_j} = \sum_{i\in s} \pi_{i(j)}^{-1} \Delta\delta_{i(j)}, \widehat{\Delta e_j} = \sum_{i\in s} \pi_{i(j)}^{-1} \Delta\delta_{i(j)} \int_{min(t_{ij}^1, t_{ij}^2)}^{t_{ij}^2} \lambda^*(u)du$, $\hat{d}_j = d_j^L - \Delta\hat{d}_j$ and $\hat{e}_j = e_j^L + \Delta\hat{e}_j$.

An application of Theorem 3 provides us with an estimator of the variance of $\widehat{SMR} - SMR$. It is given by

$$\mathtt{var}(\widehat{SMR} - SMR) = \hat{e}^{-2}\left[\mathtt{var}(\hat{d} - d) + \hat{d}^2\hat{e}^{-2}\mathtt{var}(\hat{e} - e) + 2\hat{d}\hat{e}^{-1}\mathtt{cov}(\hat{d} - d, \hat{e} - e)\right],$$

where $\mathtt{var}(\hat{d} - d) = \sum_{j=1}^J v_j$ with $v_j = \sum_i^{n_j} \sum_{k>i}^{n_j} \pi_{ik(j)}{}^{-1}(\pi_{i(j)}\pi_{k(j)} - \pi_{ik(j)})(\pi_{i(j)}^{-1}\Delta\delta_{i(j)} - \pi_{k(j)}^{-1}\Delta\delta_{k(j)})^2$. We can define $\mathtt{var}(\hat{e} - e)$ and $\mathtt{cov}(\hat{d} - d, \hat{e} - e)$ similarly.

Let us now consider estimation of the regression coefficient β in the relative risk regression model. First, we propose to adjust the log-likelihood given in Krewski *et al.* (2005). We shall find the estimator of β as a solution, say, $\hat{\beta}$, of the following score function

$$Q = Q(\beta; (\hat{d}_j, \hat{e}_j), j = 1, \cdots, J) = \sum_{j=1}^J \frac{\partial \Lambda_j(\beta)}{\partial \beta} \left\{\hat{d}_j - \mathtt{exp}\{\Lambda_j(\beta)\}\hat{e}_j\right\} = 0.$$

Since E $\left\{Q(\beta; (\hat{d}_j, \hat{e}_j), j = 1, \cdots, J) - Q(\beta; (d_j, e_j), j = 1, \cdots, J)\right\} = 0$, we can expect $\hat{\beta}$ to perform well. The covariance matrix of the proposed estimator, $\hat{\beta}$, can be estimated by $\left[\frac{\partial Q}{\partial \beta}\right]^2_{\beta=\hat{\beta}}$.

2.3. Evaluation

In this subsection, we consider the prediction of $Y = \sum_{i=1}^N \delta_i$. A naive predictor of Y that ignores the misclassification errors is given by $\hat{Y} = \sum_{i=1}^N \hat{\delta}_i$. Under the misclassification error model of Section 2.1 with $p_{ikl} = p_{kl}$, $(i = 1, \cdots, N;\ k, l = 0, 1)$, the misclassification error bias of $\hat{Y}$ can be obtained as

$$\text{Bias}_M(\hat{Y}) = N(p_{10} - p_{01}) = Nb,$$

where $b = p_{10} - p_{01}$. Thus, $\hat{Y}$ is positively (negatively) biased if the false positive probability is more (less) than the false negative probability. It is interesting to note that there is no bias in $\hat{Y}$ due to misclassification even if there is a large misclassification error as long as the false positive and false negative probabilities are identical.

Using the theory developed in Section 2.1, we can correct the misclassification bias of $\hat{Y}$ by drawing a simple random sample (SRS) of size n from U and determining the status of each sampled unit for misclassification error. We propose the misclassification bias-corrected predictor of Y as

$$\tilde{Y} = \hat{Y} - \widehat{Bias},$$

where

$$\widehat{Bias} = \frac{N}{n}\sum_{i\in s}^{n} e_i.$$

Evidently, the proposed predictor is unbiased with respect to the combined distribution of misclassification and sampling errors. But, the proposed method introduces some costs. Also, the bias correction is expected to increase the variability. Thus, we study how the above bias-correction affects the mean squared error that incorporates both the misclassification and sampling errors. We define the total mean square error of a predictor $\hat{Y}$ of Y as

$$\mathrm{MSE}(\hat{Y}) = \mathtt{E}(\hat{Y} - Y)^2,$$

where the expectation $\mathtt{E}$ is with respect to the SRS and the multinomial misclassification error model. After considerable algebra, we obtain

$$\mathrm{MSE}(\hat{Y}) = N(a + Nb^2),$$
$$\mathrm{MSE}(\tilde{Y}) = \frac{N^2}{n}(1-f)a,$$

where $a = p_{10}(1-p_{10}) + p_{01}(1-p_{01}) + 2p_{10}p_{01}$ and $f = \frac{n}{N}$.

We define the relative improvement in MSE (MSERI) as follows:

$$\mathrm{MSERI} = \frac{\mathrm{MSE}(\hat{Y}) - \mathrm{MSE}(\tilde{Y})}{\mathrm{MSE}(\tilde{Y})}.$$

It can be shown that

$$\mathrm{MSERI} = [\frac{n}{(1-f)} \times \frac{b^2}{a}] - \frac{1-2f}{1-f}, \tag{1}$$

where $a > 0$, $0 < f < 1$ and $0 < b^2 < 1$. In order for the bias-corrected predictor $\hat{Y}$ to improve on the naive predictor $\hat{Y}$ in terms of MSE, b must satisfy one of the following two conditions:

$$b > \sqrt{(\frac{1}{n} - \frac{2}{N})a} = b_2 \tag{2}$$

or

$$b < -\sqrt{(\frac{1}{n} - \frac{2}{N})a} = b_1. \tag{3}$$

In many situations, the sampling fraction f is negligible in which case MSREI $\approx n\frac{b^2}{a}$.

Define the relative bias (RB) as

$$\mathrm{RB}(\hat{Y}) = \frac{Nb}{\mathrm{E}(Y)} = \frac{Nb}{N(p_{11}+p_{01})} = \frac{p_{10}-p_{01}}{p_{11}+p_{01}}.$$

Since both MSRI and RB depend on N only through f, we arbitrarily fix N and vary f. For a numerical comparison, we fix $N = 100$. Table 2 displays RB($\hat{Y}$), MSE($\hat{Y}$), MSE($\tilde{Y}$) and MSERI for $f = 0.05,\ 0.10$ and two levels of misclassification errors (LE): High (H) and Moderate (M). Table 3 reports the results for low misclassification errors (L). First of all, we notice that the relative bias of $\hat{Y}$ depends on the configurations of false positive (p_{10}) and false negative (p_{01}) probabilities. Clearly, a high relative bias in $\hat{Y}$ is possible. In this case, the bias-corrected estimator $\tilde{Y}$ can have substantially smaller MSE than $\hat{Y}$ even when the sampling fraction f is small. When the relative bias in $\hat{Y}$ is small, one needs much higher sampling fraction for $\tilde{Y}$ to improve on $\hat{Y}$ in terms of mean squared error.

Table 2. RB($\widehat{Y^*}$), MSE($\widehat{Y^*}$), MSE($\hat{Y}$) and MSERI for high and medium misclassification errors

n	f	LE	p_{10}	p_{01}	b_2	b	b_1	RB($\hat{Y}^*$)	MSE($\hat{Y}^*$)	MSE($\hat{Y}$)	MSERI
5	0.05	H	0.80	0.15	0.31	0.65	-0.31	3.71	4277.75	1002.25	3.27
		H	0.75	0.15	0.31	0.60	-0.31	3.00	3654.00	1026.00	2.56
		H	0.60	0.10	0.28	0.50	-0.28	3.33	2545.00	855.00	1.98
		H	0.10	0.60	0.28	-0.50	-0.28	-0.77	2545.00	855.00	1.98
		H	0.20	0.65	0.34	-0.45	-0.34	-0.60	2089.75	1230.25	0.70
		H	0.50	0.35	0.39	0.15	-0.39	0.38	307.75	1572.25	-0.80
		H	0.30	0.60	0.38	-0.30	-0.38	-0.49	981.00	1539.00	-0.36
		H	0.45	0.25	0.34	0.20	-0.34	0.44	466.00	1254.00	-0.63
		M	0.10	0.40	0.27	-0.30	-0.27	-0.43	941.00	779.00	0.21
		M	0.10	0.35	0.26	-0.25	-0.26	-0.45	663.75	736.25	-0.10
		M	0.35	0.20	0.31	0.15	-0.31	0.27	277.75	1002.25	-0.72
		M	0.15	0.35	0.29	-0.20	-0.29	-0.33	446.00	874.00	-0.49
10	0.10	H	0.80	0.15	0.21	0.65	-0.21	3.71	4277.75	474.75	8.01
		H	0.75	0.15	0.21	0.60	-0.21	3.00	3654.00	486.00	6.52
		H	0.60	0.10	0.19	0.50	-0.19	3.33	2545.00	405.00	5.28
		H	0.10	0.60	0.19	-0.50	-0.19	-0.77	2545.00	405.00	5.28
		H	0.20	0.65	0.23	-0.45	-0.23	-0.60	2089.75	582.75	2.59
		H	0.50	0.35	0.26	0.15	-0.26	0.38	307.75	744.75	-0.59
		H	0.30	0.60	0.25	-0.30	-0.25	-0.49	981.00	729.00	0.35
		H	0.45	0.25	0.23	0.20	-0.23	0.44	466.00	594.00	-0.22
		M	0.10	0.40	0.18	-0.30	-0.18	-0.43	941.00	369.00	1.55
		M	0.10	0.35	0.18	-0.25	-0.18	-0.45	663.75	348.75	0.90
		M	0.35	0.20	0.21	0.15	-0.21	0.27	277.75	474.75	-0.41
		M	0.15	0.35	0.19	-0.20	-0.19	-0.33	446.00	414.00	0.08

Table 3. RB$(\hat{Y}^*)$, MSE$(\hat{Y}^*)$, MSE$(\hat{Y})$ and MSERI for low misclassification errors

n	f	p_{10}	p_{01}	b_2	b	b_1	RB$(\hat{Y}^*)$	MSE$(\hat{Y}^*)$	MSE$(\hat{Y})$	MSERI
5	0.05	0.07	0.03	0.13	0.04	-0.13	0.05	25.84	186.96	-0.86
		0.07	0.13	0.19	-0.06	-0.19	-0.11	55.64	373.16	-0.86
10	0.10	0.07	0.03	0.09	0.04	-0.09	0.05	25.84	88.56	-0.71
		0.07	0.13	0.13	-0.06	-0.13	-0.11	55.64	176.76	-0.69
15	0.15	0.07	0.03	0.07	0.04	-0.07	0.05	25.84	55.76	-0.54
		0.07	0.13	0.10	-0.06	-0.10	-0.11	55.64	111.29	-0.50
30	0.30	0.07	0.03	0.04	0.04	-0.04	0.05	25.84	22.96	0.13
		0.07	0.13	0.05	-0.06	-0.05	-0.11	55.64	48.83	0.21

Suppose $p_{10} = p_{01}$, then $b = 0$. Hence, RB$(\hat{Y})$ will always be zero. Also, the MSERI will be a function of the sampling fraction f as shown below

$$\text{MSERI} = -\frac{1-2f}{1-f}. \tag{4}$$

Figure 1 displays the MSERI for different choices of f.

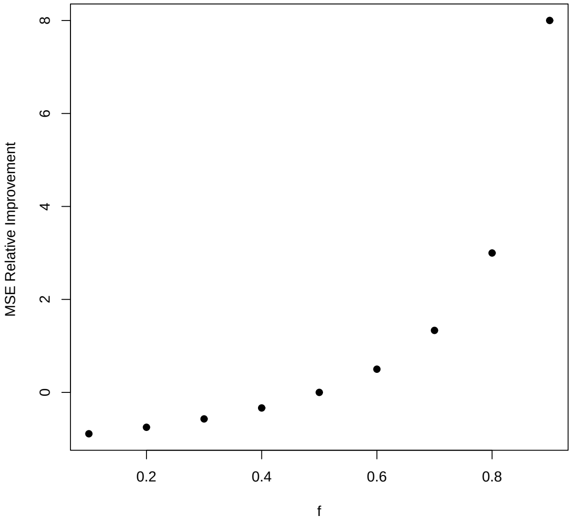


Figure 1. MSE Relative Improvement when $p_{10} = p_{01}$

3. Prediction of $\mathbf{Y}_\delta$ when $S \subset U$

In this subsection, we consider prediction of a finite population total $Y = \sum_{i=1}^{N} \delta_i$ using a sample (s_1) of size n_1 drawn by the *probability proportional to size with* *replacement* (PPSWR) sampling design, where size is defined as:

$$\phi_i = \frac{x_i}{\sum_{i=1}^{N} x_i},$$

where $x_i > 0$ known, $i \in U$. A natural estimator of Y is given by $\hat{Y}^* = \sum_{i \in s_1} \omega_i \delta_i^*$, where δ_i^* is *observed* value of δ_i with misclassification error and $\omega_i = \frac{1}{n_1 \phi_i}, i \in s_1$.

Under the PPSWR sample design and the multinomial classification model of Section 2.3, a heavy algebra yields

$$\text{Bias}(\hat{Y}^*) \quad = \quad Nb$$

and

$$\text{MSE}(\hat{Y}^*) \quad = \quad \frac{1}{n_1}\sum_{i=1}^{N}\frac{1}{\phi_i}p_{1+} - \frac{N}{n_1}p_{1+}\{1+(N-1)p_{1+}\} + N[a+b^2(N-1)],$$

where $p_{1+} = p_{11} + p_{10}$ and $p_{+1} = p_{11} + p_{01}$.

In order to correct the bias of the predictor $\hat{Y}^*$, a second sample, say s_2, of size n_2 is drawn from the first sample s_1 using a SRS design and true δ_i is measured without misclassification error. We define a bias-corrected predictor $\hat{Y}$ as follows:

$$\hat{Y} = \hat{Y}^* - \widehat{Bias},$$

where

$$\widehat{Bias} = \frac{n_1}{n_2}\sum_{i \in s_2} \omega_i(\delta_i^* - \delta_i).$$

It is easy to show that $\hat{Y}$ is an unbiased predictor of Y. Using heavy algebra, the exact MSE of $\hat{Y}$ under different sources of uncertainty is obtained as follows:

$$\begin{aligned}\text{MSE}(\hat{Y}) = \frac{1}{n_1}\{-2Np_{11} + \sum_{i=1}^{N}\frac{1}{\phi_i}[p_{+1} + a\frac{1}{f_2}(1-f_2)] \\ -Np_{1+}\frac{1}{f_2}(1-f_2)[1+p_{1+}(N-1)]\},\end{aligned}$$

where $f_2 = \frac{n_2}{n_1}$.

It can be shown that

$$\begin{aligned}&MSE(\hat{Y^*}) - MSE(\hat{Y}) = \\ &\frac{1}{n_1}\{\sum_{i=1}^{N}\frac{1}{\phi_i}[b - \frac{a(1-f_2)}{f_2}] - Np_{1+}[1+p_{1+}(N-1) - \frac{1}{f_2}(1-f_2)[1+p_{1+}(N-1)]] \\ &\qquad + 2Np_{11}\} + N[a+b^2(N-1)]. \quad (5)\end{aligned}$$

The expression for MSERI is obtained by dividing Eq.(5) by MSE$(\hat{Y})$. Also, we can show that

$$\mathrm{RB}(\hat{Y}^*) \quad = \quad \frac{Nb}{\mathtt{E}(Y)} = \frac{Nb}{N(p_{11}+p_{01})} = \frac{b}{p_{11}+p_{01}}.$$

Tables 4, 5 and 6 display RB$(\hat{Y}^*)$, MSE$(\hat{Y}^*)$, MSE$(\hat{Y})$ and MSERI for high, medium and low misclassification probabilities given by

$$P_H = \begin{pmatrix} 0.10 \\ p_{10} \\ 0.80 - p_{10} \\ 0.10 \end{pmatrix}, P_M = \begin{pmatrix} 0.25 \\ p_{10} \\ 0.50 - p_{10} \\ 0.25 \end{pmatrix}, P_L = \begin{pmatrix} 0.40 \\ p_{10} \\ 0.20 - p_{10} \\ 0.40 \end{pmatrix},$$

respectively.

Different configurations of the high, medium and low misclassification errors are considered by varying the false positive probability p_{10}. For each case, we consider $f_2 = 0.2$, $N = 1000$ when $n_1 = 300$, and $N = 100,000$ when $n_1 = 10,000$.

Table 4. RB$(\hat{Y}^*)$, MSE$(\hat{Y}^*)$, MSE$(\hat{Y})$ and MSERI for high misclassification errors

n_1	p_{10}	Bias$(\hat{Y}^*)$	RB$(\hat{Y}^*)$(%)	MSE$(\hat{Y}^*)$	MSE$(\hat{Y})$	MSERI(%)
300	0.10	-600	-75.00	387092.90	535198.40	-27.67
	0.20	-400	-57.14	200519.40	521137.70	-61.52
	0.30	-200	-33.33	93799.32	506810.60	-81.49
	0.40	0	0.00	66932.65	492217.10	-86.40
	0.50	200	50.00	119919.40	477357.20	-74.88
	0.60	400	133.33	252759.50	462230.90	-45.32
	0.70	600	300.00	465453.00	446838.20	4.17
10000	0.10	-60000	-75.00	3615422483	308209682	1073.04
	0.20	-40000	-57.14	1623101725	300300437	440.50
	0.30	-20000	-33.33	430752967	292311194	47.36
	0.40	0	0.00	38376209	284241951	-86.50
	0.50	20000	50.00	445971451	276092709	61.53
	0.60	40000	133.33	1653538694	267863468	517.31
	0.70	60000	300.00	3661077936	259554228	1310.53

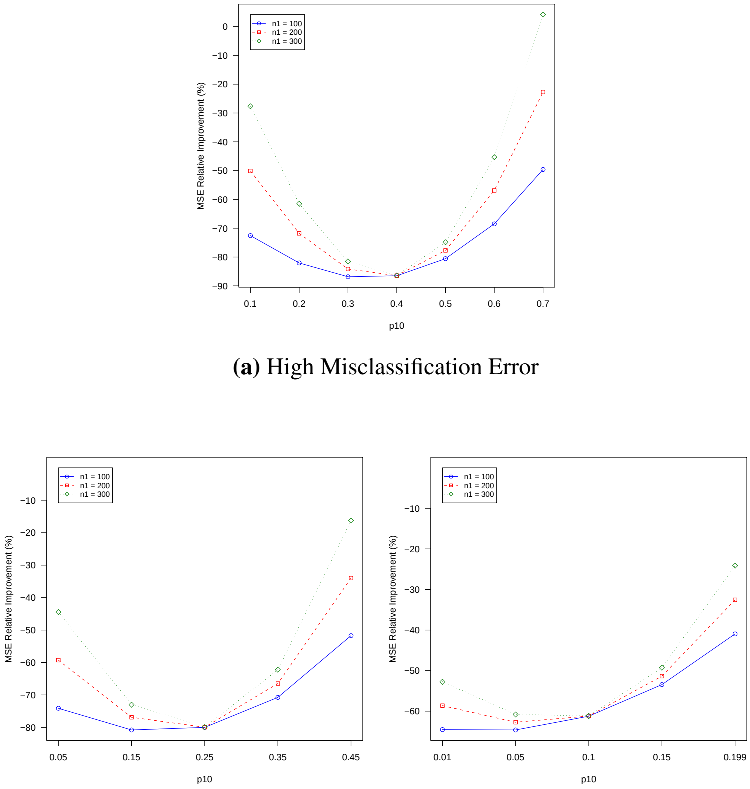
Table 5. RB$(\hat{Y}^*)$, MSE$(\hat{Y}^*)$, MSE$(\hat{Y})$ and MSERI for medium misclassification errors

n_1	p_{10}	Bias$(\hat{Y}^*)$	RB$(\hat{Y}^*)$(%)	MSE$(\hat{Y}^*)$	MSE$(\hat{Y})$	MSERI(%)
300	0.05	-400	-57.14	200219.4	360416.3	-44.45
	0.15	-200	-33.33	93499.32	346089.2	-72.98
	0.25	0	0.00	66632.65	331495.7	-79.90
	0.35	200	50.00	119619.4	316635.8	-62.22
	0.45	400	133.33	252459.5	301509.5	-16.27
10000	0.05	-40000	-57.14	1623071725	207789527	681.11
	0.15	-20000	-33.33	430722967	199800284	115.58
	0.25	0	0.00	38346209	191731041	-80
	0.35	20000	50.00	445941451	183581799	142.92
	0.45	400	133.33	1653508694	175352558	842.96

Table 6. RB$(\hat{Y}^*)$, MSE$(\hat{Y}^*)$, MSE$(\hat{Y})$ and MSERI for low misclassification errors

n_1	p_{10}	Bias$(\hat{Y}^*)$	RB$(\hat{Y}^*)(\%)$	MSE$(\hat{Y}^*)$	MSE$(\hat{Y})$	MSERI$(\%)$
	0.01	-180	-30.51	86919.25	183920.5	-52.74
	0.05	-100	-18.18	69784.31	178104.4	-60.82
300	0.10	0	0.00	66332.65	170774.4	-61.16
	0.15	100	22.22	82844.34	163377.7	-49.29
	0.199	198	49.38	118394.2	156064.4	-24.14
	0.01	-18000	-30.51	355456551	106486050	233.81
	0.05	-10000	-18.18	134508088	103264753	30.26
10000	0.10	0	0.00	38316209	99220131	-61.38
	0.15	10000	22.22	142117330	95155510	49.35
	0.199	19800	49.38	437875637	91152778	380.38

The results for MSE improvement achieved by the bias-corrected estimator over the naive estimator for different situations are plotted in Figures 2 and 3.

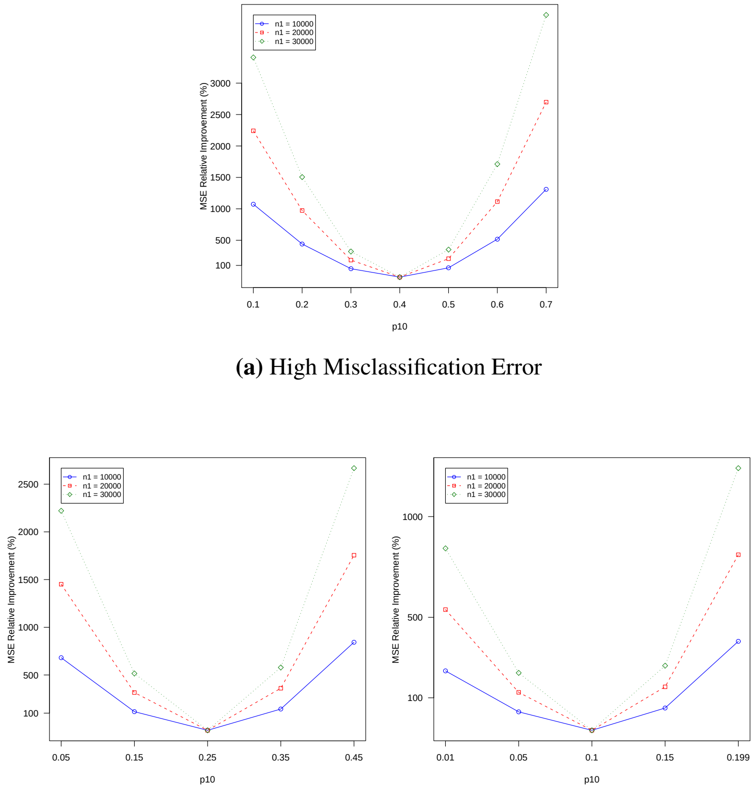


(a) High Misclassification Error

(b) Medium Misclassification Error

(c) Low Misclassification Error

Figure 2. MSE Relative Improvement for $f_2 = 0.2$, $N = 1000$ and n_1 = 100, 200, 300



(a) High Misclassification Error

(b) Medium Misclassification Error **(c)** Low Misclassification Error

Figure 3. MSE Relative Improvement for $f_2 = 0.2$, $N = 10,000$ and n_1 = 10,000, 20,000, 30,000

4. Conclusions

Our research shows that it is possible to correct bias due to misclassification in predictors by drawing a probability sample from the original data, determining the status of misclassification error for the sample and then applying the standard sample survey method. The bias-correction increases variance in the predictor, which impacts the mean squared error. The improvement depends on the distribution of the misclassification error and the sampling fraction in the drawn sample. If additional data that generates the misclassification error are available as in the record linkage literature, it may be possible to improve on the proposed method, we plan to investigate this direction in the future.

5. Acknowledgements

We are also grateful to an anonymous referee for making a number of constructive suggestions, which led to a significant improvement of our paper.

REFERENCES

BEAUCHAMP, A., TONKIN, A. M., KELSALL, H., SUNDARARAJAN, V., ENGLISH, D. R., SUNDARESAN, L., WOLFE, R., TURRELL, G., GILES, G. G., PEETERS, A., (2011). Validation of de-identified record linkage to ascertain hospital admissions in a cohort study. BMC Medical Research Methodology. 11–42.

BENNELL, C., SNOOK, B., MACDONALD, S., HOUSE, J. C., TAYLOR, P. J., (2012). Computerized crime linkage systems: a critical review and research agenda. Criminal Justice and Behavior. 39(5): 620–634.

BOESE, D. H., YOUNG, D. M., STAMEY, J. D., (2006). Confidence intervals for a binomial parameter based on binary data subject to false-positive misclassification. Computational Statistics & Data Analysis. 50: 3369–3385.

BRESLOW, N. E., LUBIN, J. H., LANGHOLZ, B., (1983). Multiplicative models and cohort analysis. Journal of the American Statistical Association. 78: 1–12.

BROSS, I., (1954). Misclassification in 2×2 tables. Biometrics. 10: 478–486.

EVANS, M., GUTTMAN, I., HAITOVSKY, Y., SWARTZ, T., (1996). Bayesian analysis of binary data subject to misclassification. In: Berry, D., Chaloner, K., Geweke, J., eds. Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner. New York: John Wiley, 67–77.

FAIR, M. E., (1989). Studies and references relating to the uses of the Canadian Mortality Data Base. Report from the Occupational and Environmental Health Research Unit, Health Division, Statistics Canada, Ottawa.

FELLIGI, I., SUNTER, A., (1969). A theory for record linkage. Journal of the American Statistical Association. 64: 1183–1210.

GABA, A., WINKLER, R. L., (1992). Implications of errors in survey data: a Bayesian model. Management Science. 38: 913–925.

GIRAUD-CARRIER, C., GOODLIFFE, J., JONES, B. M., CUEVA, S., (2015). Effective record linkage for mining campaign contribution data. Knowledge and Information Systems. 45(2): 389–416.

GOLDBERG, J. D., (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. Journal of the American Statistical Association. 70: 561–567.

GUSTAFSON, P., LE, N. D., SASKIN, R., (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. Biometrics. 57: 598–609.

HOWE, G. R., (1985). Use of computerized record linkage in follow-up studies of cancer epidemiology in Canada. National Cancer Institute Monograph. 67: 117–121.

HOWE, G., R., (1998). Use of computerized record linkage in cohort studies. Epidemiologic Reviews. 20(1): 112–121.

HERZOG, T. N., SCHEUREN, F. J., WINKLER, W. E., (2007). Data Quality and Record Linkage Techniques. Springer, New York, NY.

KABUDULA, C. W., JOUBERT, J. D., TUOANE-NKHASI, M., KAHN, K., RAO, C., GÓMEZ OLIVÉ, F. X., MEE, P., TOLLMAN, S., LOPEZ, A. D., VOS, T., BRADSHAW, D., (2014). Evaluation of record linkage of mortality data between a health and demographic surveillance system and national civil registration system in South Africa. Population Health Metrics. 12–23.

KREWSKI, D., DEWANJI, A., WANG, Y., BARTLETT, S., ZIELINSKI, J. M., MALLICK, R., (2005). The Effect of Record Linkage Errors on Risk Estimates in Cohort Mortality Studies. Survey Methodology. 31: 13–21.

LAHIRI, P., LARSEN, M. D., (2005). Regression analysis with linked data. Journal of the American Statistical Association. 100: 222–230.

LYLES, R. H., LIN, H., M., WILLIAMSON, J. M., (2004). Design and analytic considerations for single-armed studies with misclassification of a repeated binary outcome. Journal of Biopharmaceutical Statistics. 14: 229–247.

NETER, J., MAYNES, E. S., RAMANATHAN, R., (1965). The effect of mismatching on the measurement of response errors. Journal of the American Statistical Association. 60: 1005–1027.

RAHARDJA, D., YANG, Y., (2015). Maximum likelihood estimation of a binomial proportion using one-sample misclassified binary data. Statistica Neerlandica. 69(3), 272–280.

RAHARDJA, D., ZHAO, Y. D., (2013). One-way analysis of proportions for misclassified binomial data. Journal of Statistical Computation and Simulation. 1–10.

SCHEUREN, F., WINKLER, W. E., (1993). Regression Analysis of Data Files That Are Computer Matched. Survey Methodology. 19, 39–58.

STAMEY, J. D., SEAMAN, J. W., YOUNG, D. M., (2007). Bayesian estimation of intervention effect with pre- and post-misclassified binomial data. Journal of Biopharmaceutical Statistics. 17: 93–108.

TENENBEIN, A., (1970). A double sampling scheme for estimating from binomial data with misclassifications. Journal of American Statistical Association. 65(331): 1350–1361.

VIANA, M., RAMAKRISHNAN, V., LEVY, P., (1993). Bayesian analysis of prevalence from results of small screening samples. Communication Statistics Theory and Methods. 22: 575–585.

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. 15: 235–261.

ZHONG, B., (2002). Evaluating qualitative assays using sensitivity and specificity. Journal of Biopharmaceutical Statistics. 12: 409–424.

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 449–466

AN EXTENSION OF THE CLASSICAL DISTANCE CORRELATION COEFFICIENT FOR MULTIVARIATE FUNCTIONAL DATA WITH APPLICATIONS

Tomasz Górecki[1], **Mirosław Krzyśko**[2], **Waldemar Ratajczak**[3], **Waldemar Wołyński**[4]

ABSTRACT

The relationship between two sets of real variables defined for the same individuals can be evaluated by a few different correlation coefficients. For the functional data we have one important tool: canonical correlations. It is not immediately straightforward to extend other similar measures to the context of functional data analysis. In this work we show how to use the distance correlation coefficient for a multivariate functional case.
The approaches discussed are illustrated with an application to some socio-economic data.

Key words: multivariate functional data, functional data analysis, correlation analysis.

1. Introduction

In recent years methods for data representing functions or curves have received much attention. Such data are known in the literature as functional data (Ramsay & Silverman (2005), Horváth & Kokoszka (2012)). Examples of functional data can be found in several application domains, such as medicine, economics, meteorology and many others. In a great number of applications it is necessary to use statistical methods for objects characterized by many features observed in many time points (double multivariate data). In this case such data are called multivariate functional data. The pioneering theoretical paper was Besse (1979), in which random variables have values in a general Hilbert space. Berrendero et al. (2011), Górecki et

[1] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. Department of Statistics, Colorado State University, USA. E-mail: tomasz.gorecki@amu.edu.pl

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

[3] Faculty of Geographical and Geological Sciences, Adam Mickiewicz University, Poland.
E-mail: walrat@amu.edu.pl

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

al. (2014) and Jacques & Preda (2014), present an analysis of multivariate functional data from the point of view of Multivariate Principal Component Analysis (MPCA). Also functional regression models have been extensively studied; see for example James (2002), Müller and Stadmüller (2005), Reiss and Ogden (2007) and Matsui et al. (2008). Various basic classification methods have been adapted to functional data, such as linear discriminant analysis (Hastie et al. (1995)), logistic regression (Rossi 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 theory of combining classifiers has been extended to functional data (Ferraty and Vieu (2009)). Górecki et al. (2015) discussed the problem of classification via regression for multivariate functional data.

In this paper we focus on correlation analysis for multivariate functional data. In the literature, there are different strategies to explore the association between two sets of variables (p dimensional $\boldsymbol{X}$ and q dimensional $\boldsymbol{Y}$). Historically, the first approach was put forward by Hotelling (1936), who proposed the canonical correlation in the framework of Canonical Correlation Analysis (CCA). The CCA is a reference tool concerned with describing linear dependencies between two sets of variables; it seeks a linear combination of the variables of the first group which is maximally correlated with a linear combination of the variables of the second group. The correlation coefficient thus obtained is said to be canonical and the linear combinations are called canonical variables. Leurgans et al. (1993), He et al. (2004), Krzyśko & Waszak (2013) discussed this analysis in the context of functional data.

Another approach is to consider each set of variables through its individual cloud, and to compare the structures (i.e. the shapes) of the two point clouds. In this way, the so-called rV coefficient (Escoufier (1970, 1973), Robert & Escoufier (1976), Escoufier & Robert (1979)) provides an insight into the global association between the two sets of variables.

Székely et al. (2007), Székely & Rizzo (2009, 2012, 2013) defined a measure of dependence between random vectors: the distance correlation (dCor) coefficient. The authors showed that for all random variables with finite first moments, the dCor coefficient generalizes the idea of correlation 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 $p = q = 1$. They constructed their coefficient as a generalization of the simple correlation coefficient without reference to the earlier literature on the rV coefficient. Secondly, the dCor coefficient is equal to zero if and only if there is independence between the random vectors. Indeed, a correlation coefficient measures linear relationships and can be equal to 0 even when the random variables are dependent. This can be seen as a major shortcoming of the correlation coefficient and the rV coefficient.

The rest of this paper is organized as follows. We first review the concept of transformation of discrete data into multivariate functional data (Section 2). Section 3 contains the functional version of canonical correlation coefficients analysis. Section 4 describes our extension of the distance correlation coefficient to the functional case. In Section 5 the accuracy of the proposed methods is demonstrated using some empirical data. Conclusions are given in Section 6.

2. Smoothing of stochastic processes

Let us assume that $\boldsymbol{X} \in L_2^p(I_1)$ and $\boldsymbol{Y} \in L_2^q(I_2)$ are random processes, where $L_2(I)$ is a Hilbert space of square integrable functions on the interval I.

We also assume that $\mathrm{E}(\boldsymbol{X}(s)) = \mathbf{0},\ s \in I_1$ and $\mathrm{E}(\boldsymbol{Y}(t)) = \mathbf{0},\ t \in I_2$.

This fact does not cause loss of generality, because functional correlation coefficients are calculated on the basis of the covariance functions of processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ of the form

$$\mathrm{Cov}\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} = \boldsymbol{\Sigma}(s,t) = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{XX}}(s,t) & \boldsymbol{\Sigma}_{\boldsymbol{XY}}(s,t) \\ \boldsymbol{\Sigma}_{\boldsymbol{YX}}(s,t) & \boldsymbol{\Sigma}_{\boldsymbol{YY}}(s,t) \end{bmatrix},\ s \in I_1,\ t \in I_2,$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{XX}}(s,t) &= \mathrm{E}[\boldsymbol{X}(s)\boldsymbol{X}'(t)],\ s,t \in I_1,\\
\boldsymbol{\Sigma}_{\boldsymbol{XY}}(s,t) &= \mathrm{E}[\boldsymbol{X}(s)\boldsymbol{Y}'(t)],\ s \in I_1,\ t \in I_2,\\
\boldsymbol{\Sigma}_{\boldsymbol{YX}}(s,t) &= \mathrm{E}[\boldsymbol{Y}(s)\boldsymbol{X}'(t)],\ s \in I_2,\ t \in I_1,\\
\boldsymbol{\Sigma}_{\boldsymbol{YY}}(s,t) &= \mathrm{E}[\boldsymbol{Y}(s)\boldsymbol{Y}'(t)],\ s,t \in I_2.
\end{aligned}$$

We will further assume that each component X_g of process $\boldsymbol{X}$ and Y_h of process $\boldsymbol{Y}$ can be represented by a finite number of orthonormal basis functions $\{\varphi_e\}$ and $\{\varphi_f\}$ of space $L_2(I_1)$ and $L_2(I_2)$, respectively:

$$\begin{aligned}
X_g(s) &= \sum_{e=0}^{E_g} \alpha_{ge}\varphi_e(s), s \in I_1, g = 1,2,...,p,\\
Y_h(t) &= \sum_{f=0}^{F_h} \beta_{hf}\varphi_f(t), t \in I_2, h = 1,2,...,q.
\end{aligned}$$

The choice of the basis seems not crucial. We can use various orthonormal basis, but Fourier basis seems the most appropriate in most cases (Górecki & Krzyśko

(2012)) and the most common in practice. The degree of smoothness of functions X_g and Y_h depends on values E_g and F_h respectively (small values cause more smoothing). The choice of the truncation parameters is critical for the proper representation of general stochastic process. The optimal number of basis elements could be determined using the Bayesian Information Criterion (BIC) for each function separately through finding the most frequent value (modal value) over all functions. We should prefer this value to be large, particularly when the stochastic processes are observed at high frequency with little noise.

We introduce the following notation:

$$\boldsymbol{\alpha} = (\alpha_{10},...,\alpha_{1E_1},...,\alpha_{p0},...,\alpha_{pE_p})',$$
$$\boldsymbol{\beta} = (\beta_{10},...,\beta_{1F_1},...,\beta_{q0},...,\beta_{qF_q})',$$

$$\boldsymbol{\varphi}_{E_g}(s) = (\varphi_0(s),\dots,\varphi_{E_g}(s))', s \in I_1, g = 1,2,...,p,$$
$$\boldsymbol{\varphi}_{F_h}(t) = (\varphi_0(t),\dots,\varphi_{E_g}(t))', t \in I_2, h = 1,2,...,q,$$

$$\boldsymbol{\Phi}_1(s) = \begin{bmatrix} \boldsymbol{\varphi}'_{E_1}(s) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}'_{E_2}(s) & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\varphi}'_{\mathrm{E}_p}(s) \end{bmatrix},$$
$$\boldsymbol{\Phi}_2(t) = \begin{bmatrix} \boldsymbol{\varphi}'_{F_1}(t) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}'_{F_2}(t) & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\varphi}'_{F_q}(t) \end{bmatrix}.$$

Using the above matrix notation, processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be represented as:

$$\boldsymbol{X}(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{\alpha}, \quad \boldsymbol{Y}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\beta}.$$

This means that the realizations of processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ are in finite dimensional subspaces of $L_2^p(I_1)$ and $L_2^q(I_2)$ respectively. We will denote these subspaces by $\mathscr{L}_2^p(I_1)$ and $\mathscr{L}_2^q(I_2)$.

For random vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ we have:

$$\mathrm{E}(\boldsymbol{\alpha}) = \mathbf{0},\ \mathrm{E}(\boldsymbol{\beta}) = \mathbf{0}$$

and

$$\mathrm{Cov}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}=\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\alpha\alpha}} & \boldsymbol{\Sigma}_{\boldsymbol{\alpha\beta}}\\ \boldsymbol{\Sigma}_{\boldsymbol{\beta\alpha}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta\beta}}\end{bmatrix},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\alpha\alpha}}=\mathrm{E}(\boldsymbol{\alpha\alpha}')$, $\boldsymbol{\Sigma}_{\boldsymbol{\alpha\beta}}=\mathrm{E}(\boldsymbol{\alpha\beta}')$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta\alpha}}=\mathrm{E}(\boldsymbol{\beta\alpha}')$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta\beta}}=\mathrm{E}(\boldsymbol{\beta\beta}')$.

Note that

$$\boldsymbol{\Sigma}_{\boldsymbol{XX}}(s,t)=\mathrm{E}[\boldsymbol{\Phi}_1(s)\boldsymbol{\alpha\alpha}'\boldsymbol{\Phi}_1'(t)]=\boldsymbol{\Phi}_1(s)\,\mathrm{E}(\boldsymbol{\alpha\alpha}')\boldsymbol{\Phi}_1'(t)=\boldsymbol{\Phi}_1(s)\boldsymbol{\Sigma}_{\boldsymbol{\alpha\alpha}}\boldsymbol{\Phi}_1'(t).$$

Similarly

$$\boldsymbol{\Sigma}_{\boldsymbol{XY}}(s,t)=\boldsymbol{\Phi}_1(s)\boldsymbol{\Sigma}_{\boldsymbol{\alpha\beta}}\boldsymbol{\Phi}_2'(t),$$

$$\boldsymbol{\Sigma}_{\boldsymbol{YX}}(s,t)=\boldsymbol{\Phi}_2(s)\boldsymbol{\Sigma}_{\boldsymbol{\beta\alpha}}\boldsymbol{\Phi}_1'(t),$$

$$\boldsymbol{\Sigma}_{\boldsymbol{YY}}(s,t)=\boldsymbol{\Phi}_2(s)\boldsymbol{\Sigma}_{\boldsymbol{\beta\beta}}\boldsymbol{\Phi}_2'(t).$$

In fact, the correlation analysis for random processes is based on matrices $\boldsymbol{\Sigma}_{\boldsymbol{\alpha\alpha}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha\beta}}$ which are unknown. We have to estimate them on the basis of n independent realizations $\boldsymbol{X}_1,\boldsymbol{X}_2,....,\boldsymbol{X}_n$ and $\boldsymbol{Y}_1,\boldsymbol{Y}_2,....,\boldsymbol{Y}_n$ of random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$. We have $\boldsymbol{X}_i(s)=\boldsymbol{\Phi}_1(s)\boldsymbol{\alpha}_i$ and $\boldsymbol{Y}_i(t)=\boldsymbol{\Phi}_2(t)\boldsymbol{\beta}_i$, $i=1,2,\ldots,n$. This problem has been extensively studied in the literature, e.g. Beutler (1970), Lee (1976) and Masry (1978).

Typically, data are recorded at discrete moments in time. The transformation of discrete data into functional data is performed for each realization and each variable separately. Let x_{gj} denote an observed value of feature X_g, $g=1,2,\ldots p$ at the jth time point s_j, where $j=1,2,...,J$. Similarly, let y_{hj} denote an observed value of feature Y_h, $h=1,2,\ldots q$ at the jth time point t_j, where $j=1,2,...,J$. Then, our data consist of pJ pairs of (s_j,x_{gj}) and of qJ pairs of (t_j,y_{hj}).

The coefficients $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ are estimated by the least squares method. Let us denote these estimates by $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, $i=1,2,\ldots,n$.

As a result of the transformation process, we obtain functional data of the form:

$$\boldsymbol{x}_i(s)=\boldsymbol{\Phi}_1(s)\boldsymbol{a}_i,\ \boldsymbol{y}_i(t)=\boldsymbol{\Phi}_2(t)\boldsymbol{b}_i, \tag{1}$$

where $s\in I_1$, $t\in I_2$ and $i=1,2,\ldots,n$.

Let $\boldsymbol{A}=(\boldsymbol{a}_1',\boldsymbol{a}_2',\ldots,\boldsymbol{a}_n')'$, and $\boldsymbol{B}=(\boldsymbol{b}_1',\boldsymbol{b}_2',\ldots,\boldsymbol{b}_n')'$. Then

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha\alpha}}=\frac{1}{n}\boldsymbol{A}'\boldsymbol{A},\quad \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta\beta}}=\frac{1}{n}\boldsymbol{B}'\boldsymbol{B},\quad \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha\beta}}=\frac{1}{n}\boldsymbol{A}'\boldsymbol{B}. \tag{2}$$

3. Functional canonical correlation coefficient

Functional canonical variables U and V for stochastic processes $\boldsymbol{X} \in \mathscr{L}_2^p(I_1)$ and $\boldsymbol{Y} \in \mathscr{L}_2^q(I_2)$ are defined as follows:

$$U(s) = \boldsymbol{l}'(s)\boldsymbol{X}(s),\ \boldsymbol{l} \in \mathscr{L}_2^p(I_1),$$
$$V(t) = \boldsymbol{m}'(t)\boldsymbol{Y}(t),\ \boldsymbol{m} \in \mathscr{L}_2^q(I_2),$$

where $\boldsymbol{l}$ and $\boldsymbol{m}$ are weight functions.

We have

$$\mathrm{E}(U(s)) = \mathrm{E}(V(t)) = 0,\ s \in I_1,\ t \in I_2.$$

Let us denote the covariance matrix of processes U and V by

$$\boldsymbol{\Sigma}_{UV}(s,t) = \begin{bmatrix} \sigma_{UU}(s,t) & \sigma_{UV}(s,t) \\ \sigma_{VU}(s,t) & \sigma_{VV}(s,t) \end{bmatrix}.$$

Because $\boldsymbol{l} \in \mathscr{L}_2^p(I_1)$ and $\boldsymbol{m} \in \mathscr{L}_2^q(I_2)$ we have

$$\boldsymbol{l}(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{\lambda},\ \boldsymbol{m}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\mu},$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{K_1+p}$, $\boldsymbol{\mu} \in \mathbb{R}^{K_2+q}$ and $K_1 = E_1 + \ldots + E_p$, $K_2 = F_1 + \ldots + F_q$.

Moreover

$$\begin{aligned}\sigma_{UU}(s,t) &= \mathrm{E}[U(s)U'(t)] = \mathrm{E}[\boldsymbol{l}'(s)\boldsymbol{X}(s)\boldsymbol{X}'(t)\boldsymbol{l}(t)] \\ &= \mathrm{E}[\boldsymbol{\lambda}'\boldsymbol{\Phi}_1'(s)\boldsymbol{\Phi}_1(s)\boldsymbol{\alpha}\boldsymbol{\alpha}'\boldsymbol{\Phi}_1'(t)\boldsymbol{\Phi}_1(t)\boldsymbol{\lambda}] \\ &= \boldsymbol{\lambda}'\boldsymbol{\Phi}_1'(s)\boldsymbol{\Phi}_1(s)\boldsymbol{\Sigma}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\boldsymbol{\Phi}_1'(t)\boldsymbol{\Phi}_1(t)\boldsymbol{\lambda}.\end{aligned}$$

Similarly

$$\begin{aligned}\sigma_{UV}(s,t) &= \boldsymbol{\lambda}'\boldsymbol{\Phi}_1'(s)\boldsymbol{\Phi}_1(s)\boldsymbol{\Sigma}_{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{\Phi}_2'(t)\boldsymbol{\Phi}_2(t)\boldsymbol{\mu}, \\ \sigma_{VU}(s,t) &= \boldsymbol{\mu}'\boldsymbol{\Phi}_2'(s)\boldsymbol{\Phi}_2(s)\boldsymbol{\Sigma}_{\boldsymbol{\beta}\boldsymbol{\alpha}}\boldsymbol{\Phi}_1'(t)\boldsymbol{\Phi}_1(t)\boldsymbol{\lambda}, \\ \sigma_{VV}(s,t) &= \boldsymbol{\mu}'\boldsymbol{\Phi}_2'(s)\boldsymbol{\Phi}_2(s)\boldsymbol{\Sigma}_{\boldsymbol{\beta}\boldsymbol{\beta}}\boldsymbol{\Phi}_2'(t)\boldsymbol{\Phi}_2(t)\boldsymbol{\mu}.\end{aligned}$$

The functional canonical coefficient $\rho_{\boldsymbol{X},\boldsymbol{Y}}$ is defined as

$$\rho_{\boldsymbol{X},\boldsymbol{Y}} = \max_{\boldsymbol{l},\boldsymbol{m}} \int_{I_1}\int_{I_2} \sigma_{UV}(s,t)dsdt = \max_{\boldsymbol{l},\boldsymbol{m}} \int_{I_1}\int_{I_2} \mathrm{E}[\boldsymbol{l}'(s)\boldsymbol{X}(s)\boldsymbol{Y}'(t)\boldsymbol{m}(t)]dsdt,$$

subject to the constraint

$$\int_{I_1}\int_{I_2}\sigma_{UU}(s,t)dsdt=\int_{I_1}\int_{I_2}\sigma_{VV}(s,t)dsdt=1.$$

Because

$$\int_{I_1}\int_{I_2}\sigma_{UV}(s,t)dsdt=\boldsymbol{\lambda}'\boldsymbol{\Sigma}_{\boldsymbol{\alpha\beta}}\boldsymbol{\mu},$$
$$\int_{I_1}\int_{I_2}\sigma_{UU}(s,t)dsdt=\boldsymbol{\lambda}'\boldsymbol{\Sigma}_{\boldsymbol{\alpha\alpha}}\boldsymbol{\lambda},$$
$$\int_{I_1}\int_{I_2}\sigma_{VV}(s,t)dsdt=\boldsymbol{\mu}'\boldsymbol{\Sigma}_{\boldsymbol{\beta\beta}}\boldsymbol{\mu}$$

we have

$$\rho_{\boldsymbol{X},\boldsymbol{Y}}=\max_{\boldsymbol{\lambda},\boldsymbol{\mu}}\boldsymbol{\lambda}'\boldsymbol{\Sigma}_{\boldsymbol{\alpha\beta}}\boldsymbol{\mu}=\rho_{\boldsymbol{\alpha},\boldsymbol{\beta}},$$

subject to the restriction

$$\boldsymbol{\lambda}'\boldsymbol{\Sigma}_{\boldsymbol{\alpha\alpha}}\boldsymbol{\lambda}=\boldsymbol{\mu}'\boldsymbol{\Sigma}_{\boldsymbol{\beta\beta}}\boldsymbol{\mu}=1.$$

From the above we see that the functional canonical correlation coefficient $\rho_{\boldsymbol{X},\boldsymbol{Y}}$ of the pair of random processes $\boldsymbol{X}\in\mathscr{L}_2^p(I_1)$ and $\boldsymbol{Y}\in\mathscr{L}_2^q(I_2)$ is equivalent to the canonical correlation coefficient $\rho_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ of the pair of the random vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Note that if $\boldsymbol{X}\in L_2^p(I_1)$ and $\boldsymbol{Y}\in L_2^q(I_2)$ there exist weight functions such that the functional canonical coefficient is equal to one. This means that, with an increasing size of a number of basis functions, the functional canonical coefficient will tend to one. To avoid this problem Leurgans et al. (1993) proposed some additional regularization. However, as for many correlation coefficients, it is difficult to evaluate the magnitude of the relationship just by considering its values.

The canonical correlation coefficient $\rho_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ of the pair of random vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is based on matrices $\boldsymbol{\Sigma}_{\boldsymbol{\alpha\alpha}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha\beta}}$. If they are not known, we have to use their estimators (2).

Hence

$$\hat{\rho}_{\boldsymbol{\alpha},\boldsymbol{\beta}}=\max_{\boldsymbol{\lambda},\boldsymbol{\mu}}\boldsymbol{\lambda}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha\beta}}\boldsymbol{\mu},$$

under the condition

$$\boldsymbol{\lambda}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha\alpha}}\boldsymbol{\lambda}=\boldsymbol{\mu}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta\beta}}\boldsymbol{\mu}=1.$$

4. Functional distance correlation

First, let us define the joint characteristic function of the pair of random processes $(\boldsymbol{X},\boldsymbol{Y})$. If for all functions $\boldsymbol{l}\in L_2^p(I_1)$ the integral $\int_{I_1}\boldsymbol{l}'(s)\boldsymbol{X}(s)ds$ converges for almost all realization of $\boldsymbol{X}$, and for all functions $\boldsymbol{m} \in L_2^q(I_1)$ the integral $\int_{I_2} \boldsymbol{m}'(t)\boldsymbol{Y}(t)dt$ converges for almost all realizations of $\boldsymbol{Y}$, then the characteristic function of the pair of random processes $(\boldsymbol{X}, \boldsymbol{Y})$ has the following form:

$$\varphi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{m}) = \mathrm{E}\{\exp[i < \boldsymbol{l},\boldsymbol{X} >_p + i < \boldsymbol{m},\boldsymbol{Y} >_q]\},$$

where

$$< \boldsymbol{l},\boldsymbol{X} >_p = \int_{I_1} \boldsymbol{l}'(s)\boldsymbol{X}(s)ds, \quad < \boldsymbol{m},\boldsymbol{Y} >_q = \int_{I_2} \boldsymbol{m}'(t)\boldsymbol{Y}(t)dt$$

and $i^2 = -1$. Moreover, we define the marginal characteristic function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows: $\varphi_{\boldsymbol{X}}(\boldsymbol{l}) = \varphi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{0})$ and $\varphi_{\boldsymbol{Y}}(\boldsymbol{m}) = \varphi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{0},\boldsymbol{m})$.

Now, let us assume that $\boldsymbol{X} \in \mathscr{L}_2^p(I_1)$ and $\boldsymbol{Y} \in \mathscr{L}_2^q(I_2)$. Then, the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be represented as:

$$\boldsymbol{X}(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{\alpha}, \quad \boldsymbol{Y}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\beta},$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{K_1+p}$ and $\boldsymbol{\beta} \in \mathbb{R}^{K_2+q}$.

In this case, we may assume (Ramsey & Silverman (2005)) that the vector function $\boldsymbol{l}$ and the process $\boldsymbol{X}$ are in the same space, i.e. function $\boldsymbol{l}$ can be written in the form

$$\boldsymbol{l}(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{K_1+p}$.

We may assume the same for the vector function $\boldsymbol{m}$ and the process $\boldsymbol{Y}$. Then, we have

$$\boldsymbol{m}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\mu},$$

where $\boldsymbol{\mu} \in \mathbb{R}^{K_2+q}$.

Hence

$$< \boldsymbol{l},\boldsymbol{X} >_p = \int_{I_1} \boldsymbol{l}'(s)\boldsymbol{X}(s)ds = \boldsymbol{\lambda}'[\int_{I_1} \boldsymbol{\Phi}_1'(s)\boldsymbol{\Phi}_1(s)ds]\boldsymbol{\alpha} = \boldsymbol{\lambda}'\boldsymbol{\alpha}$$

and

$$< \boldsymbol{m},\boldsymbol{Y} >_q = \int_{I_2} \boldsymbol{m}'(t)\boldsymbol{Y}(t)dt = \boldsymbol{\mu}'[\int_{I_2} \boldsymbol{\Phi}_2'(t)\boldsymbol{\Phi}_2(t)dt]\boldsymbol{\beta} = \boldsymbol{\mu}'\boldsymbol{\beta}$$

then

$$\varphi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{m}) = \mathrm{E}\{\exp[i\boldsymbol{\lambda}'\boldsymbol{\alpha} + i\boldsymbol{\mu}'\boldsymbol{\beta}]\} = \varphi_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{\lambda},\boldsymbol{\mu}),$$

where $\varphi_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{\lambda},\boldsymbol{\mu})$ is the joint characteristic function of the pair of random vectors

$(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

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 processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ as a nonnegative number $\nu_{\boldsymbol{X},\boldsymbol{Y}}$ defined by

$$\nu_{\boldsymbol{X},\boldsymbol{Y}} = \nu_{\boldsymbol{\alpha},\boldsymbol{\beta}},$$

where

$$\nu^2_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \frac{1}{C_{K_1+p}C_{K_2+q}} \int_{\mathbb{R}^{K_1+K_2+p+q}} \frac{|\phi_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{\lambda},\boldsymbol{\mu}) - \phi_{\boldsymbol{\alpha}}(\boldsymbol{\lambda})\phi_{\boldsymbol{\beta}}(\boldsymbol{\mu})|^2}{\|\boldsymbol{\lambda}\|^{K_1+p+1}_{K_1+p} \|\boldsymbol{\mu}\|^{K_2+q+1}_{K_2+q}} d\boldsymbol{\lambda} d\boldsymbol{\mu},$$

and $|z|$ denotes the modulus of $z \in \mathbb{C}$, $\|\boldsymbol{\lambda}\|_{K_1+p}$, $\|\boldsymbol{\mu}\|_{K_2+q}$ the standard Euclidean norms on the corresponding spaces, and

$$C_r = \frac{\pi^{\frac{1}{2}(r+1)}}{\Gamma(\frac{1}{2}(r+1))}.$$

The functional distance correlation between random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ is a nonnegative number defined by

$$\mathscr{R}_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\nu_{\boldsymbol{X},\boldsymbol{Y}}}{\sqrt{\nu_{\boldsymbol{X},\boldsymbol{X}} \nu_{\boldsymbol{Y},\boldsymbol{Y}}}}$$

if both $\nu_{\boldsymbol{X},\boldsymbol{X}}$ and $\nu_{\boldsymbol{Y},\boldsymbol{Y}}$ are strictly positive, and zero otherwise. For distributions with finite first moments, distance correlation characterizes independence in that $0 \leq \mathscr{R}_{\boldsymbol{X},\boldsymbol{Y}} \leq 1$ with $\mathscr{R}_{\boldsymbol{X},\boldsymbol{Y}} = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.

We can estimate functional distance covariance using functional data of the form (1).

On the basis of the result of Székely et al. (2007), we have

$$\hat{\nu}^2_{\boldsymbol{X},\boldsymbol{Y}} = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl},$$

where $a_{kl} = \|\boldsymbol{a}_k - \boldsymbol{a}_l\|_{K_1+p}$, $\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}$, $\bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}$, $\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}$ and $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$, and similarly for $b_{kl} = \|\boldsymbol{b}_k - \boldsymbol{b}_l\|_{K_2+q}$, $\bar{b}_{k\cdot}$, $\bar{b}_{\cdot l}$, $\bar{b}_{\cdot\cdot}$, and B_{kl}, where $\boldsymbol{a}_k$, $\boldsymbol{a}_l$, $\boldsymbol{b}_k$, $\boldsymbol{b}_l$ are given by (1) and $k,l = 1,\ldots,n$. Thus, the squared sample distance covariance equals an average entry in the component-wise or Schur product of the centered distance matrices for the two vectors.

The sample functional distance correlation is then defined by

$$\hat{\mathscr{R}}_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\hat{\nu}_{\boldsymbol{X},\boldsymbol{Y}}}{\sqrt{\hat{\nu}_{\boldsymbol{X},\boldsymbol{X}}\hat{\nu}_{\boldsymbol{Y},\boldsymbol{Y}}}}$$

if both $\hat{\nu}_{\boldsymbol{X},\boldsymbol{X}}$ and $\hat{\nu}_{\boldsymbol{Y},\boldsymbol{Y}}$ are strictly positive, and zero otherwise.

The problem of testing the independence between the random processes $\boldsymbol{X} \in \mathscr{L}_2^p(I_1)$ and $\boldsymbol{Y} \in \mathscr{L}_2^q(I_2)$ is equivalent to the problem of testing H_0: $\mathscr{R}_{\boldsymbol{X},\boldsymbol{Y}} = 0$. Székely et al. (2007) showed that under the null hypothesis of independence, $n\hat{\mathscr{R}}_{\boldsymbol{X},\boldsymbol{Y}}$ converges to $\sum_{j=1}^{\infty} \nu_j Z_j^2$, where Z_j are i.i.d $N(0,1)$, and ν_j depends on the distribution of $(\boldsymbol{X},\boldsymbol{Y})$. In practice, permutation tests are used to assess the significance of the functional distance correlation (Josse & Holmes (2014)).

5. Empirical application

In this Section we offer an illustrative example of applying correlation analysis to functional data. This method was employed here to cluster the twenty groups (pillars) of variables of 127 countries of the world in the period 2008-2014. The list of countries used in correlation analysis is contained in Table 1. Table 2 describes the variables used in the analysis divided into pillars. For this purpose, 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. The data were transformed into functional data by the method described in Section 2. Calculations were performed using the Fourier basis. The time interval $[0,T] = [0,6]$ was divided into moments of time in the following way: $t_1 = 0.5(2008/2009), t_2 = 1.5(2009/2010), \ldots, t_6 = 5.5(2013/2014)$. Moreover, in view of the small number of time periods ($J = 6$), for each variable the maximum number of basis components was taken to be equal to five. Table 3 contains the values of functional canonical correlation coefficients. As expected, they are all close to one. But a high value of this coefficient does not necessarily mean that there is a significant relationship between the two groups of variables. Table 4 contains the values of functional distance correlation coefficients. This time the values are rather moderate and easier to interpret. It is readily visible that the coefficients assume the highest values for the following pairs of pillars: 2 - infrastructure and 10 - marker size; 11 - business sophistication and 12 - innovation; 5 - higher education and training and 12 - innovation; 5 - higher education and training and 11 - business sophistication; as well as 6 - goods market efficiency and 11 - business sophistication. In turn, the coefficients have the lowest values for the pillars: 4 - health and primary education and

10 - market size, as well as 4 - health and primary education and 2 - infrastructure. Both the highest and the lowest values of distance correlation coefficients have an obvious empirical foundation.

We performed permutation tests for the correlation coefficients discussed. For all tests p-values were close to zero, so we can infer that we have some significant relationship between the groups (pillars) of variables.

Finally, we joined these pillars using Ward's hierarchical clustering algorithm with a distance measure of the form $1-\hat{\rho}_{\boldsymbol{X},\boldsymbol{Y}}$ and $1-\hat{\mathscr{R}}_{\boldsymbol{X},\boldsymbol{Y}}$ respectively. The results are shown in Figures 1 and 2. As can be observed, given the wide differences in the $\hat{\mathscr{R}}_{\boldsymbol{X},\boldsymbol{Y}}$ values, functional distance correlations permit arranging the various groups of variables into pillars in a logical way, e.g. $(4,9)$, $(11,12)$, etc. This allows analysing the examined reality in a deeper way, which is not possible when using canonical correlation coefficients.

During the numerical calculation process we used R software (R Core Team (2015)) and packages: `CCP` (Menzel (2012)), `energy` (Rizzo & Székely (2014)) and `fda` (Ramsay et al. (2014)).

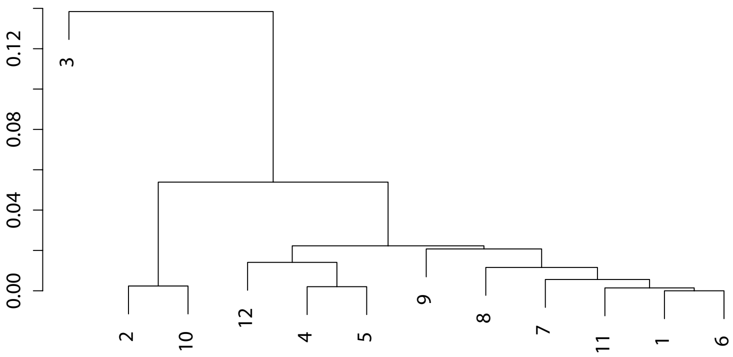


Figure 1: Dendrogram based on the functional canonical correlation coefficients.

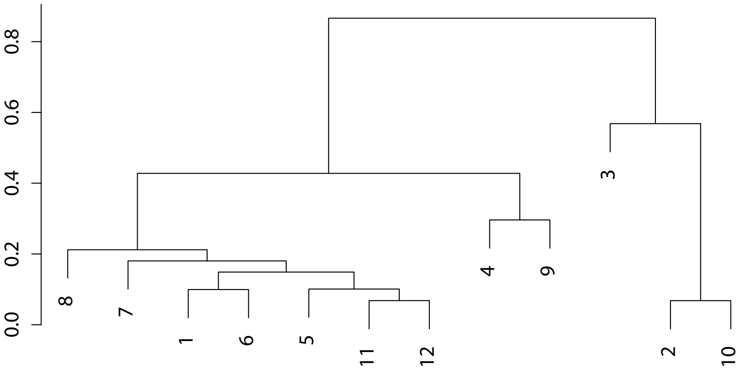


Figure 2: Dendrogram based on the functional distance correlation coefficients.

Table 1. Countries used in correlation analysis, 2008-2014

1	Albania	44	Germany	87	Nicaragua
2	Algeria	45	Ghana	88	Nigeria
3	Argentina	46	Greece	89	Norway
4	Armenia	47	Guatemala	90	Oman
5	Australia	48	Guyana	91	Pakistan
6	Austria	49	Honduras	92	Panama
7	Azerbaijan	50	Hong Kong SAR	93	Paraguay
8	Bahrain	51	Hungary	94	Peru
9	Bangladesh	52	Iceland	95	Philippines
10	Barbados	53	India	96	Poland
11	Belgium	54	Indonesia	97	Portugal
12	Benin	55	Ireland	98	Puerto Rico
13	Bolivia	56	Israel	99	Qatar
14	Bosnia and Herzegovina	57	Italy	100	Romania
15	Botswana	58	Jamaica	101	Russian Federation
16	Brazil	59	Japan	102	Saudi Arabia
17	Brunei Darussalam	60	Jordan	103	Senegal
18	Bulgaria	61	Kazakhstan	104	Serbia
19	Burkina Faso	62	Kenya	105	Singapore
20	Burundi	63	Korea Rep	106	Slovak Republic
21	Cambodia	64	Kuwait	107	Slovenia
22	Cameroon	65	Kyrgyz Republic	108	South Africa
23	Canada	66	Latvia	109	Spain
24	Chad	67	Lesotho	110	Sri Lanka
25	Chile	68	Lithuania	111	Sweden
26	China	69	Luxembourg	112	Switzerland
27	Colombia	70	Macedonia FYR	113	Taiwan China
28	Costa Rica	71	Madagascar	114	Tanzania
29	Côte d'Ivoire	72	Malawi	115	Thailand
30	Croatia	73	Malaysia	116	Timor-Leste
31	Cyprus	74	Mali	117	Trinidad and Tobago
32	Czech Republic	75	Malta	118	Turkey
33	Denmark	76	Mauritania	119	Uganda
34	Dominican Republic	77	Mauritius	120	Ukraine
35	Ecuador	78	Mexico	121	United Arab Emirates
36	Egypt	79	Mongolia	122	United Kingdom
37	El Salvador	80	Montenegro	123	United States
38	Estonia	81	Morocco	124	Uruguay
39	Ethiopia	82	Mozambique	125	Venezuela
40	Finland	83	Namibia	126	Vietnam
41	France	84	Nepal	127	Zambia
42	Gambia The	85	Netherlands		
43	Georgia	86	New Zealand		

Table 2. Variables used in correlation analysis, 2008-2014

No.	Variables	Pillars
1	Property rights	1. Institutions
2	Intellectual property protection	
3	Diversion of public funds	
4	Public trust of politicians	
5	Judicial independence	
6	Favoritism in decisions of government officials	
7	Wastefulness of government spending	
8	Burden of government regulation	
9	Transparency of government policymaking	
10	Business costs of terrorism	
12	Business costs of crime and violence	
11	Organized crime	
12	Reliability of police services	
13	Ethical behavior of firms	
14	Strength of auditing and reporting standards	
15	Efficacy of corporate boards	
16	Protection of minority shareholders' interests	
17	Quality of overall infrastructure	2. Infrastructure
18	Quality of roads	
19	Quality of port infrastructure	
20	Quality of air transport infrastructure	
21	Available airline seat kilometers	
22	Quality of electricity supply	
23	Inflation	3. Macroeconomic environment
24	Government debt	
25	Business impact of tuberculosis	4. Health and primary education
26	Tuberculosis incidence	
27	Business impact of HIV/AIDS	
28	HIV prevalence	
29	Infant mortality	
30	Life expectancy	
31	Quality of primary education	
32	Quality of the educational system	5. Higher education and training
33	Quality of math and science education	
34	Quality of management schools	
35	Internet access in schools	
36	Local availability of specialized research and training services	
37	Extent of staff training	
38	Intensity of local competition	6. Goods market efficiency
39	Extent of market dominance	
40	Effectiveness of anti-monopoly policy	
41	Agricultural policy costs	
42	Prevalence of trade barriers	
43	Prevalence of foreign ownership	
44	Business impact of rules on FDI	
45	Burden of customs procedures	
46	Degree of customer orientation	
47	Buyer sophistication	

Table 2. Variables used in correlation analysis, 2008-2014 (continuation)

No.	Variables	Pillars
48	Cooperation in labor-employer relations	7. Labor market efficiency
49	Flexibility of wage determination	
50	Hiring and firing practices	
51	Pay and productivity	
52	Reliance on professional management	
53	Female participation in labor force	
54	Financing through local equity market	8. Financial market development
55	Ease of access to loans	
56	Venture capital availability	
57	Soundness of banks	
58	Regulation of securities exchanges	
59	Availability of latest technologies	9. Technological readiness
60	Firm-level technology absorption	
61	FDI and technology transfer	
62	Internet users	
63	Domestic market size index	10. Market size
64	Foreign market size index	
65	GDP valued at PPP	
66	Exports as a percentage of GDP	
67	Local supplier quantity	11. Business sophistication
68	Local supplier quality	
69	State of cluster development	
70	Nature of competitive advantage	
71	Value chain breadth	
72	Control of international distribution	
73	Production process sophistication	
74	Extent of marketing	
75	Willingness to delegate authority	
76	Capacity for innovation	12. Innovation
77	Quality of scientific research institutions	
78	Company spending on R&D	
79	Government procurement of advanced technology products	
80	Availability of scientists and engineers	

Table 3. Functional canonical correlation coefficients

	1	2	3	4	5	6	7	8	9	10	11
2	0.9997										
3	0.9711	0.9137									
4	0.9999	0.9903	0.9036								
5	0.9993	0.9928	0.9186	0.9980							
6	1.0000	0.9927	0.9440	0.9921	0.9941						
7	0.9995	0.9795	0.8687	0.9822	0.9913	0.9947					
8	0.9988	0.9701	0.8773	0.9778	0.9781	0.9917	0.9878				
9	0.9988	0.9872	0.8714	0.9744	0.9927	0.9922	0.9846	0.9683			
10	0.9949	0.9976	0.8558	0.9518	0.9699	0.9828	0.9561	0.9408	0.9391		
11	1.0000	0.9934	0.9274	0.9924	0.9973	0.9982	0.9941	0.9885	0.9915	0.9816	
12	0.9984	0.9795	0.8782	0.9849	0.9915	0.9928	0.9794	0.9763	0.9795	0.9540	0.9937

Table 4. Functional distance correlation coefficients

	1	2	3	4	5	6	7	8	9	10	11
2	0.4961										
3	0.5116	0.5199									
4	0.6162	0.4064	0.4255								
5	0.8941	0.5128	0.5034	0.7062							
6	0.9006	0.5480	0.5550	0.6582	0.8972						
7	0.8480	0.5654	0.5760	0.6380	0.8443	0.8760					
8	0.8508	0.4921	0.4662	0.6371	0.8026	0.8681	0.8062				
9	0.7482	0.4524	0.4882	0.7038	0.8540	0.7510	0.6880	0.6750			
10	0.4752	0.9319	0.4938	0.4163	0.5066	0.5419	0.5506	0.4631	0.4400		
11	0.8671	0.5864	0.5381	0.7047	0.9008	0.9110	0.8398	0.8257	0.8118	0.5713	
12	0.8606	0.5616	0.5466	0.6466	0.9121	0.8652	0.8194	0.7757	0.7892	0.5585	0.9318

6. Conclusions

We proposed an extension of the classical correlation coefficients for two sets of variables for multivariate functional data. We suggested permutation tests to examine the significance of the results because the values of the proposed coefficients are rather hard to interpret. The presented method has been proved to be useful as it was tested on a real data set, in investigating the correlation between two sets of variables. This example confirms its usefulness in revealing the hidden structure of the co-dependence between groups (pillars) of variables representing various fields of socio-economic activity.

REFERENCES

ANDO, T., (2009). Penalized optimal scoring for the classification of multi-dimensional functional data. Statistcal Methodology 6, 565–576.

BERRENDERO, J. R., JUSTEL, A., SVARC, M., (2011). Principal components for multivariate functional data. Computational Statistics & Data Analysis 55(9), 2619–2634.

BESSE, P., (1979). Étude descriptive d'un processus: Aproximation et interpolation, Ph.D. thesis, Université Paul Sabatier, Toulouse III.

BEUTLER, F. J., (1970). Alias-free randomly timed sampling of stochastic process. IEEE Transactions on Information Theory 16, 147–152.

ESCOUFIER, Y., (1970). Echantillonnage dans une population de variables aléatoires réelles, Ph.D thesis, Université des sciences et techniques du Languedoc, Montpellier.

ESCOUFIER, Y., (1973). Le traitement des variables vectorielles. Biometrics 29(4), 751–760.

ESCOUFIER, Y., ROBERT, P., (1979). Choosing variables and metrics by optimizing the RV coefficient. In Optimizing Methods in Statistics, Rustagi, J.S., Ed., Academic: New York, 205–219.

FERRATY, F., VIEU, P., (2003). Curve discrimination. A nonparametric functional approach. Computational Statistics & Data Analysis 44 161–173.

FERRATY, F., VIEU, P., (2009). Additive prediction and boosting for functional data. Computational Statistics & Data Analysis 53(4), 1400–1413.

GÓRECKI, T., KRZYŚKO, M., WASZAK, Ł., WOŁYŃSKI, W., (2014). Methods of reducing dimension for functional data. Statistics in Transition new series 15(2), 231–242.

GÓRECKI, T., KRZYŚKO, M., WOŁYŃSKI, W., (2015). Classification problems based on regression models for multi-dimensional functional data. Statistics in Transition new series 16(1), 97–110.

HASTIE, T. J., TIBSHIRANI, R. J., BUJA, A., (1995). Penalized discriminant analysis. Annals of Statistics 23, 73–102.

HE, G., MULLER, H. G., WANG, J. L., (2004). Methods of canonical analysis for functional data. Journal of Statistical Planning and Inference 122(1-2), 141–159.

HOTELLING, H., (1936). Relation between two sets of variables. Biometrika 28(3/4), 321–377.

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

JAMES, G. M., (2002). Generalized linear models with functional predictors. Journal of the Royal Statistical Society 64(3), 411–432.

JACQUES, J., PREDA, C., (2014). Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis 71(C), 92–106.

JOSSE, J., HOLMES, S.,(2014). Tests of independence and beyond. arXiv:1307.7383v3.

KRZYŚKO, M., WASZAK, Ł., (2013). Canonical correlation analysis for functional data. Biometrical Letters 50(2), 95–105.

LEE, A. J., (1976). On band-limited stochastic processes. SIAM Journal on Applied Mathematics 30, 169–177.

LEURGANS, S. E., MOYEED, R. A., SILVERMAN, B. W., (1993). Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society. Series B (Methodological) 55(3), 725–740.

MASRY, E., (1978). Poisson sampling and spectral estimation of continuous time processes. IEEE Transactions on Information Theory 24, 173–183.

MATSUI, H., ARAKI, Y., KONISHI, S., (2008). Multivariate regression modeling for functional data. Journal of Data Science 6, 313–331.

MENZEL, U., (2012). CCP: Significance Tests for Canonical Correlation Analysis (CCA). R package version 1.1. http://CRAN.R-project.org/package=CCP.

MÜLLER, H. G., STADMÜLLER, U., (2005). Generalized functional linear models. Annals of Statistics 33, 774–805.

R CORE TEAM (2015). 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, Second Edition, Springer.

RAMSAY, J. O., WICKHAM, H., GRAVES, S., HOOKER, G., (2014). fda: Functional Data Analysis. R package version 2.4.4. http://CRAN.R-project.org/package=fda.

REISS, P. T., OGDEN, R.T., (2007). Functional principal component regression and functional partial least squares. Journal of the American Statistcal Assosiation 102(479), 984–996.

RIZZO, M. L., SZÉKELY, G. J., (2014). energy: E-statistics (energy statistics). R package version 1.6.2. http://CRAN.R-project.org/package=energy.

ROBERT, P., ESCOUFIER, Y., (1976). A unifying tool for linear multivariate statistical methods: the RV coefficient. Journal of the Royal Statistical Society. Series C (Applied Statistics) 25(3), 257–265.

ROSSI, F., DELANNAYC, N., CONAN-GUEZA, B., VERLEYSENC, M., (2005). Representation of functional data in neural networks. Neurocomputing 64, 183–210.

ROSSI, F., VILLA, N., (2006). Support vector machines for functional data classification. Neural Computing 69, 730–742.

SZÉKELY, G. J., RIZZO, M. L., BAKIROV, N.K., (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics 35(6), 2769–2794.

SZÉKELY, G. J., RIZZO, M. L., (2009). Brownian distance covariance. Annals of Applied Statistics 3(4), 1236–1265.

SZÉKELY, G. J., RIZZO, M. L., (2012). On the uniqueness of distance covariance. Statistical Probability Letters 82(12), 2278–2282.

SZÉKELY, G. J., RIZZO, M. L., (2013). The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis 117, 193–213.

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 467–496

CHANGING MORTALITY DISTRIBUTION IN DEVELOPED COUNTRIES FROM 1970 TO 2010: LOOKING AT AVERAGES AND BEYOND THEM

Adam Szulc[1]

ABSTRACT

The methods typically developed in income inequality and poverty research are employed to observe changes in life spans distribution in 35 developed countries. The analyses are performed at two levels, using the same methods when possible: i/ taking the countries as the units with a mean length of life being a single parameter representing the distribution, ii/ utilizing the country life tables (taking people as the units) in order to compare other than mean length of life attributes of mortality distribution. Increasing divergence in the mean length of life across the countries is due to growing distance of the countries below the median, mainly the post-communist ones, to the upper half. The comparisons of the within-country distributions of ages at death by means of the Kullback-Leibler divergence provides similar results. However, poverty and inequality indices calculated at this level yield opposite conclusions. Hence, most of the between-country variation might be attributed to the variation in the mean length of life while the changes in within-country inequality reduced this effect. At the same time, huge alterations in the within-country mortality rankings can be observed. Australia, Japan, Taiwan, Austria and Luxembourg may be said to be the "winners" while most of the post-communist countries are among the "losers".

Key words: mean length of life, mortality distribution, poverty and inequality indices.

1. Introduction

Social inequality has attracted the attention of scholars since the ancient times. It was Plato who declared in "The Republic": "If a state is to avoid (…) civil disintegration (…) extreme poverty and wealth must not be allowed to rise in any section of the citizen-body, because both lead to disasters." (Plato quoted after Cowell, 1977, page 26). Since the end of the XIX Century and V. Pareto's works on income distribution (Pareto, 1896), research on inequality has become one of

[1] Warsaw School of Economics, Institute of Statistics and Demography; Madalińskiego 6/8, 02-513 Warszawa, Poland. E-mail: aszulc@sgh.waw.pl.

the inherent elements of the modern economics. Considering some opinions that the length of life is one of most comprehensive single indicators of economic development (Hicks and Streeten, 1979, Silber, 1983), as well as attention paid to this indicator by the United Nations (United Nations Development Programme, 2014) or the European Union (European Commission, 2009), the growing interest in inequality among demographers can hardly surprise (see Muszyńska et al., 2014 for a review). In this study some concepts of income inequality and poverty measurement are applied to the changes in mortality distribution observed in 35 developed countries between 1970 and 2010. We employ a set of inequality scalar measures intended to capture various aspects and ranges of mortality distribution. Moreover, indicators developed in well-being poverty research are also utilized for the same purpose. Following Ravallion (1994), poverty may be defined generally as an inability to "attain a level of material well-being deemed to constitute a reasonable minimum by the standards of that society". Employing poverty measures in demographic research was proposed by Silber (1992) who calculated three poverty indices using life tables for 27 countries, setting the "poverty line" at 60 years of age, i.e. typical retirement age for women (in that way, people were considered poor if they "did not have opportunity to enjoy retirement", Silber 1992, p. 415). In the present study the definition of poverty could be interpreted in two ways. When the country mean lengths of life are compared, poverty denotes falling below the broadly conceived "international standard". When individual data on mortality is analysed (as in Silber's study), poverty could be interpreted as "dying too early". It should be said, however, that in the present study all poverty measures following Silber's approach are relative, i.e. they should be considered a vehicle for comparing the life spans distribution, taking the "international standards" as a benchmark. As the final results strongly depend on the poverty line setting, we explore its various level(s), producing indicators for each one. The poverty indices are intended to answer the question of how many people or countries are exposed to demographic poverty but also what the depth and severity of this type of poverty is. Beyond the poverty indices mentioned above we apply the Kullback-Leibler divergence (Kullback and Leibler, 1951) in order to compare the mortality distributions in their whole ranges.

Although the number of studies comparing mortality between countries is large, only a few of them investigate other than mean length of life parameters of within-country distribution. Smits and Monden (2009), and Edwards (2011) calculated inequality indices for the whole world, then decomposing it into between- and within-country inequality. Edwards and Tuljapurkar (2005) and d'Albis et al. (2014) employed the Kullback-Leibler divergence in order to compare within-country distributions. In Muszyńska et al. (2014) the concept of Equivalent Length of Life (Silber, 1983) was incorporated to take into consideration within-country inequality and asymmetry, and to find relations between these parameters and the mean length of life. The latter studies utilized variance and Gini-based measures of inequality/asymmetry while the present one employs a broader set of distribution characteristics, which makes the conclusions potentially more robust. To our knowledge Silber (1992) is the only study comparing mortality poverty indices calculated by means of individual data. The novel feature of this study is matching comparisons between countries ("macro level") and within countries ("micro level") using, when possible, the same or similar methods at both levels. The following questions are specifically addressed:

1. How did mortality distribution change over time?
2. Were the changes at both levels similar?
3. What parameters should be used to depict those changes?
4. What is the "explanatory power" of the mean length of life in the above-mentioned context?
5. How is the mean length of life related to other characteristics of the distribution?

The data comes from the Human Mortality Database (2011) and covers 35 developed countries (for a complete list see Table A1 in Appendix). For each country the life tables including individual information on the age at death are available (more precisely, aggregate information on the age at death is equivalent to individual information, as the variable of interest is discrete). The calculations are performed for the years 1970, 1980, 1990, 2000 and 2010[2]. All the analyses are carried out for women and men separately. Following some other authors (Edwards, 2011, Edwards and Tuljapurkar, 2005, Smits and Monden, 2009) we decided to eliminate differences in infant and childhood mortality, taking 10 years of age as a cutting point. Moreover, it was our intention to make this study comparable with that by Muszyńska et al. (2014), in which inequality is also analysed for people aged 10 or over. Therefore, all the calculations, including the country mean lengths of life also have been performed for such subsets. Starting mortality analysis form 10 years of age seems to be hardly justifiable when demographic poverty, i. e. early deaths, are considered. Nevertheless, such a choice is less critical in comparing distributions of the deaths, which is the case in the present study: all poverty measures are in fact relative, not absolute. To check the impact of the above-mentioned choice on the results, some calculations were supplementary performed for the whole populations.

The remaining part of this paper is organized as follows. Theoretical issues in inequality and poverty measurement as well as the concept of the Kullback-Leibler divergence are discussed in Section 2. Section 3 reports the results of comparisons of the mean lengths of life across the countries in the form of inequality and poverty indices. Section 4 is devoted to comparisons of within-country distributions based on the life tables. In Section 5 the results displayed in the previous two sections are compared. Section 6 concludes.

[2] More precisely, for all countries the most recent data were used, not necessarily of 2010. For some of them the last available year is 2009 (see Table A1 for the details).

2. Theoretical concepts

In a standard inequality and poverty research the units of measurement are constituted by households or persons. The latter choice is applied in the present study in within-country analyses ("micro level") while at the "macro level" the countries are the units. In the first case the individual asset, being a counterpart of well-being, is the age at death. In the latter case, it is the country mean length of life.

2.1. Inequality indices

The Gini index has gained wide recognition due to its clear geometric interpretation based on Lorenz curve. It may be calculated for a variable y (income or length of life) by means of the following equation:

$$G(y) = \frac{n+1}{n-1} - \frac{2}{n(n-1)\overline{Y}} \sum_{i=1}^{n} r_i y_i \qquad (1)$$

where $\overline{Y}$ is the mean value of y, n stands for the population (or sample) size, r_i denotes the rank of *i-th* unit, after ranking the incomes/lengths of life in descending order. Many other inequality indices have been presented further on. In the present study also Theil's (1967) entropy measure and three indices belonging to Atkinson's (see Cowell, 1977) family representing three levels of inequality aversion have been calculated. However, as they yield conclusions that are generally consistent with those derived from the Gini indices, they are not displayed in the empirical part of this paper.

The Gini index and variation/standard deviation, unlike the Atkinson indices, do not assume any inequality aversion. Hence, equal changes in inequality both in upper and low ranges of distribution result in similar changes in the index value. This may be a limitation, but it may be resolved by using supplementary inequality measures in the form of percentile ratios. Those typically applied in income distribution analysis are based on lower and upper deciles (90/10) or quartiles (75/25), although there are no logical restrictions on these selections. The main disadvantage of those measures is poor responsivity to income transfers. If they do not result in changes of the predefined percentiles, a transfer from a "poor" to a "rich" person does not affect them. In other words, the strong version of the Dalton-Pigou axiom (see Fishburn, 1984) is not passed by these ratios[3].

Inequality measures in the form of percentile ratios offer a possibility of focusing on selected ranges of the distribution, although it is the researcher's responsibility to decide on which ones. As it is rather hard to choose particular range(s), it may be useful to generalize by employing some poverty indices based on a variable poverty threshold, as discussed in the next section.

[3] They pass only the weak version of the axiom: a transfer from a "poor" to a "rich" does not decrease the index.

2.2. Poverty lines and poverty indices

In its most basic approach, poverty measurement requires a decision on two concepts: i/ the poverty threshold (usually referred to as the poverty line) separating the poor from the non-poor, ii/ the method of poverty aggregation over the units, in the present study people or countries. The latter concept refers to the poverty indices theory, which provides index formulas displaying various aspects of poverty, e.g. its incidence (how many poor?) or depth (how poor are the poor?). This issue is addressed in more details below.

Income concepts of the poverty line referring its value to basic needs, mainly defined in terms of consumption, is of scarce utility in demographic context. It seems to be more reasonable to set the poverty line in relative terms, referring its level to the actual population parameters. As it is hard to find a rationale for setting the poverty line at any particular level (e.g. 50% of the mean or 60% of the median is frequently done in income poverty research), it may be justified to produce a wide range of age thresholds to observe changes in poverty indices with respect to those levels. Especially, poverty indices might be computed for the whole range of ages observed. In that way, the poverty rates would take values from zero (below the minimum age observed) to one (over the maximum age). The shape of the curve thus obtained may be one of the forms of presentation of the distribution of deaths with a clear and intuitive interpretation. This issue is discussed in more details in Section 3.3, together with the presentation of the empirical results. An answer to the question “how to aggregate individual measures of poverty into poverty indices?” depends on the type of information we are interested in. In the poverty literature, the objects of interest usually include three aspects of poverty: incidence, depth and severity. The respective indices are presented below.

The proportion of poor units (households, persons, groups, countries), referred to as a head count ratio is a measure of poverty incidence and represents the most common poverty indicator. Formally, it is defined as:

$$H = \frac{q}{n} \tag{2}$$

where q is a number of poor units while n stands for the population size. The head count ratio is not responsive to changes in poverty depth: it remains unchanged if the poor become more poor. Moreover, if some poor improve their positions, however without reaching the poverty line, it also would not affect the head count ratio. When the head count ratio is calculated using country life tables to describe within-country distributions, it may be rewritten using customary demographic symbols:

$$H = \frac{\sum_{X=\omega_{\min}}^{\omega_z} d_X}{\sum_{X=\omega_{\min}}^{\omega_{\max}} d_X}$$

where ω_Z denotes an age set as the poverty line.

In the next types of indices the question addressed is: how poor are the poor or what is the poverty severity? The Dalton index, measuring poverty depth, is defined as a relative difference between the poverty line (z) and the mean value (income or age) obtained for the poor $\overline{Y}_P$ [4]:

$$D = \frac{z - \overline{Y}_P}{z} \tag{3}$$

It is worth mentioning that leaving the poverty zone by some poor households may increase poverty depth measured by D. Consequently, this index may decrease with respect to the increase in the poverty line. This is the case for the present data, as displayed by Figures 2a and 2b.

Indices intended to measure poverty severity take into consideration not only poverty incidence and depth but also inequality among the poor. Out of several indices of this type the Sen formula has gained wide recognition due to passing a set of axioms expected to be held by poverty measures (Sen, 1976). Moreover, due to its definition it is possible to find which components, a head count ratio (H), the Dalton (D) or Gini index, are responsible for changes in poverty severity. The Sen index is defined as follows:

$$S = H[D + (1 - D) \cdot G_P] \tag{4}$$

where G_P stands for the Gini inequality index calculated for the poor.

2.3. From comparing parameters to comparing distribution functions: Kullback-Leibler divergence

All the methods described above allow comparisons of single parameters characterizing the distributions. Hence, the resulting differences in mortality distributions depend on the choice of those parameters. Using a set of parameters allows relaxing this impact, although at the cost of clarity of the final results. The scalar measure presented in this section, the Kullback-Leibler divergence (Kullback and Leibler, 1951), is intended to compare the whole ranges of distributions. Informally speaking, it measures the average distance between two probability functions. Mathematically, for two discrete distributions defined over domains from 1 to m, with probability functions P and Q, the Kullback-Leibler divergence (hereafter: KLD) is defined as follows:

$$KLD = \sum_{i=\min}^{\max} \ln\left(\frac{P_i}{Q_i}\right) P_i \tag{5}$$

[4] When income or consumption are variables of interest, a system of weights, reflecting different household sizes, should be applied in calculations. This is not applicable in demographic studies, unless one decides to take into account sub-group (e. g. countries) sizes. This makes the Dalton index equivalent to a popular poverty gap measure.

i.e. as an expected value of a logarithmic difference between probabilities, using probabilities of distribution P (in the present case: min=10 and max=110). KLD is nonnegative and equals zero if and only if the two distributions are identical. The higher the KLD, the larger divergence between the distributions. When comparing more than two distributions (the present case), one can use an average probability function, as it was done by d'Albis et al. (2014), and then calculate the mean value of all KLD with the average distribution as a point of reference.

3. Empirical results: comparisons of mean lengths of life across countries

Using the mean lengths of life for people aged 10 and over in 35 developed countries (see Introduction), the set of inequality and poverty indices presented in the previous section has been calculated for five selected years between 1970 and 2010.

3.1. Indices of mortality inequality

The trend in inequality may be generally said to be increasing over the period investigated, although for females a minor drop between 2000 and 2010 occurred for most of the indices. There is one important exception from that rule, however. The ratio of the ninth and the fifth decile is relatively stable and 2010 levels are even slightly below those of 1970. This means that the prevailing portion of the increase in inequality has occurred due to the increase in inequality below the median length of life. In other words, in some countries the mean length of life was increasing slower than the average pace or even decreased, in spite of the general increases. The results reported in Section 4.3, especially in Tables 6a and 6b, confirm this hypothesis. Moreover, one can indicate the group of post-communist, especially ex-Soviet countries as the main source of that divergence. The inequality for all years has been much higher for males than for females. Also, the increases were stronger for males.

Table 1a. Inequality in mean length of life: females, age 10+.

Inequality measure	1970	1980	1990	2000	2010
Standard deviation	1.341	1.968	2.341	2.862	2.800
Coefficient of variation	0.018	0.026	0.030	0.036	0.035
Gini index·100	0.973	1.422	1.667	1.974	1.843
Q_{75}/Q_{25}	1.023	1.049	1.058	1.055	1.055
Q_{90}/Q_{10}	1.042	1.061	1.073	1.100	1.089
Q_{90}/Q_{50}	1.027	1.031	1.024	1.024	1.023
Q_{50}/Q_{10}	1.014	1.029	1.048	1.074	1.064
Mean length of life	75.58	76.84	78.12	79.52	81.48

Table 1b. Inequality in mean length of life: males, age 10+.

Inequality measure	1970	1980	1990	2000	2010
Standard deviation	2.239	2.970	3.508	4.916	5.546
Coefficient of variation	0.032	0.043	0.049	0.068	0.074
Gini index·100	1.683	2.326	2.730	3.592	3.788
Q_{75}/Q_{25}	1.042	1.060	1.097	1.098	1.115
Q_{90}/Q_{10}	1.070	1.114	1.136	1.188	1.228
Q_{90}/Q_{50}	1.039	1.040	1.029	1.038	1.031
Q_{50}/Q_{10}	1.030	1.0718	1.104	1.145	1.192
Mean length of life	69.24	69.74	70.91	72.44	74.74

3.2. Indices of length of life poverty

The poverty lines are set at the median and first quartile in the length of life distributions obtained for all countries. All indices are calculated twice: i/ using the current information on the country mean length of life for each year and ii/ as an appropriate percentile in 1970 distribution (the poverty lines are then fixed over the whole period observed). Consequently, in the first case all the head count ratios are close to 0.5 (not 0.5 exactly, as the number of countries is odd) or to 0.25. In the second case these values are reached for 1970 only. Using the above-mentioned percentiles as the poverty lines is an arbitrary choice, although the problem of the poverty line selection may be resolved by making the poverty line variable, as presented further on.

Another goal of this part of the study is to compare the head count ratios that are fixed over time, with changes of the remaining indices. The Dalton index values demonstrate that the poor, in terms of mortality, countries have become poorer over the period observed, i.e. the distance between their mean length of life and the percentiles selected as the poverty lines has increased. This is also true, though with some exceptions, for the results based on the first quartile. Both single year values and increases of the Dalton index were higher for males than for females and higher when median poverty line was used. Given stable, by the definition, head count ratios and increases in the Dalton and Gini indices, the Sen indices, being measures of poverty severity (or comprehensive poverty), also displayed an increasing trend. Nevertheless, in 1990 for males and the poverty line set at the first quartile, a massive drop in the Sen index occurred due to the drops in poverty depth and inequality among the poor countries. On the other hand, 2010 value was much above 1970 level, due to strong increases occurring in the remaining years. For females, in 2010 all of them (naturally, except for the head count ratios) dropped as compared to 2000 levels, while the changes for males were in opposite direction.

The head count ratios based on the fixed (at 1970 levels) poverty lines demonstrate serious decreases in poverty incidence. In other words, the number of countries that do not pass “1970 standards” decreased. On the other hand, for males approximately one sixth of the countries (i.e. six of them) did not reach 1970 median value in 2010. For females such a proportion was three times lower. Moreover, for males the drops observed under the lower poverty line were less important than those observed for the higher one: in 2010 14.3% of the countries were below 1970 quartile. Relative comparisons between two poverty lines are not conclusive for females, as in 2010 all countries were placed at or above 1970 quartile.

Including mortality below 10 years of age slightly changes the absolute values of poverty (and inequality) indices, with no regularities observed. Nevertheless, the basic conclusions on trends remain unchanged.

Table 2a. Length of life poverty (in %), poverty line at median and first quartile of current distribution: females, age 10+.

Indicator	1970	1980	1990	2000	2010
Median	75.34	76,79	78.89	81.78	82.52
Head count ratio	48.571	48.571	48.571	48.571	48.571
Dalton	1.089	2.104	3.582	4.194	3.900
Sen	0.812	1.421	2.236	2.837	2.732
$\text{Gini}_{\text{poor}}\cdot 100$	0.590	0.840	1.059	1.718	1.795
First quartile	74.91	75.17	75.88	77.55	79.01
Head count ratio	22.857	22.857	22.857	22.857	22.857
Dalton	1.374	1.110	1.191	3.047	2.302
Sen	0.457	0.432	0.438	0.913	0.787
$\text{Gini}_{\text{poor}}\cdot 100$	0.635	0.788	0.732	0.976	1.169

Table 2b. Length of life poverty (in %), poverty line at median and first quartile of current distribution: males, age 10+.

Indicator	1970	1980	1990	2000	2010
Median	69.12	70.27	72.38	74.39	77.13
Head count ratio	48.571	48.571	48.571	48.571	48.571
Dalton	2.227	4.107	6.164	7.781	8.495
Sen	1.715	2.956	3.983	5.475	6.065
$\text{Gini}_{\text{poor}}\cdot 100$	1.334	2.064	2.170	3.786	4.361
First quartile	68.07	67.58	67.19	69.15	70.85
Head count ratio	22.857	22.857	22.857	22.857	22.857
Dalton	2.507	3.325	2.023	6.496	6.641
Sen	0.958	1.108	0.665	1.973	2.290
$\text{Gini}_{\text{poor}}\cdot 100$	1.729	1.577	0.905	2.284	3.618

Table 3a. Length of life poverty (in %), poverty line at median and first quartile of 1970 distribution: females, age 10+.

Indicator	1970	1980	1990	2000	2010
1970 median			75.34		
Head count ratio	48.571	25.714	11.429	11.429	5.714
Dalton	1.089	1.211	1.337	1.544	0.338
Sen	0.812	0.499	0.271	0.256	0.025
$\text{Gini}_{\text{poor}} \cdot 100$	0.590	0.737	1.052	0.702	0.995
1970 first quartile			74.91		
Head count ratio	22.857	14.286	5.714	8.571	0.000
Dalton	1.374	1.374	1.976	1.366	0.000
Sen	0.457	0.337	0.194	0.171	0.000
$\text{Gini}_{\text{poor}} \cdot 100$	0.635	0.995	1.440	0.634	-

Table 3b. Length of life poverty (in %), poverty line at median and first quartile of 1970 distribution: males, age 10+.

Indicator	1970	1980	1990	2000	2010
1970 median			69.12		
Head count ratio	48.571	31.429	31.429	22.857	17.143
Dalton	2.227	4.470	4.059	6.444	6.396
Sen	1.715	1.900	1.597	1.961	1.623
$\text{Gini}_{\text{poor}} \cdot 100$	1.334	1.648	1.064	2.284	3.279
1970 first quartile			68.07		
Head count ratio	22.857	28.571	28.571	22.857	14.286
Dalton	2.507	3.314	2.849	5.006	6.138
Sen	0.958	1.396	1.083	1.640	1.291
$\text{Gini}_{\text{poor}} \cdot 100$	1.729	1.624	0.970	2.284	3.092

3.3. Poverty incidence and depth as a function of the poverty line

The results displayed in the previous section depend on an arbitrary selection of the poverty line. Given a lack of ground for setting this threshold at any particular level(s), it seems to be justified to calculate indices for the whole range of variability, obtaining in this way a type of distribution function. For clarity of the plots only the datasets for the years 1970, 1990 and 2010 were applied. Changes in the head count ratios and the Dalton indices measuring poverty depth are displayed by means of Figures 1a - 1b, and 2a - 2b, respectively. The range of the poverty line is set from the lowest to the highest values observed in the whole dataset (the head count ratios equal zero or one, respectively, at these values).

The most obvious observation is on growing ranges of distribution in the succeeding years, as may be deducted from Figures 1a and 1b. This coincides with inequality growths reported in the previous section. For females, this happened solely due to the increases in the maximum age but for males also due to the drops in the minimum value in 2010, as compared to both previous years. Another conclusion is on definitely non-linear growth of the poverty incidence with respect to the poverty line. At the bottom ranges of the distributions, the head count ratios were relatively stable or were growing at moderate pace and then experienced sharp growths, reaching the maximum value, i.e. one. This indicates the existence of the relatively homogenous groups of countries with low life expectancy. For females, for all years such sharp growths may be observed starting from 74 to 75 years of age, but for males the turning points were absolutely different, growing considerably during each 20 year period. Widening ranges of sharp growth may be interpreted as growing polarization of the mean lengths of life, which is reinforced for males by the previously mentioned occurrence. For the Dalton indices no regularities can be observed. Unlike the head count ratios, this index for all years and for both sexes suffered some drops due to the poverty line growth, though the general trend was increasing. As the Sen index depends also on inequality, which is not related directly to the poverty line, the resulting changes would not be informative.

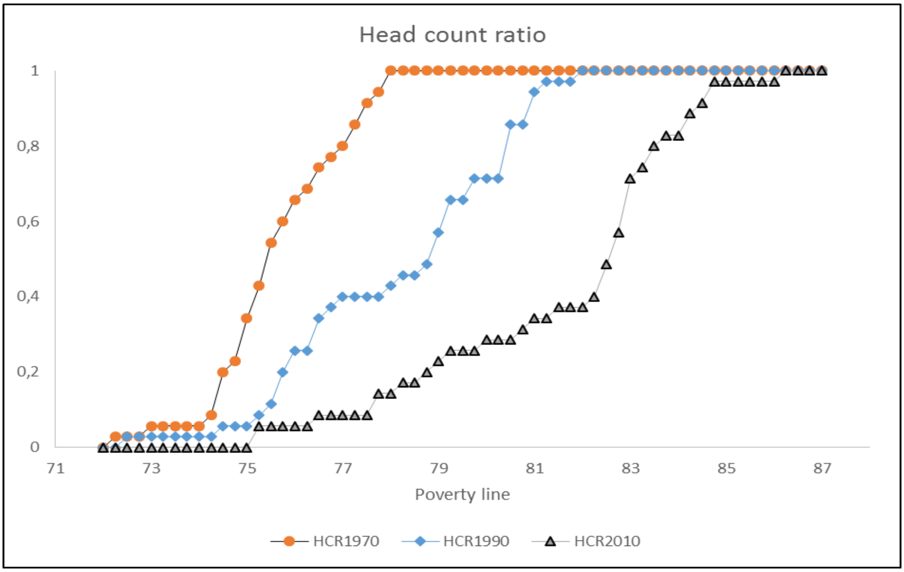


Figure 1a. Length of life poverty incidence depending on poverty line for 1970, 1990 and 2010: females, age 10+.

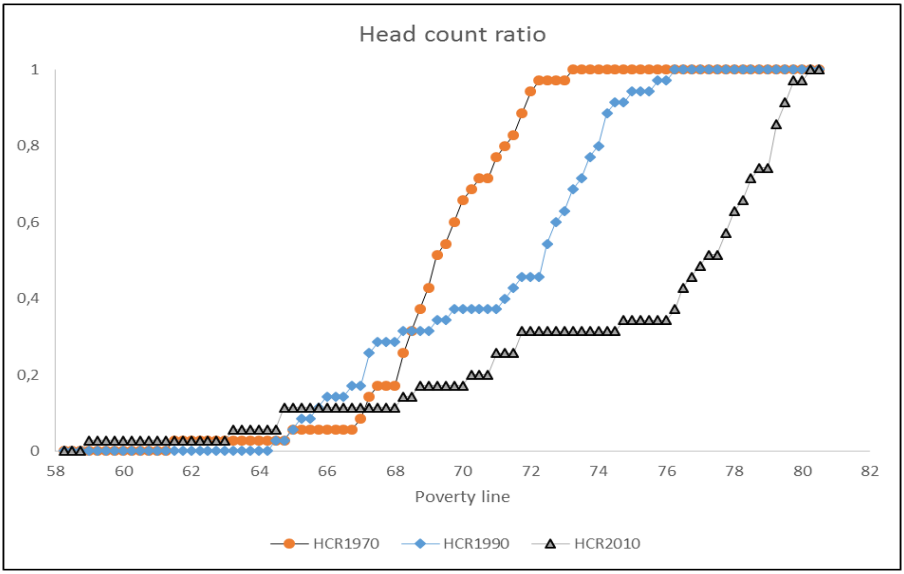


Figure 1b. Length of life poverty incidence depending on poverty line for 1970, 1990 and 2010: males, age 10+.

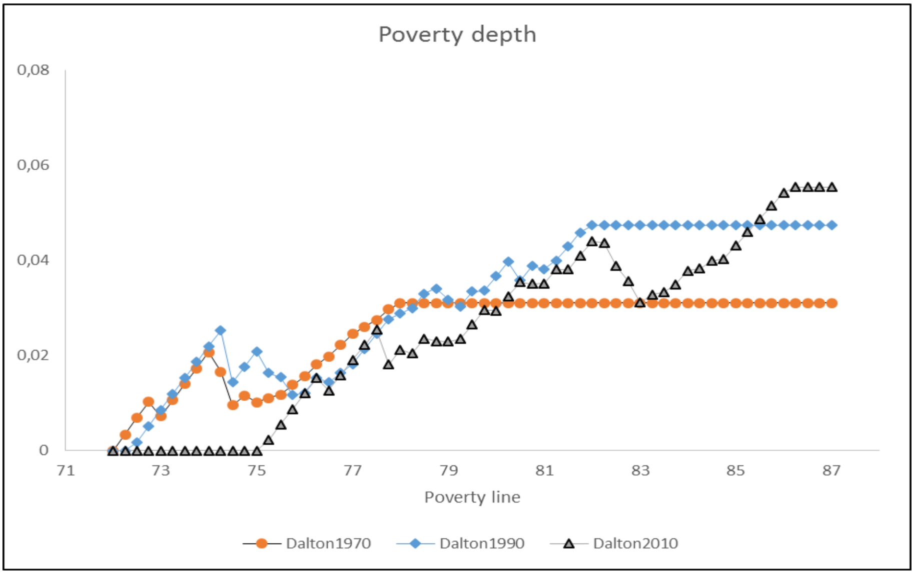


Figure 2a. Length of life poverty depth (Dalton index) for 1970, 1990 and 2010: females, age 10+.

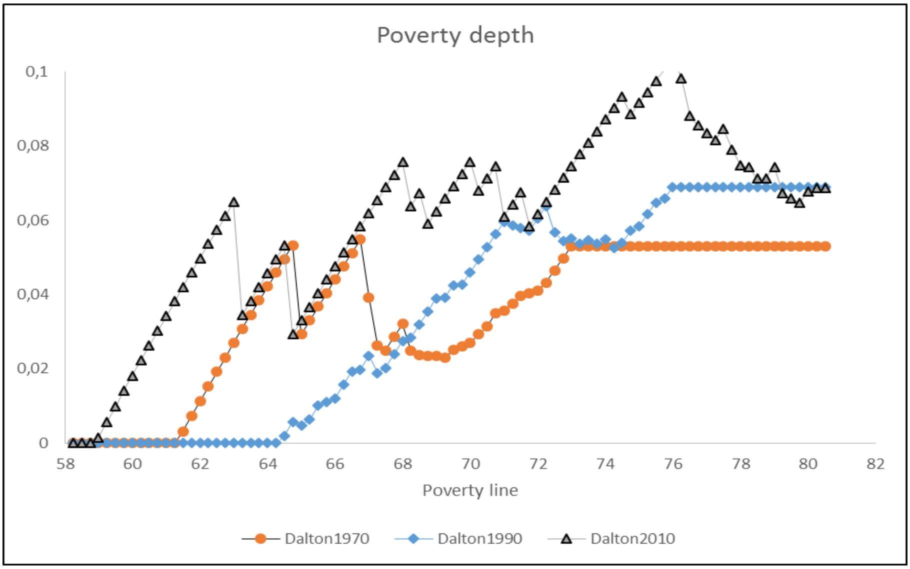


Figure 2b. Length of life poverty depth (Dalton index) for 1970, 1990 and 2010: males, age 10+.

4. Empirical results: comparing distributions within the countries

4.1. Kullback-Leibler divergence

In this section probability functions representing mortality distributions within-country are compared in their whole ranges, using country life tables. Consequently, the results are not sensitive to the choice of a single parameter supposed to represent the distribution (in other parts of this study: the mean length of life or poverty/inequality index). First, the mean probability function for the whole population is constructed and then for each country the Kullback-Leibler divergence (KLD) is calculated. The average indicator is a measure of “overall divergence” across the countries under comparison. Tables 4a and 4b display the mean KLDs for the years 1970, 1980, 1990, 2000 and 2010 (bottom row). Moreover, for each year a list of five countries with highest values (i.e. highest distance from the mean distribution) is included. Divergence in distributions does not necessary result in divergence in the mean values, although makes them very likely.

The trends in the mean KLD values are generally consistent with the trends in the country mean values, reported in the previous sections: the average distance between countries has been growing over the whole period observed. This is true for both sexes and, again, the intensity of this process appeared to be much higher for males. The countries with largest distances from the mean distribution include those with the low and high mean length of life. For instance, in 1970 for males two high mortality countries (Czech Republic and Russia) and three low mortality ones (Luxembourg, Iceland and Sweden) are present. For two succeeding years only one low mortality country (Iceland) is ranked among those most distant from the mean distribution, while for 2000 and 2010 all countries in Top 5 are high mortality countries. Similar process may be found for females, although in 2000 and 2010 one low mortality country (Japan) was ranked on the top. Generally, for most of the years and for both sexes high mortality countries are more distant from the mean distribution than the low mortality ones.

Table 4a. Kullback-Leibler divergence for five countries with highest distance to mean distribution: females, age 10+.

KLD rank	1970		1980		1990		2000		2010	
	Country	KLD	Country	KLD	Country	KLD	Country	KLD	Country	KLD
1	DEN	0.0451	DEN	0.0965	DEN	0.1228	RUS	0.1483	UKR	0.1720
2	BLR	0.0359	CZE	0.0597	BUL	0.0600	UKR	0.1364	RUS	0.1499
3	ICE	0.0331	ICE	0.0516	JAP	0.0555	BUL	0.1173	BUL	0.1295
4	CAN	0.0272	HUN	0.0459	CZE	0.0542	DEN	0.1067	BLR	0.1187
5	CZE	0.0262	EGE	0.0386	HUN	0.0539	JAP	0.1038	JAP	0.1060
Mean-all countries	-	0.0136	-	0.0221	-	0.0278	-	0.0378	-	0.0380

Table 4b. Kullback-Leibler divergence for five countries with highest distance to mean distribution: males, age 10+.

KLD rank	1970		1980		1990		2000		2010	
	Country	KLD	Country	KLD	Country	KLD	Country	KLD	Country	KLD
1	RUS	0.0444	RUS	0.0954	RUS	0.0854	RUS	0.3423	RUS	0.3224
2	SWE	0.0353	ICE	0.0572	LAT	0.0847	UKR	0.2325	BLR	0.3077
3	ICE	0.0330	LAT	0.0530	HUN	0.0785	BLR	0.1973	UKR	0.2770
4	CZE	0.0296	CZE	0.0477	EST	0.0734	LAT	0.1296	LAT	0.1208
5	LUX	0.0282	EST	0.0465	ICE	0.0729	EST	0.1191	LIT	0.1191
Mean-all countries	-	0.0145	-	0.0231	-	0.0337	-	0.0607	-	0.0673

Note: EGE denotes Eastern Germany; for a full list of countries and abbreviations see Appendix.

4.2. Indices of within-country poverty and inequality

In this section, the within-country mortality distributions are compared by means of poverty and inequality indices. They are intended to indicate other than the mean length of life parameters responsible for growing differences between mortality distribution functions reported in the previous section. The indices are calculated using within-country information on individual age at death while the poverty lines are based on the "international" median length of life in a given year. Hence, the resulting indices depend on both mean values and shapes of distribution.

The head count ratio is a non-decreasing function of the poverty line and, therefore, is indirectly related to the mean length of life. This is not necessarily true for the Dalton index measuring poverty depth (see Figures 2a and 2b) and, consequently, for the Sen index neither. The head count ratio depends also on the shape of distribution over the whole range while the Dalton index is influenced by distribution below the poverty line only. The Sen index, as a combination of these two measures, as well as the Gini index among the "poor", takes into account the largest set of the distribution attributes.

In Tables 5a – 5b the average values of the three above-mentioned poverty indices are shown. They are supplemented by four inequality indices, namely the Gini index and three decile ratios 9th/1st, 9th/5th and 5th/1st. Unlike in the between-country mean length of life comparisons (see Tables 1a and 1b) all measures indicate decreasing trend in inequality. The largest relative drops may be observed for the Gini indices and the smallest for 9th and 5th decile ratios, which suggests that inequality decline affected mainly middle and lower ranges of the distributions. In other words, the greatest progress was made in early mortality reduction, at least for those aged 10 years and over.

Table 5a. Average within-country length of life poverty (in %) and inequality: females, age 10+.

Index	1970	1980	1990	2000	2010
Poverty line at the international median					
Head count ratio	48.086	46.967	49.383	47.455	47.250
Dalton	16.248	15.976	15.654	15.344	14.715
Sen	11.541	11.096	11.437	10.863	10.418
Poverty line at the international first quartile					
Head count ratio	24.890	24.432	0.24034	24.747	24.104
Dalton	18.035	17.625	17.265	16.976	16.499
Sen	6.652	6.368	6.135	6.240	5.914
Gini·100	9.351	9.097	8.858	8.561	8.134
Q91	1.531	1.517	1.498	1.483	1.453
Q95	1.148	1.142	1.137	1.130	1.122
Q51	1.333	1.328	1.316	1.311	1.295

Table 5b. Average within-country length of life poverty (in %) and inequality: males, age 10+.

Index	1970	1980	1990	2000	2010
Poverty line at the international median					
Head count ratio	48.456	49.357	48.702	49.203	47.270
Dalton	19.737	19.780	19.601	19.327	18.452
Sen	14.058	14.375	14.100	14.232	13.065
Poverty line at the international first quartile					
Head count ratio	23.831	23.501	23.453	24.451	23.622
Dalton	22.025	21.750	21.544	21.193	19.634
Sen	7.768	7.571	7.461	7.679	6.903
Gini·100	11.659	11.567	11.368	11.018	10.313
Q91	1.729	1.728	1.701	1.683	1.628
Q95	1.192	1.187	1.185	1.177	1.166
Q51	1.451	1.453	1.434	1.425	1.392

Decreasing average mortality inequality does not necessarily result in decreases in average poverty as the latter depend also on the mean length of life. Nevertheless, declines in within-country inequality were strong enough to compensate divergence in average life spans between the countries. As might be expected, the highest reduction has been observed for poverty depth (as a result of reduction of early mortality), the lowest for poverty incidence (as a result of increase in “international” poverty line). The latter peaked in 1990 (females) or in 1980 (males), although 2010 values were slightly below the initial ones. Similar trends are revealed when the first quartile (and also the third one, although these results are not produced here) is applied as the poverty line. The only important differences are in the positions of the peaks and in the magnitude of drops in the head count ratio that have been smaller for the first (and third) quartile. Considering the results reported in this section, together with the results of between-country comparisons, one can conclude that the mean length of life is the only parameter of distribution for which growing divergence between the countries may be observed. In Section 4.4 some relations between this parameter and other distribution attributes are investigated.

Table 6a. Mean length of life and length of life poverty rankings in 1970 and 2010: females, age 10+.

Country	Ranking for:											
	Mean			Head count ratio			Dalton index			Sen index		
	1970	2010	Difference	1970	2010	Difference	1970	2010	Difference	1970	2010	Difference
AUS	19	6	**13**	20	5	**15**	25	15	**10**	25	7	**18**
AUT	27	10	**17**	26	11	**15**	16	8	**8**	20	9	**11**
BEL	18	21	**-3**	19	19	**0**	17	22	**-5**	17	20	**-3**
BLR	5	33	**-28**	8	33	**-25**	31	33	**-2**	11	33	**-22**
BUL	22	31	**-9**	28	32	**-4**	5	24	**-19**	19	31	**-12**
CAN	4	9	**-5**	4	10	**-6**	29	23	**6**	8	17	**-9**
CZE	33	25	**8**	33	25	**8**	8	9	**-1**	30	23	**7**
DEN	35	32	**3**	35	31	**4**	34	29	**5**	35	32	**3**
EGE	30	17	**13**	30	17	**13**	12	7	**5**	26	14	**12**
ENG	12	14	**-2**	13	16	**-3**	19	17	**2**	10	18	**-8**
EST	16	24	**-8**	15	24	**-9**	21	27	**-6**	14	24	**-10**
FIN	23	12	**11**	25	9	**16**	3	20	**-17**	13	15	**-2**
FRA	8	2	**6**	6	2	**4**	24	26	**-2**	9	5	**4**
HUN	32	30	**2**	32	30	**2**	13	28	**-15**	29	30	**-1**
ICE	2	7	**-5**	1	8	**-7**	2	1	**1**	2	6	**-4**
IRE	29	22	**7**	31	22	**9**	23	11	**12**	31	22	**9**
ITA	10	5	**5**	10	6	**4**	10	3	**7**	7	4	**3**
JAP	17	1	**16**	17	1	**16**	18	21	**-3**	16	1	**15**
LAT	21	29	**-8**	18	28	**-10**	28	31	**-3**	27	29	**-2**
LIT	11	27	**-16**	11	27	**-16**	26	32	**-6**	15	28	**-13**
LUX	31	16	**15**	24	15	**9**	32	12	**20**	33	13	**20**
NED	6	18	**-12**	5	18	**-13**	6	16	**-10**	5	19	**-14**
NOR	1	13	**-12**	3	14	**-11**	1	10	**-9**	1	12	**-11**
NZL	13	11	**2**	14	12	**2**	22	14	**8**	12	10	**2**
POL	20	26	**-6**	21	26	**-5**	14	25	**-11**	18	26	**-8**
POR	25	20	**5**	27	20	**7**	15	6	**9**	23	16	**7**
RUS	26	35	**-9**	22	34	**-12**	30	35	**-5**	32	35	**-3**
SPA	9	4	**5**	9	4	**5**	9	2	**7**	6	3	**3**
SUI	7	3	**4**	7	3	**4**	4	5	**-1**	4	2	**2**
SVK	28	28	**0**	29	29	**0**	11	19	**-8**	24	27	**-3**
SWE	3	8	**-5**	2	7	**-5**	7	4	**3**	3	8	**-5**
TAI	34	19	**15**	34	21	**13**	33	18	**15**	34	21	**13**
UKR	15	34	**-19**	16	35	**-19**	27	34	**-7**	22	34	**-12**
USA	14	23	**-9**	12	23	**-11**	35	30	**5**	28	25	**3**
WGE	24	15	**9**	23	13	**10**	20	13	**7**	21	11	**10**

Table 6b. Mean length of life and length of life poverty rankings in 1970 and 2010: males, age 10+.

Country	Ranking for:											
	Mean			Head count ratio			Dalton index			Sen index		
	1970	2010	Difference	1970	2010	Difference	1970	2010	Difference	1970	2010	Difference
AUS	23	2	**21**	25	3	**22**	16	15	**1**	20	6	**14**
AUT	26	16	**10**	28	16	**12**	19	17	**2**	23	16	**7**
BEL	18	19	**-1**	22	19	**3**	8	18	**-10**	16	19	**-3**
BLR	16	34	**-18**	12	34	**-22**	32	33	**-1**	26	33	**-7**
BUL	7	29	**-22**	7	29	**-22**	9	26	**-17**	6	28	**-22**
CAN	10	9	**1**	11	9	**2**	20	14	**6**	13	11	**2**
CZE	31	25	**6**	33	25	**8**	17	20	**-3**	28	25	**3**
DEN	2	17	**-15**	5	18	**-13**	4	6	**-2**	2	15	**-13**
EGE	17	20	**-3**	18	20	**-2**	5	19	**-14**	14	20	**-6**
ENG	12	8	**4**	17	8	**9**	1	9	**-8**	8	9	**-1**
EST	34	28	**6**	32	27	**5**	31	30	**1**	33	29	**4**
FIN	33	23	**10**	34	23	**11**	25	22	**3**	30	22	**8**
FRA	15	14	**1**	13	13	**0**	24	23	**1**	17	17	**0**
HUN	22	30	**-8**	21	30	**-9**	18	29	**-11**	19	30	**-11**
ICE	5	4	**1**	3	2	**1**	28	5	**23**	11	3	**8**
IRE	13	18	**-5**	16	17	**-1**	3	16	**-13**	12	18	**-6**
ITA	9	6	**3**	9	7	**2**	13	3	**10**	9	4	**5**
JAP	11	5	**6**	10	6	**4**	10	13	**-3**	10	7	**3**
LAT	32	31	**1**	29	32	**-3**	33	31	**2**	34	31	**3**
LIT	27	32	**-5**	20	31	**-11**	34	32	**2**	32	32	**0**
LUX	28	15	**13**	30	14	**16**	23	10	**13**	25	14	**11**
NED	4	11	**-7**	6	11	**-5**	2	1	**1**	3	5	**-2**
NOR	3	10	**-7**	2	10	**-8**	7	7	**0**	4	10	**-6**
NZL	14	7	**7**	15	5	**10**	12	11	**1**	15	8	**7**
POL	20	26	**-6**	19	26	**-7**	21	28	**-7**	21	27	**-6**
POR	19	22	**-3**	14	21	**-7**	26	21	**5**	22	21	**1**
RUS	35	35	**0**	35	35	**0**	35	35	**0**	35	35	**0**
SPA	6	12	**-6**	4	12	**-8**	14	12	**2**	5	13	**-8**
SUI	8	1	**7**	8	1	**7**	11	4	**7**	7	1	**6**
SVK	24	27	**-3**	24	28	**-4**	27	24	**3**	24	26	**-2**
SWE	1	3	**-2**	1	4	**-3**	6	2	**4**	1	2	**-1**
TAI	30	24	**6**	31	24	**7**	22	25	**-3**	27	24	**3**
UKR	29	33	**-4**	26	33	**-7**	30	34	**-4**	31	34	**-3**
USA	25	21	**4**	27	22	**5**	29	27	**2**	29	23	**6**
WGE	21	13	**8**	23	15	**8**	15	8	**7**	18	12	**6**

4.3. Poverty rankings: changes between 1970 and 2010

While the results reported in the previous sections focus on the overall size of changes in mortality distributions, in this part of the analysis the alterations in relative mortality are observed for every country separately. More precisely, the changes in the country rankings for the mean length of life and poverty indices based on country data life tables are reported. The poverty lines are set at median values in the actual aggregate (i.e. capturing all countries) distributions. In Tables 6a and 6b for every country rankings (in ascending order for the poverty indices and in descending order for the mean; 1 means "best value") the initial and the last year of observation are compared. A positive difference between the rankings indicates a relative improvement in the ranking, i.e. a decrease in mortality, as compared to the remaining countries.

The comparison of 1970 and 2010 rankings reveals huge alterations in terms of all measures. For all indicators and for both sexes most of the post-communist countries' relative positions worsened seriously. To the highest extent this can be said for some ex-Soviet countries: Russia, Belarus and Ukraine as well as Bulgaria, to the least for the Czech Republic and Slovakia. The Czech Republic's rankings improved in terms of most of measures. Among the richest countries the highest relative deterioration in rankings occurred for Norway and the Netherlands. The main winners are three non-European countries: Australia, Japan and Taiwan, as well as Luxembourg and Austria, and, to less extent, Western Germany. In Eastern Germany, serious ranking improvement occurred for females only. It is interesting that for many countries the Dalton index measuring poverty depth yields rankings quite different from those constructed with the use of the means or the poverty incidence. It may be supposed that this results from different changes in mortality in various age groups. This influenced also changes in the Sen indices. Changes in the mean values and poverty incidence were in most of the cases close to each other. Including mortality below 10 years of age does not alter rankings by the head count ratios and slightly changes those by the mean length of life. More important changes may be observed in rankings by the Dalton and Sen indices, nevertheless even in that case they are moderate.

Additionally, the intensity of the ranking changes during each decade included into the study is measured. For each two succeeding decades Spearman correlation coefficients are calculated (the higher the value, the closer the ranking). It may be observed from the results reported in Tables 7a - 7b that the most important ranking alterations have occurred between 1970 and 1980. All respective correlations were below 0.9 or even 0.8. The results observed after 1990 for men and after 2000 for women indicate much higher stability of the rankings than during the previous periods. It may be also observed that the Dalton and Sen indices appeared to be much more stable over the whole period observed, however with some exceptions occurring during selected decades. Moreover, the rankings for women were less stable than those for men.

Table 7a. Spearman correlations for length of life rankings (mean value and poverty) from 1970 to 2010: females, age 10+.

Measure	1970-1980	1980-1990	1990-2000	2000-2010	1970-2010
Mean	0.7924	0.9230	0.9291	0.9759	0.4431
Head count ratio	0.7992	0.9258	0.9227	0.9678	0.4524
Dalton	0.8599	0.8725	0.7884	0.9347	0.5996
Sen	0.7686	0.9244	0.9392	0.9759	0.5286

Table 7b. Spearman correlations for length of life rankings (mean value and poverty) from 1970 to 2010: males, age 10+.

Measure	1970-1980	1980-1990	1990-2000	2000-2010	1970-2010
Mean	0.7955	0.9608	0.9762	0.9818	0.6462
Head count ratio	0.7515	0.9513	0.9627	0.9711	0.5555
Dalton	0.8826	0.9269	0.9538	0.9006	0.7104
Sen	0.8821	0.9527	0.9499	0.9798	0.7490

4.4. Poverty and inequality versus mean length of life

Referring to the findings presented in Section 3.1, indicating a positive correlation between the mean length of life and inequality at the aggregate levels, one can rise a similar question at the level of single countries. In the present study this question is addressed by means of simple graphical methods as well as using estimates of linear equations taking the general form:

$$I_i = \alpha_1 e_i + \alpha_0 + \varepsilon_i \tag{6}$$

where I_i is a poverty or inequality index for i-th country, e_i stands for a mean length of life while ε_i represents a stochastic disturbance. As may be observed in Figures 3a – 3d, all three measures employed: the Gini inequality index as well as poverty incidence and depth measures[5] are negatively correlated with the mean length of life. This is true for both sexes and for both years observed (1970 and

[5] For the sake of the clarity of the plots, Sen index, which takes values similar to the Gini index, is not included.

2010). It is also evident that the head count ratio is much more sensitive to changes in the mean length of life than two remaining indicators. This finding might be definitely expected, as rising the mean value (i.e. the independent variable in eqn 6) is equivalent to lowering the poverty line. It should be noted, however, that within-country mortality distribution also affects the poverty incidence. As a result, within some small groups of countries, a higher mean may be matched with higher poverty rate. In other words, contrary to the results obtained for the aggregate measures (see Figures 1a - 1b), the relation between poverty rate and the mean length of life is not monotonic. A negative correlation with the mean length of life may be also observed for the Gini and Dalton indices, although the elasticities are much lower than in the case of the previous indicator.

Table 8a. Spearman correlations for length of life poverty and mean length of life rankings from 1970 to 2010: females, age 10+.

Poverty index	1970	1980	1990	2000	2010
Head count ratio	-0.9782	-0.9824	-0.9891	-0.9947	-0.9944
Dalton	-0.2647	-0.4913	-0.6797	-0.6440	-0.6901
Sen	-0.9028	-0.9661	-0.9804	-0.9933	-0.9759

Table 8b. Spearman correlations for length of life poverty and mean length of life rankings from 1970 to 2010: males, age 10+.

Poverty index	1970	1980	1990	2000	2010
Head count ratio	-0.9692	-0.9840	-0.9829	-0.9919	-0.9964
Dalton	-0.6902	-0.7412	-0.8126	-0.8059	-0.8933
Sen	-0.9555	-0.9706	-0.9762	-0.9840	-0.9866

Supplementary Spearman correlation coefficients between rankings produced for the mean length of life and three poverty indices are calculated. They are displayed in Tables 8a – 8b. As might be expected, all correlations are strongly negative, with the lowest absolute value observed for the Dalton index. In that case large disparities between sexes appeared – absolute values for males were much higher than for females. This finding is consistent with the results of regressions of the Dalton index on the mean length of life reported in Tables 10a and 10b.

Finally, for inequality and poverty indices simple regressions generally defined by (6) were run. The estimates are reported in Tables 9a – 12b. They confirm much higher, negative elasticity of the head count ratios than those obtained for remaining indices. Moreover, their absolute values for males appeared to be much lower than for females, although this observation is not confirmed in the case of other indices. R-squared, being a type of measure of linear correlation between the dependent and independent variables, is the only parameter for which a clear increasing trend occurred between 1970 and 2010. This may be interpreted as increasing importance of the mean length of life in determining within-country mortality inequality and poverty.

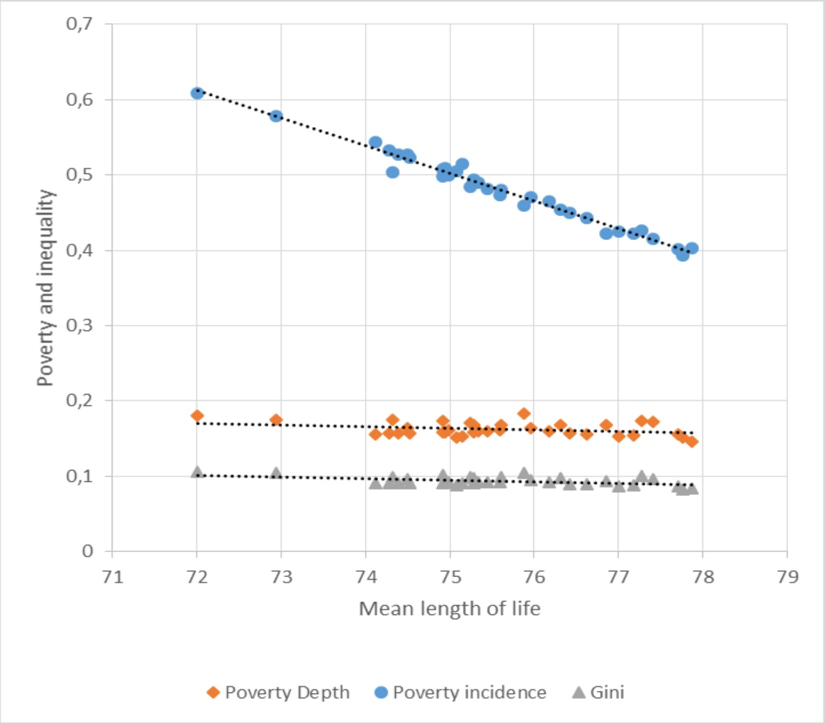


Figure 3a. Length of life poverty and inequality versus mean length of life in 1970: females, age 10+.

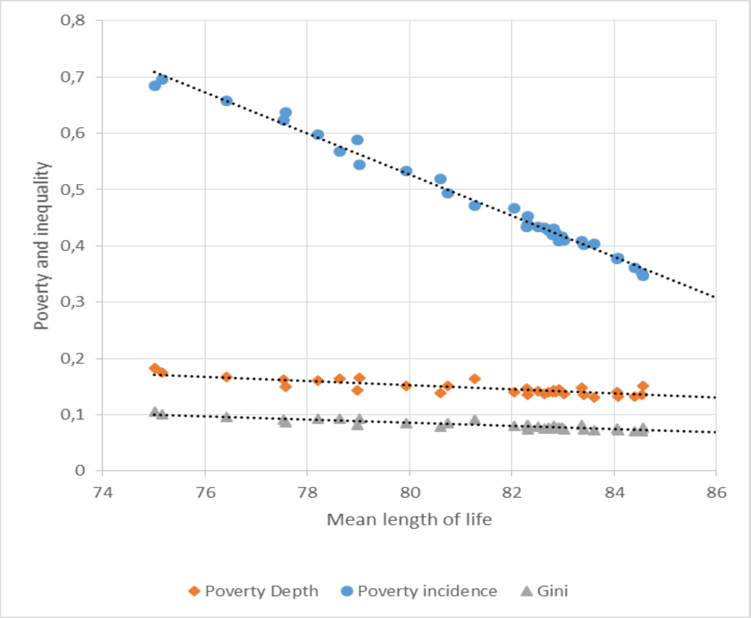


Figure 3b. Length of life poverty and inequality versus mean length of life in 2010: females, age 10+.

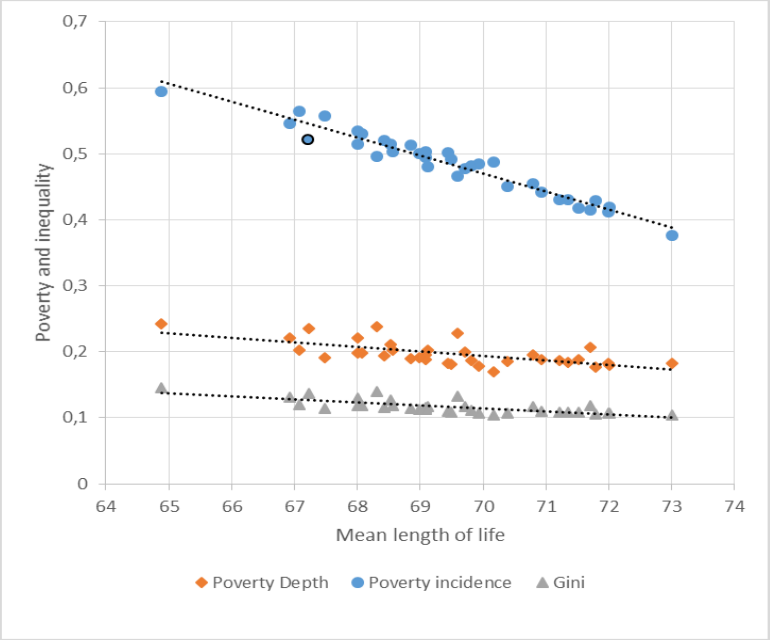


Figure 3c. Length of life poverty and inequality versus mean length of life in 1970: males, age 10+.

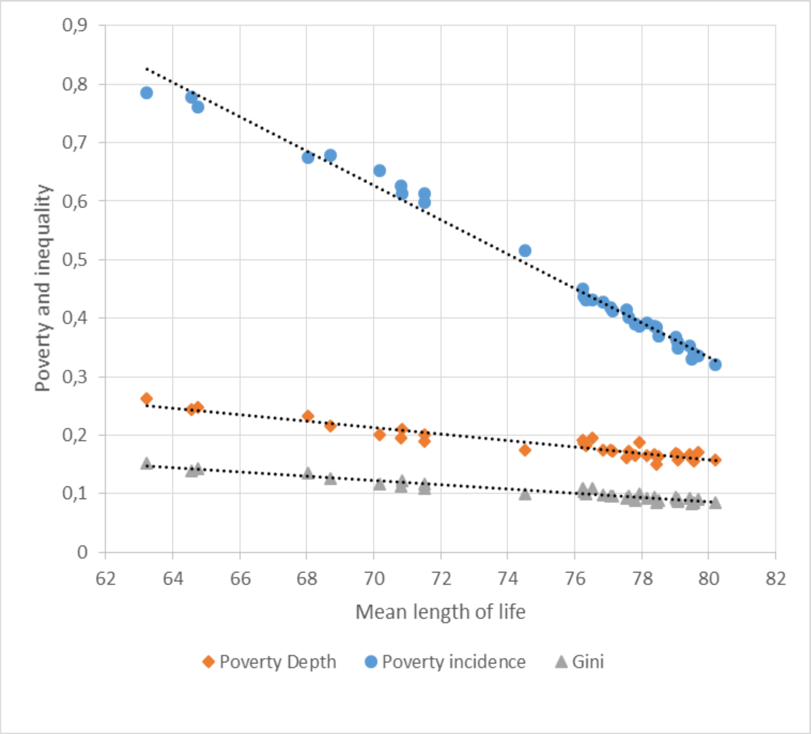


Figure 3d. Length of life poverty and inequality versus mean length of life in 2010: males, age 10+.

Table 9a. Linear regression of length of life poverty incidence on mean length of life: females, age 10+.

Parameter	1970	1980	1990	2000	2010
Slope	-0.03677	-0.03549	-0.03692	-0.03600	-0.03650
Intercept	3.26002	3.19668	3.37783	3.33781	3.44603
R-squared	0.97919	0.97910	0.98866	0.98917	0.98983

Table 9b. Linear regression of length of life poverty incidence on mean length of life: males, age 10+.

Parameter	1970	1980	1990	2000	2010
Slope	-0.02729	-0.02658	-0.02887	-0.02789	-0.02928
Intercept	2.38019	2.35323	2.54120	2.52031	2.67654
R-squared	0.94597	0.96957	0.98789	0.98558	0.98992

Table 10a. Linear regression of poverty depth (Dalton index) on mean length of life: females, age 10+.

Parameter	1970	1980	1990	2000	2010
Slope	-0.00233	-0.00272	-0.00298	-0.00350	-0.00376
Intercept	0.33864	0.36843	0.38911	0.43159	0.45371
R-squared	0.12327	0.29678	0.53951	0.66694	0.68125

Table 10b. Linear regression of poverty depth (Dalton index) on mean length of life: males, age 10+.

Parameter	1970	1980	1990	2000	2010
Slope	-0.00679	-0.00733	-0.00586	-0.00598	-0.00549
Intercept	0.66931	0.71104	0.61271	0.62818	0.59755
R-squared	0.43252	0.70653	0.75217	0.87555	0.90255

Table 11a. Linear regression of poverty severity (Sen index) on mean length of life: females, age 10+.

Parameter	1970	1980	1990	2000	2010
Slope	-0.01023	-0.01009	-0.01047	-0.01070	-0.01082
Intercept	0.88865	0.88615	0.93240	0.95973	0.98590
R-squared	0.88087	0.94346	0.97617	0.98008	0.97873

Table 11b. Linear regression of poverty severity (Sen index) on mean length of life: males, age 10+.

Parameter	1970	1980	1990	2000	2010
Slope	-0.01251	-0.01290	-0.01215	-0.01252	-0.01208
Intercept	1.00993	1.04652	1.00585	1.05304	1.04003
R-squared	0.90656	0.96158	0.97801	0.98823	0.99069

Table 12a. Linear regression of Gini index on mean length of life: females, age 10+.

Parameter	1970	1980	1990	2000	2010
Slope	-0.00221	-0.00227	-0.00240	-0.00276	-0.00286
Intercept	0.26021	0.26559	0.27619	0.30476	0.31407
R-squared	0.25835	0.47329	0.70437	0.84075	0.83241

Table 12b. Linear regression of Gini index on mean length of life: males, age 10+.

Parameter	1970	1980	1990	2000	2010
Slope	-0.00445	-0.00468	-0.00395	-0.00390	-0.00364
Intercept	0.42544	0.44288	0.39465	0.39375	0.37709
R-squared	0.54525	0.77642	0.85083	0.93614	0.93714

5. Between-country versus within-country measures: discussion

Differences, both quantitative and qualitative, between the results on inequality at international and within-country levels are the most obvious findings of the present study. There are two types of such differences. First, an increase in mortality inequality between the countries is accompanied by a decrease in average inequality within the countries. Second, the correlation between the average length of life and inequality is positive at the “macro level” (both indicators increased over the period investigated) while at the “micro level” the countries with higher length of life are generally characterized by lower mortality inequality. Both cases are discussed below.

As mentioned previously, the growth in the mean length of life between 1970 and 2010 was not equal. Although the prevailing part of the countries improved considerably their scores, some others (mainly post-communist ones) experienced relatively small growths, and for some of them even a decline in absolute values was observed (Russia and Ukraine for males). This type of changes obviously explains increases of the overall means and inequality and also of poverty depth under fixed poverty incidence. The decreases in average within-country inequality and poverty may be explained by changes in their mortality patterns. All countries were relatively successful in reducing mortality in the low age groups, which obviously contributed to the reduction of the inequality. Main differences between high and low mortality countries in terms of age-at-death distribution occurred for higher age groups. In the first type of the countries (mainly post-communist ones) serious mortality increases between 1970 and 2010 might be observed for middle age groups (see Billingsley, 2011), especially 40-59 years of age. This type of changes resulted in low increases (or even decreases) in the mean length of life and a decrease in inequality, as those age groups are relatively close to the mean values. In the countries characterized by high length of life the changes in the age-at-death distribution were of different type: usually the distribution functions moved towards the upper tails, while the shape was relative stable. In that case the main factor behind decreases in inequality were the above-mentioned decreases in mortality in lowest age groups.

Another type of difference between the results at the “macro” and “micro” level, i.e. the type of the correlation between length of life and inequality and poverty is consistent with some findings on relations between a health care system and mortality. It is confirmed (Hisnanick and Coddington, 1995, Korda and Butler, 2006) that universal healthcare is generally effective in reducing mortality, especially the so-called avoidable mortality. Hence, it reduces inequality as well as improves the length of life. This type of relation is observed in the present research.

5. Concluding remarks

During the four decades observed (1970 to 2010) the mean ages at death in the countries included in the study were getting more and more apart. This process was much more intensive for males. The main source of those changes was in relative, and in some cases also absolute, deterioration of the position of most of the post-communist countries. Using the income distribution nomenclature one can describe the above-mentioned process as a polarization (see Esteban and Ray, 1994 or Wolfson, 1994), resulting in the emergence of two relatively homogeneous groups with growing distance between them. Divergence in mortality distributions has been explored at two levels. First, a set of inequality and poverty indices has been calculated, taking the countries as the units with the mean length of life as a scalar indicator of development. Second, the within-country distributions using country life tables have been compared by means of the Kullback-Leibler divergence and three poverty indices with the poverty lines set at the median values calculated for all countries, supplemented by within-country inequality measures. These results only partly confirmed the conclusions on growing disparities across the countries. The whole distributions became more dissimilar, although the mean length of life appeared to be the only parameter for which growing divergence could be observed. Both poverty and inequality indices, evaluating selected aspects of mortality distribution, converged over the period investigated. This suggests the growing importance of the mean length of life as a scalar representation of the mortality distribution. Another type of differences between the results in between-country and within-country comparisons refers to the signs of correlation between the average length of life and inequality: it is positive at the “macro level” (both the mean length of life and inequality increased between 1970 and 2010) and negative at the “micro level” (the countries with higher length of life are generally characterized by lower mortality inequality). Poverty country rankings supplemented by the mean length of life rankings have indicated a prevailing part of the post-communist countries as main “losers” of the above-mentioned changes, although a few countries among them improved their initial relative positions. Among the remaining European countries also considerable changes in the ranking, in both directions, could be observed in some cases. The countries that might be said to be the main “winners” are usually non-European ones.

Acknowledgement

The study was supported by the National Science Centre of Poland, grant no. 2012/05/B/HS4/00550 entitled “Measures of differences in mortality distributions”. The author is grateful to two anonymous referees for the comments on the earlier draft of the manuscript. All remaining errors are the responsibility of the author.

REFERENCES

BILLINGSLEY, S., (2011). Exploring the conditions for a mortality crisis: bringing context back into debate. Population, Space and Place 17, pp. 267–289.

COWELL, F., (1977). Measuring Inequality, Philip Allan, Deddington, Oxon.

D'ALBIS, H., ESSO, L. J., PIFARRÉ, H., AROLAS, I., (2014). Persistent differences in mortality patterns across industrialized countries. PLoS ONE 9(9): e106176. doi:10.1371/journal.pone.0106176

EDWARDS, R. D., (2011). Changes in world inequality in length of life: 1970-2000. Population and Development Review 37, pp. 499–528.

EDWARDS, R. D., TULJAPURKAR, S., (2005). Inequality in life spans and a new perspective on mortality convergence across industrialized countries. Population and Development Review 31, pp. 645–674.

ESTEBAN, J., D., RAY, (1994). On the measurement of polarization, Econometrica 62, pp. 819–851.

EUROPEAN COMMISSION, (2009). Portfolio of indicators for the monitoring of the European strategy for social protection and social inclusion – 2009 update. The Eurostat website.

FISHBURN, P. C., (1984). Transfer principles in income distribution. Journal of Public Economics 25, pp. 323–328.

HICKS, N., STREETEN, P., (1979). Indicators of development: the search for a basic yardstick. World Development 7, pp. 567–580.

HISNANICK, J. J., CODDINGTON, D. A., (1995). Measuring human betterment through avoidable mortality: a case for universal health care in the USA. Health Policy 34, pp. 9–19.

HUMAN MORTALITY DATABASE, (2011). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), www.mortality.org.

KORDA, R. J., BUTLER, J. R. G., (2006). Effect of healthcare on mortality: Trends in avoidable mortality in Australia and comparisons with Western Europe. Public Health 120, pp. 95–105.

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

MADDEN, D., (2000). Relative or absolute poverty lines: A new approach, Review of Income and Wealth 46, pp. 181–199.

MUSZYŃSKA, M., SZULC, A., JANSSEN, F., (2014). An index of inequality in age-at-death distributions across a group of countries based on the concept of the Equivalent Length of Life, paper presented at the European Population Conference 2014, 25–28 June, Budapest. http://epc2014.princeton.edu/papers/140389

PARETO, V., (1896). La courbe des revenus. Le Monde Economique, 1896.

RAVALLION, M., (1994). Poverty Comparisons. A Guide to Concepts and Methods, Living Standards Measurement Study, Working Paper, No. 88, The World Bank, Washington, D. C.

SEN, A., (1976). Poverty: an ordinal approach to measurement, Econometrica 44, pp. 219–231.

SEN, A., (1998). Mortality as an indicator of economic success and failure. The Economics Journal 108, pp. 1–25.

SILBER, J., (1983). ELL (the equivalent length of life) or another attempt at measuring development. World Development 11, pp. 21–29.

SILBER, J., (1992). Life tables and the measurement of the extent of poverty, Poverty Measurement for Economies in Transition in Eastern European Countries. International Scientific Conference, Warsaw, 7-9 October 1991, Warsaw 1992, pp. 411–431.

SMITS, J., MONDEN, C., (2009). Length of life inequality around the globe. Social Science and Medicine 68, pp. 1114–1123.

UNITED NATIONS DEVELOPMENT PROGRAMME, (2014). Human Development Reports, http://hdr.undp.org/en.

THEIL, H., (1967). Economics and Information Theory. Chicago: Rand McNally and Company.

WOLFSON, M. C., (1994). When inequalities diverge, The American Economic Review 84, pp. 353–358.

APPENDIX

Table A1. Countries included in the study and the abbreviations

Country	Abbreviation
Australia*	AUS
Austria	AUT
Belgium*	BEL
Belarus	BLR
Bulgaria	BUL
Canada*	CAN
Czech Republic	CZE
Denmark	DEN
Eastern Germany	EGE
England and Wales	ENG
Estonia	EST
Finland*	FIN
France	FRA
Hungary*	HUN
Iceland	ICE
Ireland*	IRE
Italy*	ITA
Japan*	JAP
Latvia	LAT
Lithuania	LIT
Luxembourg*	LUX
Netherlands*	NED
Norway*	NOR
New Zealand*	NZL
Poland*	POL
Portugal*	POR
Russia	RUS
Spain*	SPA
Switzerland	SUI
Slovakia*	SVK
Sweden	SWE
Taiwan	TAI
Ukraine*	UKR
United States	USA
Western Germany	WGE

Asterisks denote countries for which 2009 data instead of 2010 data have been used.

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 497–514

QUALITY OF INSTITUTIONS AND TOTAL FACTOR PRODUCTIVITY IN THE EUROPEAN UNION

Adam P. Balcerzak[1], **Michał Bernard Pietrzak**[2]

ABSTRACT

The key challenge for mid- and long-term policy in the European Union countries is to use the potentials of knowledge-based economy (KBE), which is a condition for maintaining high total factor productivity in Europe. For this reason, the relationship between the quality of an institutional system and total factor productivity in the EU countries has been examined. The quality of the institutional system is defined here from the perspective of incentives that influence the use of the potential of KBE. In order to determine the level of effectiveness of the institutional system in the analysed countries the method for linear ordering of objects was applied based on data from Fraser Institute. The main hypothesis of the article states that the quality of the institutional system in the context of KBE has a significant influence on the level of total factor productivity in the EU. In order to verify this hypothesis, the parameters of the Cobb-Douglas production function were estimated, which allowed the evaluation of TFP for the EU countries. The calculation made in the article based on Eurostat data. In order to identifying the relationship between the quality of the institutional system and the level of TFP a panel model was applied using data from a conducted for years 2000-2010.

Key words: KBE, TFP, quality of intuitions, European Union, panel model.

1. Introduction

The efforts devoted to the research on the determinants of Total Factor Productivity (TFP) growth both at the international (Coe and Helpman, 1995, pp. 859–887, Coe *et al.*, 2008; Aiyar and Dalgaard, 2005, pp. 82–102) and regional level, for example for Poland (Tokarski 2008, pp. 38–53; Tomaszkiewicz and Świeczewska 2011, pp. 36–55), have been significant for the last decade. However, due to many structural changes in the world economy that have been

[1] Nicolaus Copernicus University, Department of Economics. E-mail: adam.balcerzak@umk.pl.
[2] Nicolaus Copernicus University, Department of Econometrics and Statistics.
E-mail: michal.pietrzak@umk.pl.

observed for the last years, the famous postulate of Edward Prescott on the need to strengthen the research intensity on the theory of TFP is still valid (see Presott, 1998, pp. 525–551). This is especially important in the context of a growing role of many intangible and difficult to measure factors, which affect productivity differences between developed countries in a globalized, knowledge-based economy (KBE) (see: van Ark, 2014, pp. 17–19; Fraumeni, 2014, pp. 20–21). This article can be treated as a proposal for research in this field. Thus, the aim of the paper is to analyse TFP in the European Union countries in the years 2000-2010 and to evaluate the influence of the quality of institutions in the context of KBE on the productivity growth in the years 2000–2010.

The first hypothesis of the article was formed as follows: the quality of the institutional system in the context of KBE has a significant influence on the level of TFP in the EU countries.

The second hypothesis of the research concerned the institutional literature that indicated many important weaknesses of the institutional order in former transformation countries of Central Europe and the concept of an institutional lag in the case of these economies, especially in regard to formal regulations and governance influencing the speed of diffusion of technology and new organizational ideas. As a result, the second hypothesis was formed as follows: countries that joined the European Union after 2004, under the conditions of improving the quality of their institutions for KBE, can use the potential of reducing institutional lag for increasing the speed of TFP growth more than proportionally in comparison with "old" member states. This factor is important from the perspective of guidelines for policy that is aimed at increasing the speed of convergence process in the EU. The article is a continuation of previous research presented in Balcerzak and Pietrzak (2015a, pp. 71–91; 2015b, 2016, pp. 312–337).

2. Total factor productivity in the European Union countries

Obtaining high level of productivity and improving the effectiveness of utilization of production factors is considered as an important mid- and long-term aim of economic policy in the European Union (European Commission, 2010; Forgo and Jevcak 2015). Based on the aim and objective of this paper, the analysis of TFP changes for 24 European Union countries for the years 2000–2010 was carried out.

Luxemburg, Malta, Cyprus and Croatia were not included in the research. The first three countries were excluded due to lack of data, additionally Croatia became a member state in 2013. In the analysis the following data was used: total employment (annual averages in thousands of persons), real gross value added (million euro, reference year 2000) and gross fixed capital formation (million euro, reference year 2000). Eurostat was the source of the data.

The starting point of the research was the assessment of the productivity level for all the countries in the years 2000-2010 based on the Cobb-Douglas production function. The Cobb-Douglas production function, after taking the logarithm of both sides of equation into account, can be written as follows:

$$\ln GVA_{it} = \boldsymbol{\eta}_i + gt + \alpha \ln GCF_{it} + (1-\alpha)\ln E_{it} + \boldsymbol{\varepsilon}_{it} \tag{1}$$

where:

GVA_{it} – vector of real gross value added in country i and period t,
$GFCF_{it}$ – vector of gross fixed capital formation in country i and period t,
E_{it} – vector of employment in country i and period t,
$\boldsymbol{\eta}_i$ – vector of values of individual effects that determine the average value of total factor productivity, in period t,
t – time trend,
α – elasticity of labour productivity to the capital to labour ratio,
g – rate of technological progress in the sense of Hicks,
$\boldsymbol{\varepsilon}_{it}$ – vector of disturbances.

After subtracting the expression ln(E) from both sides of equation (1), equation (2) is obtained. It describes the level of labour productivity relative to the capital to labour ratio.

$$\ln GVA / E_{it} = \boldsymbol{\eta}_i + gt + \alpha \ln GCF / E_{it} + \boldsymbol{\varepsilon}_{it} \tag{2}$$

where:

GVA/E – vector of value GVA/E – labour productivity,
GFCF/E – vector of the capital to labour ratio,
The remaining variables are understood in the same way as in the case of equation 1.

In the literature, one can find many empirical approaches to evaluating TFP (see. Welfe (ed.) 2007; Severgnini and Burda, 2010, pp. 447–466; Gehringer *et al.*, 2014). In the article, the method proposed by Tokarksi was applied. Estimation of the parameters of model (2) for labour productivity enables the determination of the value of total factor productivity TFP_{it} for the EU countries. To calculate TFP_{it} the estimated value of parameter α is used. It can be done based on the formula (see Tokarski, 2008, pp. 39–53):

$$TFP_{it} = \frac{GVA/E_{it}}{\left(GFCF/E_{it}\right)^{\hat{\alpha}}}\ . \tag{3}$$

The general assumption commonly used in many proposals for estimation of TFP at the national or regional level is the application of homogeneous production functions for all the countries or regions. However, in spite of the convergence process the member countries are still characterized by significant development differences in the case of the European Union. There are relatively big differences in labour productivity and the factors that can influence TFP. As a result, the analysed countries are heterogeneous. The most obvious structural differences can be seen between the so-called "old" and "new" member states. Thus, the assumption on homogeneous production functions is unrealistic here. In order to face the problem of heterogeneity, the authors divided the EU countries into two groups – the "old" member states and the ones that joined the EU after the year 2004. As a result, separate parameters α_1 and α_2 for „old" and „new" member states were introduced in the case of the model for labour productivity (equation 2). They can be written as follows[3]:

$$\ln GVA/E_{it} = \boldsymbol{\eta}_i + gt + \alpha_1 \ln GCF/E_{it}^1 + \alpha_2 \ln GCF/E_{it}^2 + \boldsymbol{\varepsilon}_{it}\,, \tag{4}$$

where:

α_1 and α_2 – elasticity of labour productivity to the capital to labour ratio for the group of "old" and "new" member states, respectively.

The remaining variables are understood in the same way as in the case of equation 1 and 2.

Additionally, the model given in equation 4 was estimated with a FE panel model estimator with individual effects in order to observe the country-specific factors. The results of the estimation of the parameters of the model can be found in Table 1. Individual effects for all 24 countries were statistically significant, which is consistent with previous argumentation. The estimations of parameters α_1, α_2 and g were statistically significant too[4]. The value of estimates of the parameter α_1 and α_2 indicates that the flexibility of labour productivity to capital to labour ratio equals 0.129 in the case of the countries that joined the EU before 2004 and 0.290 for "new" members. The value of the estimate of the parameter g at the level of 0.017 indicates that the European Union economies are characterized by 2% rate of technological progress in the sense of Hicks. These results are generally consistent with previous research both at the national (Gehringer *et al.* 2014; Severgnini and Burda, 2010, pp. 447–466) and regional

[3] The components of GCF/E_{it}^1 vector make the value of capital to labor ratio for the countries in the first group and 0 otherwise. A similar situation occurs in the case of GCF/E_{it}^2 vector.

[4] The statistics for Durbin-Watson test points to statistically significant autocorrelation in the residuals.

level in the case of Poland (Dańska-Borsiak and Laskowska, 2012, pp. 17–29; Tokarski, 2008, pp. 38–53; Tokarski, 2010, pp. 23–39).

Table 1. The results of estimation of parameters of FE panel model with individual effects for labour productivity

Parameter	Estimate	Standard error	Parameter	Estimate	Standard error
α_1	0.129	0.034	g	0.017	0.003
α_2	0.290	0.050	-	-	-
Individual effects	**Estimate**	**Standard error**	**Individual effects**	**Estimate**	**Standard error**
Austria	3.550	0.096	Ireland	3,640	0,098
Belgium	3.610	0.097	Italy	3.497	0.091
Bulgaria	1.436	0.021	Lithuania	1.982	0.049
Czech Republic	2.138	0.072	Latvia	1.907	0.059
Germany	3.550	0.091	Netherlands	3.480	0.090
Denmark	3.595	0.096	Poland	2.142	0.056
Estonia	1.930	0.071	Portugal	2.813	0.070
Spain	3.172	0.089	Romania	1.405	0.020
Finland	3.551	0.093	Sweden	3.667	0.095
France	3.597	0.093	Slovenia	2.508	0.091
Greece	3.122	0.080	Slovak Republic	1.972	0.056
Hungary	2.089	0.060	United Kingdom	3.626	0.089
Coefficient of determination		0.99	Durbin-Watson Statistics		0.42

[all the parameters are statistically significant with 5% significance level]

Source: own estimation based on Eurostat data.

The estimated value of the parameter α₁ and α₁ enable the estimation of the logarithm of TFP_{it} for every country in the years 2000-2010 according to the formula:

$$\ln TFP_{it} = \ln GVA/E_{it} - \hat{\alpha}_j \ln GCF/E_{itj}. \tag{4}$$

Table 2 presents the logarithm of TFP for the years 2000 and 2010 and the percentage change of this value in the period 2000–2010. Additionally, Table 2 shows the results of grouping the countries into classes, which was done with the use of natural breaks method. The results presented in Figure 1 confirm the heterogeneity between “old” and “new” member states.

Table 2. Total factor productivity in the European Union member countries in 2000 and 2010 (lnTFP)

2000			2010			2000-2010		
Country	**lnTFP**	**Class**	**Country**	**lnTFP**	**Class**	**Country**	**Percentage difference**	**Class**
Sweden	3.70	5	Ireland	3.86	5	Romania	38.71%	5
France	3.69	5	Sweden	3.82	5	Lithuania	23.47%	4
Denmark	3.68	5	United Kingdom	3.76	5	Latvia	22.24%	4
Ireland	3.68	5	Denmark	3.72	4	Bulgaria	17.56%	4
Belgium	3.67	5	Belgium	3.72	4	Slovak Republic	16.76%	4
United Kingdom	3.66	5	France	3.70	4	Estonia	14.49%	3
Italy	3.61	5	Finland	3.68	4	Czech Republic	11.95%	3
Germany	3.60	5	Austria	3.67	4	Poland	10.87%	3
Austria	3.59	5	Germany	3.66	4	Hungary	9.65%	3
Finland	3.58	5	Netherlands	3.64	4	Slovenia	8.11%	3
Netherlands	3.54	5	Italy	3.60	4	Portugal	5.55%	3
Spain	3.29	4	Spain	3.34	3	Greece	5.37%	2
Greece	3.13	4	Greece	3.30	3	Ireland	4.93%	2
Portugal	2.85	3	Slovenia	3.02	3	Sweden	3.15%	2
Slovenia	2.80	3	Portugal	3.01	3	Finland	2.89%	2
Czech Republic	2.34	2	Czech Republic	2.61	2	United Kingdom	2.75%	2
Poland	2.30	2	Poland	2.55	2	Netherlands	2.63%	2
Hungary	2.26	2	Hungary	2.48	2	Austria	2.29%	2
Estonia	2.10	2	Lithuania	2.45	2	Germany	1.52%	2
Slovak Republic	2.09	2	Slovak Republic	2.44	2	Denmark	1.28%	1
Lithuania	1.98	2	Estonia	2.40	2	Spain	1.26%	1
Latvia	1.96	2	Latvia	2.40	2	Belgium	1.17%	1
Bulgaria	1.49	1	Romania	1.75	1	France	0.34%	1
Romania	1.26	1	Bulgaria	1.75	1	Italy	-0.29%	1

Source: own estimation based on Eurostat data.

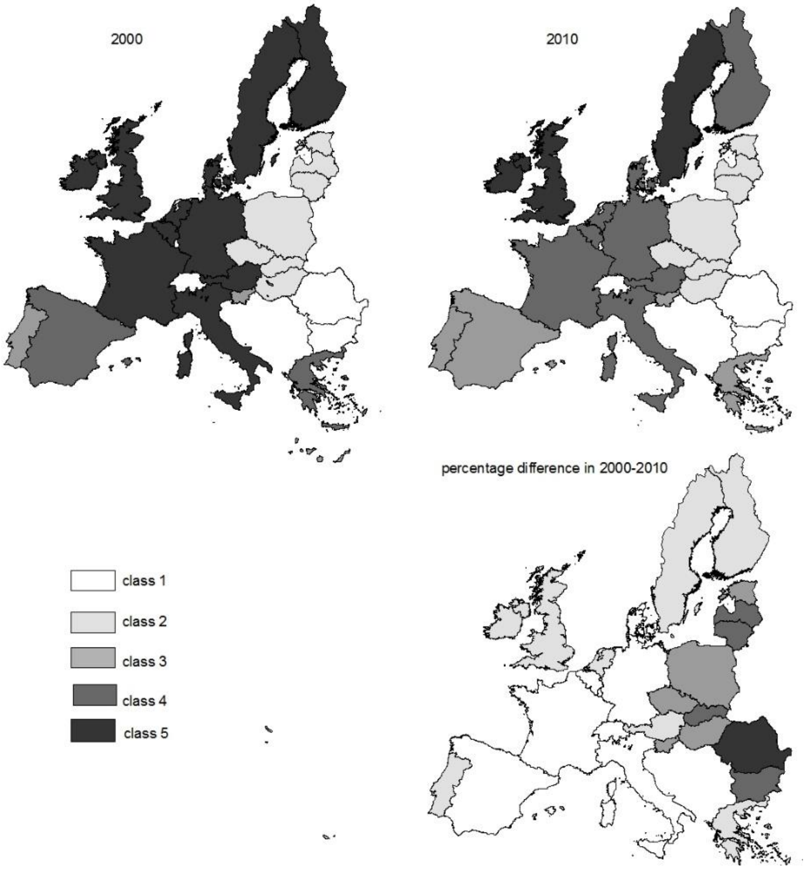


Figure 1. Total factor productivity in the European Union member countries in 2000 and 2010 (lnTFP)

Source: own estimation.

3. Quality of institutions in the context of knowledge-based economy as a determinant of total factor productivity

The mid- and long-term growth potential of developed countries is currently dependent on the ability to use the potential of KBE (Welfe (ed.), 2007; Balcerzak, 2009b, pp. 711–739, OECD, 1995; Ciborowski, 2014, pp. 57–72, Wronowska 2013, pp. 71–80). In the case of developed economies, empirical research, which has been carried out for the last two decades, confirmed the influence of institutional conditions affecting transaction costs of technological changes on the number of enterprises, which are able to use new ideas and knowledge effectively and to achieve further technological breakthroughs (OECD, 2001; McKinsey Global Institute 2002). Thus, the quality of institutions in the context of KBE should be a significant factor influencing total factor productivity in the case of developed countries. The verification of the influence of this factor is the main objective of this analysis. Its confirmation, from the point of view of policy guidelines, means that in the reality of KBE the creation of high quality institutions and their constant improvement should be treated as an essential condition for maintaining the high rate of productivity growth[5].

The analysis of empirical research in the context of the theory of new institutional economics enables one to indicate four fundamental segments of institutional systems, which on the one hand can be modified by governments in relatively short time, and which on the other hand have significant influence on the speed of technological change[6]. Additionally, based on the arguments of new institutional economics, high quality institutions are defined here as the ones that tend to lower the transaction costs of technological progress and diffusion of new organizational ideas.

The first segment of the institutional system is the effectiveness of legislation influencing entrepreneurship. High level of entrepreneurship is conducive to increasing the supply of companies with high growth potential, and increases the likelihood of the emergence of new innovative start-ups.

The second institutional segment relates to the effectiveness of juridical system in keeping the low level of transaction costs and supporting effectiveness of market mechanism. Formal regulation that reduces the level of transaction costs favours the elimination of formal barriers to the diffusion of new organizational and technological solutions in the economy.

The third segment of the institutional system is the competitive pressure and effectiveness of labour markets. The high level of competitive pressure under conditions of relatively effective labour markets creates incentives for reorganization activities, which is conducive to improving microeconomic efficiency of enterprises. It increases the potential of enterprises that are able to find and implement new technological and organizational solutions.

The fourth institutional segment refers to financial market institutions, which should act as a stimulator of development of enterprises with high growth potential. The financial markets should support a faster reallocation of capital from industries with low to new sectors with high growth potential.

These four instructional segments can be treated as the incentive pillar of the concept of pillars of KBE according to the World Bank (see Chen and Dahlman 2005, 2004, Madrak-Grochowska 2015, 7–21).

For the identified key institutional segments, the authors selected a set of diagnostic variables, which are presented in Table 3. Detailed data for all the variables were obtained from the database of Fraser Institute[7]. Due to the design

[5] From the institutional perspective the analysis proposed in the research concentrates on the institutions that can be influenced by policy action in relatively short or medium term (Williamson, 2000, pp. 595–613; North, 1994, pp. 359–368). The influence of institutions that are the result of long-term evolutionary process is not the subject or the analysis.

[6] A more detailed discussion on the research which gave the theoretical and empirical background for highlighting these segments of institutional systems as important elements of institutional matrix influencing the possibility of utilization of the potential of, KBE is available in Balcerzak, Pietrzak (2016, 2015, pp. 71–91), Balcerzak (2015b, pp. 51–63).

[7] http://www.freetheworld.com/reports.html (1.10.2014).

of the database all diagnostic variables were stimulants with the values from 0 to 10. It should be emphasized that the variables presented in Table 3 enable one to quantify the quality of the segments of the institutional system only, which are essential in the context of a country's ability to exploit the potential of KBE. This research should not be interpreted as a proposal for holistic quantification of all segments of institutional matrix influencing economic activity and welfare in the analysed countries (see Gruszewska, 2011, pp. 103–120).

Table 3. The potential variables concerning quality of institutions from the perspective of KBE potential

Y_1 – formal regulations influencing entrepreneurship
X_{1t}^1 – Administrative requirements for entrepreneurs
X_{2t}^1 – Bureaucracy costs for entrepreneurs
X_{3t}^1 – The cost of starting business
X_{4t}^1– Extra payments/bribes/favouritism
Y_2 – effectiveness of juridical system in keeping low level of transaction costs and supporting effectiveness of market mechanism
X_{1t}^2 – Judicial independence
X_{2t}^2 – Impartial courts
X_{3t}^2 – Protection of property rights
X_{4t}^2 – Integrity of the legal system
Y_3 – competitive pressure and effectiveness of labour markets
X_{1t}^3 – Revenue from trade taxes (% of trade sector)
X_{2t}^3 – Mean tariff rate
X_{3t}^3 – Standard deviation of tariff rates
X_{4t}^3 – Non-tariff trade barriers
X_{5t}^3 – Compliance costs of importing and exporting
X_{6t}^3 – Regulatory trade barriers
X_{7t}^3 – Foreign ownership/investment restrictions
X_{8t}^3 – Capital controls
X_{9t}^3 – Controls of the movement of capital and people
X_{10t}^3 – Hiring regulations and minimum wage
X_{11t}^3 – Hiring and firing regulations
X_{12t}^3 – Centralized collective bargaining
Y_4 – financial markets institutions as a stimulator of development of enterprises with high growth potential
X_{1t}^4 – Private sector credit
X_{2t}^4 – Interest rate controls/negative real interest rates

Source: own work based on the discussion presented in Balcerzak (2015b, pp. 51–63, 2009a, pp. 71–106, 2009b, pp. 711–739), Balcerzak and Pietrzak (2016), Balcerzak and Rogalska (2008, pp. 71–87).

At the next stage, the ability of the variables to differentiate the objects was verified. Then, based on the diagnostic variables describing the discussed four segments of an instructional system a taxonomic measure of development (TMR_{it}) was calculated. The TMR measure enables the evaluation of the quality of institutions for 24 EU countries for the years 2000-2010. The applied method of taxonomic measure of development was proposed by Zdzisław Hellwig 1968 (1968, pp. 307-327; 1972, pp. 131-134). It is based on the comparison of the distance of the object from a pattern of economic development. The application of the method enables one to order the objects and divide them into homogenous classes. The value of taxonomic measure of development is influenced by many variables describing different elements of a multivariate phenomenon, thus it enables to measure it synthetically.

The value of taxonomic measure of development (TMR_{it}) was evaluated in two stages. At the first stage, after normalization of the values of the variable with classic normalization formula, the values of TMR_{it}^{k} for every institutional segment showed in Table 1 were calculated based on Hellwig's method. In the case of every variable the pattern of economic development was set as a maximum value for the years 2000-2010. As a result a fixed pattern of development was used here, which enabled a dynamic comparison of the final results in the whole period[8]. The values of four measures for every institutional segment were calculated: TMR_{it}^{1} describing formal regulations influencing entrepreneurship, TMR_{it}^{2} measuring the effectiveness of juridical system in keeping low level of transaction costs and supporting effectiveness of market mechanism, TMR_{it}^{3} for competitive pressure and effectiveness of labour markets, and TMR_{it}^{4} for financial markets institutions as a stimulator of development of enterprises with high growth potential.

At the second stage, an arithmetic mean for all the four measures TMR_{it}^{k} was calculated according to the following formula:

$$TMR_{it} = \sum_{k=1}^{4} TMR_{it}^{k} / 4, \tag{5}$$

where:

i – index for the object (country),
t – index for time.

Based on the values of TMR_{it} the European Union countries were grouped to one of five classes. As in the case of TFP, it was done with the application of natural breaks method. Some sets of countries, which are relatively homogenous from the perspective of the quality of intuitions in the context of KBE, were obtained. The results for the year 2000 and 2010 are presented in Table 4 and Figure 2.

[8] The detailed formal description of the applied procedure is available in Balcerzak (2011, pp. 456–467).

Table 4. The values of TMR for the quality of institutions in the EU countries in the year 2000 and 2010

2000			2010			2000-2010		
Country	**TMR**	**Class**	**Country**	**TMR**	**Class**	**Country**	**Percentage difference**	**Class**
United Kingdom	0.78	5	Denmark	0.81	5	Romania	65.25%	5
Netherlands	0.78	5	Finland	0.78	5	Bulgaria	23.51%	4
Finland	0.77	5	Sweden	0.76	5	Slovak Republic	22.08%	4
Denmark	0.76	5	Estonia	0.72	5	Estonia	18.61%	4
Belgium	0.72	4	Netherlands	0.71	5	Poland	10.91%	3
Sweden	0.71	4	United Kingdom	0.69	4	Latvia	9.88%	3
Germany	0.70	4	Belgium	0.66	4	Sweden	8.03%	3
Ireland	0.69	4	Ireland	0.64	4	Lithuania	6.92%	3
Austria	0.68	4	Austria	0.64	4	Denmark	6.49%	3
France	0.64	3	France	0.62	4	Hungary	5.89%	3
Spain	0.63	3	Germany	0.59	3	Czech Republic	2.59%	3
Estonia	0.60	3	Hungary	0.56	3	Finland	1.71%	3
Portugal	0.55	2	Latvia	0.53	3	France	-2.96%	2
Italy	0.54	2	Spain	0.53	3	Austria	-6.65%	2
Hungary	0.53	2	Romania	0.52	2	Ireland	-7.46%	2
Slovenia	0.51	2	Czech Republic	0.52	2	Italy	-8.50%	2
Czech Republic	0.50	2	Slovak Republic	0.51	2	Netherlands	-8.61%	2
Latvia	0.48	2	Bulgaria	0.50	2	Belgium	-9.22%	2
Lithuania	0.46	2	Italy	0.49	2	Greece	-9.43%	2
Poland	0.42	1	Lithuania	0.49	2	Slovenia	-9.70%	2
Slovak Republic	0.42	1	Portugal	0.48	2	United Kingdom	-11.85%	1
Greece	0.42	1	Poland	0.46	2	Portugal	-11.87%	1
Bulgaria	0.41	1	Slovenia	0.46	2	Spain	-15.16%	1
Romania	0.31	1	Greece	0.38	1	Germany	-15.53%	1

Source: own estimation based on the data from Fraser Institute.

As it could be seen in the case of TFP presented in Table 1 and Figure 2, the results presented in Table 4 and Figure 2 confirm analogous heterogeneity between the EU countries in the case of the quality of institutions for KBE. The "old" member states are generally grouped in classes 5 and 4, whereas the "new" member states can be found in classes from 3 to 1 with the exception of Estonia in 2010.

The highest values of TMR for the quality of institutions in the context of KBE were obtained by Scandinavian countries grouped in class 5, followed by

Austria, France, Germany and Spain grouped in classes 4 and 3. The southern European countries: Portugal, Italy and Greece are characterized by a lower quality of institutions for KBE.

Central European "new" member states are grouped in classes 3 to 1, with the lowest values of TMR for Poland, Bulgaria and Romania. As a result, the biggest improvement in the sphere of the quality of institutions was obtained by the countries that joined the EU after 2004 (especially Romania, Bulgaria, Slovakia and Estonia), which was due to "the benefits" of institutional lag and the institutional convergence process in the analysed period (see more Balcerzak, 2011, pp. 17–34).

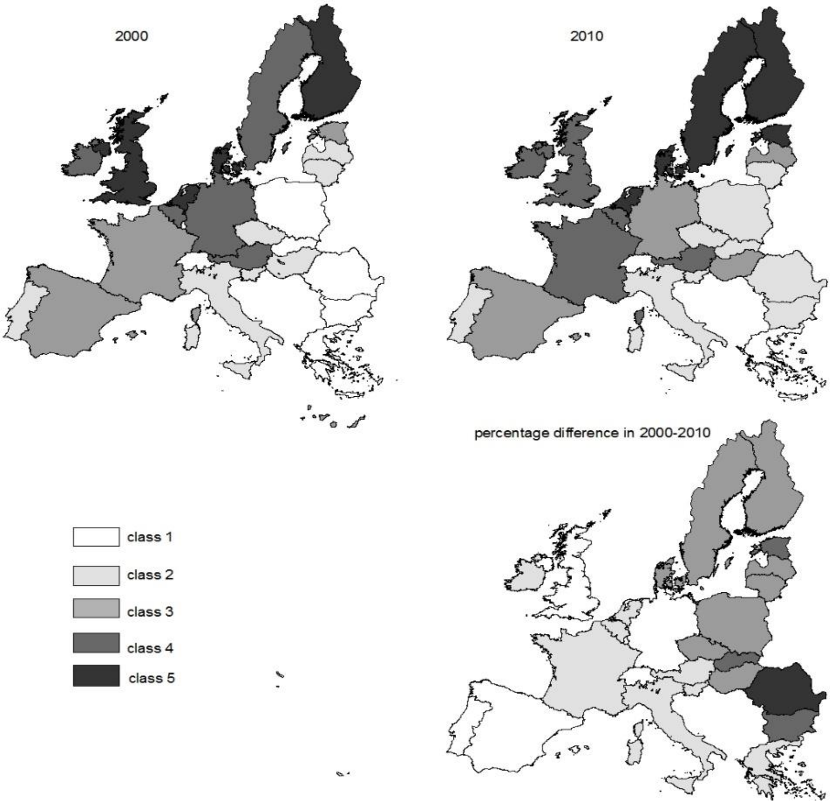


Figure 2. Quality of institutions for KBE in the EU countries in the year 2000 and 2010

Source: own estimation based on the data from Fraser Institute.

The calculated values of TMR were used at the next stage of the research, where the impact of the level of the quality of institutions on TFP was examined.

For this purpose, a specification of a FE panel model with individual effects was drawn up. The model was written with the following equation[9]:

$$\ln FTP_{it} = \boldsymbol{\eta}_i + gt + \beta_1 TMR_{it}^1 + \beta_2 TMR_{it}^2 + \boldsymbol{\varepsilon}_{it} \,, \tag{6}$$

where due to the heterogeneity of the analysed countries separate parameters β$_1$ and β$_2$ were used for two groups of economies; the dependent variable is logarithm of TFP while logarithm of TMR_{it}^1 and TMR_{it}^2 serve as independent variables.

Table 5. The results of estimation of parameters of FE panel model with individual effects for determinants of TFP

Parameter	Estimate	Standard error	Parameter	Estimate	Standard error
β$_1$	0.379	0.229	g	0.018	0.003
β$_2$	0.898	0.280	-	-	-
Individual effects	**Estimate**	**Standard error**	**Individual effects**	**Estimate**	**Standard error**
Austria	3.289	0.169	Ireland	3.374	0.172
Belgium	3.362	0.161	Italy	3.306	0.127
Bulgaria	1.070	0.127	Lithuania	1.697	0.134
Czech Republic	1.911	0.141	Latvia	1.612	0.148
Germany	3.317	0.152	Netherlands	3.204	0.178
Denmark	3.294	0.193	Poland	1.943	0.115
Estonia	1.568	0.181	Portugal	2.617	0.129
Spain	2.957	0.141	Romania	1.075	0.114
Finland	3.259	0.187	Sweden	3.389	0.179
France	3.363	0.152	Slovenia	2.349	0.139
Greece	2.957	0.111	Slovak Republic	1.709	0.135
Hungary	1.813	0.143	United Kingdom	3.352	0.176
Coefficient of determination		0.99	Durbin-Watson Statistics		0.55

[all the parameters are statistically significant at 5% significance level]

Source: own calculations based on Eurostat data.

The results of the estimation of the parameters of a panel model (6) are presented in Table 5. The parameters for individual effects for all the countries, the parameters β$_1$ and β$_2$, were statistically significant[10]. Positive values of the

[9] The components of TMR_{it}^1 vector consist of the values of *TMR* for the first group of „old" member states and 0 otherwise. A similar situation occurs in the case of TMR_{it}^2 vector and the second group of "new" member states.

[10] The statistics for Durbin-Watson test indicates statistically significant autocorrelation in the residuals.

estimation of the parameters β_1 and β_2 at the level 0.379 and 0.898 confirm the significant influence of the quality of institutions in the context of KBE on the level of TFP in the case of both groups of “old” and “new” member states. This allows the verification of the first hypothesis of the research. Additionally, a higher value of parameter β_2 for “new” member states can indicate that in the case of effective institutional policy and reforms of regulations, which will lead to a significant improvement in the quality of institutions in the context of KBE, the “new” member states would be able to improve their TFP more than proportionally in relation to “old” member states. This factor can become a significant contributor in the process of reducing development differences and supporting the convergence process of the European Union countries. This result is consistent with the previous analysis of the influence of the quality of institutions for KBE on the convergence process in Europe, which was done by the authors within conditional β-convergence framework (Balcerzak and Pietrzak 2015b). Thus, the outcome of the analysis enable one to verify the second hypothesis of the research.

4. Theoretical reference and policy implications

From the theoretical perspective, the presented results are consistent with the argumentation of new institutional economics in the context of evolutionary research on the determinants of technological changes. Additionally, the formal quantitative methodology applied in the research can be a complementary proposal to the qualitative approach, which dominates in the case of institutional framework.

From the policy perspective, the results of the research highlight the importance of institutional reforms in the European Union. The modifications of formal regulations that are up to the requirements of KBE would improve the productivity of the European countries and the European economy as a whole. It is consistent with the discussion concerning the implementation of Europe 2020 strategy (see Hobza and Mourre, 2010, Denis *et al.* 2005, Balcerzak 2015a, pp. 190–210). In the case of the “new” member states the institutional reforms are essential for increasing the speed of catching up with the “old” members. From the perspective of the common European market, they are important for keeping and eventually improving the European competitive position in the global economy.

5. Conclusions

The article concentrates on the issue of the evaluation of TFP changes in the EU countries in the years 2000-2010. Special attention was given here to the influence of the quality of intuitions on TFP. In the first part, TFP at the national level for the EU was evaluated. Then, a method for quantifying the quality of institutions in the context of KBE was proposed. From the institutional perspective, the presented approach was rooted in the transaction cost theory. Form the numerical point of view, it was based on Hellwig's concept of the pattern of economic development. Referring to the research on the determinants of productivity changes in OECD countries, the authors proposed four segments of an national institutional system that are important form the perspective of the utilization of the potential of technological changes and KBE as a whole. Based on these four segments the total measure of development for the quality of institutions in the EU countries was calculated.

Then, the influence of the quality of institutions for KBE on TFP in the EU countries was estimated with the application of a panel model. The results of the econometric analysis enabled the verification of both hypotheses of the article, the first one concerning the positive influence of the quality of institutions on TFP in the EU countries, and the second one, which indicated especially high potential in the case of the "new" member states in improving their TFP under the condition of effective institutional reforms.

REFERENCES

AIYAR, S., DALGAARD, C.-J., (2005). Total Factor Productivity Revisited: A Dual Approach to Development Accounting, IMF Staff Papers Vol. 52, No. 1, pp. 82–102.

BALCERZAK, A. P., (2009a). Wpływ działalności regulacyjnej państwa w obszarze kreowania ładu konkurencyjnego na rozwój nowej gospodarki [The impact of the regulatory activities of the state in the sphere of creation of competitive order on the new economy development]. In: Aktywność regulacyjna państwa a potencjał rozwojowy gospodarki. eds. Adam P. Balcerzak and Michał Moszyński, Polskie Towarzystwo Ekonomiczne Oddział w Toruniu, Toruń, pp. 71–106.

BALCERZAK, A. P., (2009b). Efektywność systemu instytucjonalnego a potencjał gospodarki opartej na wiedzy [Effectiveness of the Institutional System Related to the Potential of the Knowledge-Based Economy], Ekonomista, No. 6, pp. 711–739.

BALCERZAK, A. P., (2009c). Structure of Financial Systems and Development of Innovative Enterprises with High Grow Potential, In: M. Piotrowska, L. Kurowski ed., Global Challenges and Politics of the European Union – Consequences for the "New Member States", Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 59, Wrocław.

BALCERZAK, A. P., (2011). Integracja instytucjonalna w krajach Unii Europejskiej. Propozycja Pomiaru [Institutional integration in the European Union. Suggestion for measurement], Ekonomia i Prawo, Vol. 7, No. 1, pp. 17–34.

BALCERZAK, A .P., (2011). Taksonomiczna analiza jakości kapitału ludzkiego w Unii Europejskiej w latach 2002-2008 [Taxonomic Analysis of the Quality of Human Capital in the European Union in 2002-2008], Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu Taksonomia 18 Klasyfikacja I analiza danych – teoria i zastosowania, pp. 456–467.

BALCERZAK, A. P., (2015a). Europe 2020 Strategy and Structural Diversity Between Old and New Member States. Application of zero unitarization method for dynamic analysis in the years 2004–2013, Economics & Sociology, Vol. 8, No. 2, pp. 190–210.

BALCERZAK, A. P., (2015b). Wielowymiarowa analiza efektywności instytucjonalnej w krajach Europy Środkowo-Wschodniej w relacji do standardów OECD [Multivariate Analysis of Institutional Effectiveness in Central European Countries in Relation to OECD Standards], Optimum. Studia Ekonomiczne, Vol. 1(74), pp. 51–63.

BALCERZAK, A. P., PIETRZAK, M. B., (2015a). Wpływ efektywności instytucji na jakość życia w Unii Europejskiej. Badanie panelowe dla lat 2004-2010 [The Impact of the Efficiency of the Institutions on the Quality of Life in the European Union. Panel Data Evidence for the Years 2004-2010], Przegląd Statystyczny, 2015, Vol. 62, No. 1, pp. 71–91.

BALCERZAK, A. P., PIETRZAK, M. B., (2015b). Quality of institutional systems for global knowledge-based economy and convergence process in the European Union, Ekonomia. Rynek, Gospodarka, Społeczeństwo, 2015, No. 42, pp. 9–22.

BALCERZAK, A. P., PIETRZAK, M. B., (2016). Efektywność instytucjonalna krajów Unii Europejskiej [Institutional effectiveness of the European Union countries] , Ekonomista, No. 3, pp. 312–337.

BALCERZAK, A. P., ROGALSKA, E., (2008). Ochrona praw własności intelektualnej w warunkach nowej gospodarki [Protection of copyright in the conditions of the new economy], Ekonomia i Prawo, No.4, pp. 71–87.

CHEN, D. H. C., DAHLMAN, C. J., (2004). Knowledge and Development: A Cross-Section Approach. World Bank Policy Research Working Paper No. 3366.

CHEN, D. H. C., DAHLMAN, C. J., (2004). The Knowledge Economy, the KAM Methodology and World Bank Operations. World Bank Institute Working Paper No. 37256.

CIBOROWSKI, R., (2014). Innovation Process Adjusting in Peripheral Regions. The Case of Podlaskie Voivodship, Equilibrium. Quarterly Journal of Economics and Economic Policy, Vol. 9, No. 2, pp. 57–72.

COE, D. T., HELPMAN, E., (1995). International R&D spillovers, European Economic Review, Vol. 39, No. 5, pp. 859–887.

COE, D. T., HELPMAN, E., HOFFMAISTER, A. W., (2008). International R&D Spillovers and Institutions, IMF Working Paper, WP/08/104, April.

DAŃSKA-BORSIAK, B., LASKOWSKA, I., (2012). The Determinants of Total Factor Productivity in Polish Subregions. Panel Data Analysis, Comparative Economic Research. Central and Eastern Europe, Vol.15, No. 4, pp. 17–29.

DENIS, C., MORROW, K., MC, ROGER, W., VEUGELERS, R., (2005). The Lisbon Strategy and the EU's structural productivity problem, European Economy, Directorate-General for Economic and Financial Affairs, Economic Papers, N° 221 February.

EUROPEAN COMMISSION, (2010). Europe 2020 A strategy for smart, sustainable and inclusive growth, Communication from the commission, Brussels, 3.3.2010 COM(2010) 2020.

FORGO, B., JEVCAK, A., (2015). Economic Convergence of Central and Eastern European EU Member States over the Last Decade (2004-2014), European Commission, European Economy Discussion Papers, DISCUSSION PAPER 001, JULY 2015.

FRAUMENI, B. M., (2014). Frontiers and Opportunities in Productivity Research, International Productivity Monitor, No. 27, pp. 20–21.

GEHRINGER, A., MARTINEZ-ZARZOSO, I., DANZINGER, F. N.-L., (2014). TFP Estimation and Productivity Drivers in the European Union, Center for European, Governance and Economic Development Research, Discussion Papers, Number 189 – February.

GRUSZEWSKA, E., (2011). Matryca instytucjonalna a innowacyjność [An Institutional Matix and Innovation], Optimum. Studia Ekonomiczne, Vol. 2(50), pp. 103–120.

HELLWIG, Z., (1968). Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr [Procedure of Evaluating High Level Manpower Data and Typology of Countries by Means of the Taxonomic Method], Przegląd Statystyczny, No. 4, pp. 307–327.

HELLWIG, Z., (1972). Procedure of Evaluating High-Level Manpower Data and Typology of Countries by Means of the Taxonomic Method, In: Z. Gostkowski (ed.), Towards a System of Human Resources Indicators for Less Developed Countries. Papers prepared for a UNESCO Research Project, Ossolineum, The Polish Academy of Sciences, Wrocław, pp. 131–134.

HOBZA, A., MOURRE, G., (2010). Quantifying the potential macroeconomic effects of the Europe 2020 strategy: stylised scenarios, European Economy, Economic Papers 424, September.

MADRAK-GROCHOWSKA, M., (2015). The Knowledge-based Economy as a Stage in the Development of the Economy. Oeconomia Copernicana, Vol. 6, No. 2, pp. 7–21.

PRESCOTT, E. C., (2014). Needed: a Theory of Total Factor Productivity, International Economic Review, Vol. 39, No. 3, pp. 525–551.

SEVERGNINI, B., BURDA, M. C., (2010). TFP Growth in Old and New Europe, Comparative Economic Studies, Vol. 51, No. 4, 2010, pp. 447–466.

TOKARSKI, T., (2008). Oszacowania regionalnych funkcji produkcji [Estimations of Regional Production Functions], Wiadomości Statystyczne, No. 10, pp. 38–53.

TOKARSKI, T., (2010). Przestrzenne zróżnicowanie łącznej produkcyjności czynników produkcji w Polsce [The Regional Diversification of Total Factor Productivity in Poland]. Gospodarka Narodowa. Vol. 3, pp. 24–39.

TOMASZEWICZ, Ł., ŚWIECZEWSKA, I., (2011). Czynniki wzrostu efektywności sektorów polskiej gospodarki [Factors Determining Growth in Efficiency of Different Sectors of Polish Economy], Optimum. Studia Ekonomiczne, 2(50), pp. 36–55.

VAN ARK, B., (2014). Priorities and Directions for Future Productivity Research: The Need for Historical Perspective, International Productivity Monitor, No. 27, pp. 17–19.

WELFE, W., (ed.) (2007). Gospodarka oparta na wiedzy [Knowledge-based Economy], PWE, Warszawa.

WRONOWSKA, G., (2013). Innovations and Sustainable Development Outline of Problems, Equilibrium. Quarterly Journal of Economics and Economic Policy, Vol. 8, No. 1, pp. 71–80.

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 515–524

LOCALLY REGULARIZED LINEAR REGRESSION IN THE VALUATION OF REAL ESTATE

Mariusz Kubus [1]

ABSTRACT

Regression methods are used for the valuation of real estate in the comparative approach. The basis for the valuation is a data set of similar properties, for which sales transactions were concluded within a short period of time. Large and standardized databases, which meet the requirements of the Polish Financial Supervision Authority, are created in Poland and used by the banks involved in mortgage lending, for example.

We assume that in the case of large data sets of transactions, it is more advantageous to build local regression models than a global model. Additionally, we propose a local feature selection via regularization. The empirical research carried out on three data sets from real estate market confirmed the effectiveness of this approach. We paid special attention to the model quality assessment using cross-validation for estimation of the residual standard error.

Key words: large transactional data, local regression, feature selection, regularization, cross-validation.

1. Introduction

Objective and highly detailed valuation of real estate supports business decisions and the risk associated with these decisions. It affects the proper functioning of lending activities of the banks involved in mortgage financing of real estate. In newly amended Recommendation J, the Polish Financial Supervision Authority has stipulated the rules of collecting and processing data of real estate by banks. The recommendation pays special attention to the use of a reliable and standardized database. There are such databases in Poland, which are created and offered for commercial use. They include transaction prices, locations and property characteristics, which are obtained from several sources: notarial deeds, county authorities, valuations and vision of local experts. An example of this is the system for Analysis and Monitoring of Real Estate Market (AMRON),

[1] Department of Mathematics and Applied Computer Science, Opole University of Technology. E-mail: m.kubus@po.opole.pl

which was initiated by the Polish Bank Association. This interbank and standardized database contains a complex scope of information about Polish real estate market. Until now, AMRON have collected over a million records, whereas E-Valuer database offers information about real estate sale prices from across Poland. It contains over 400,000 records.

The legal basis for the valuation of real estate in Poland are: Real Estate Management Act of August 21, 1997, and Regulation of the Council of Ministers of September 21, 2004 on the valuation of real estate and preparing appraisal report. These regulations do not contain specific calculation procedures. They show only certain principles that should be followed. Some models of valuation are published by the Polish Federation of Real Estate Appraiser Associations in the Universal of National Valuation Rules and in the Interpretative Notes.

One of the four approaches to the valuation task is a comparative approach. The basis for the valuation is a data set of similar properties for which sale transactions were concluded within a short period of time (usually two years). Apart from the information about the transaction prices, real estate appraiser has also information on selected characteristics of the properties (so-called market features). It is assumed that they affect the level of prices. Thus, the valuation can be formulated as a regression task. The transaction database plays the role of the training set with multidimensional observations. The market features are the predictors, and price is the dependent variable. The goal is to predict the value of a new object. The Interpretative Note No. 1 “The application of a comparative approach in the valuation of real estate” specifies two of the three valuation methods: paired comparison and average price adjustment. The third one - the method of statistical analysis of the market – is not clarified, which gives the analyst freedom to use various econometric models. Recommendation J of the Polish Financial Supervision Authority, aimed at the banking sector, pays attention to the need of accurate and stabile models.

In the literature there are several propositions of application of the market statistical analysis. Undoubtedly, the most popular are classical multiple regression model and multiple simple regression model (Hozer 2008). Foryś (2010) applied linear ordering methods in order to select a subset of objects most similar to the valued real estate. Linear ordering methods were applied to construct the aggregated measure of attractiveness, which was used as the predictor in simple regression (Lis 2005; Doszyń 2012). Lis (2005) used information on the cluster structure of the data acquired by k-means method. The dummy variables which identified the clusters were included in the linear regression model. Mach (2012) investigated the impact of regional development on the price of a square meter of residential real estate using factor analysis, and multiple regression in the space of reduced dimension. Trzęsiok (2013) compared various nonparametric regression models on the data from the Warsaw housing market. In the literature on real estate valuation, propositions of neural networks application can also be found in (Lis 2001, Morajda 2005).

In this paper we propose the application of locally estimated regression functions in the neighbourhood of the points that represent the valued real estate. This approach is dedicated to the large data sets of transactions. Additionally, we propose a local feature selection, assuming that the validity of the feature can differ in various regions of the feature space. Feature selection is made with a use of regularization in the estimation criterion. Unfortunately, we do not have access to the large data set from the Polish market. Thus, the proposition of locally regularized linear regression was verified on three publicly accessible data sets from the US market.

2. The estimates of the linear model coefficients

Suppose we are given a set of multivariate observations with known values of a quantitative dependent variable Y (training set):

$$U = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N) : \boldsymbol{x}_i \in \boldsymbol{X} = (X_1, ..., X_p), y_i \in Y, i \in \{1, ..., N\}\}. \quad (1)$$

The goal of regression is to discover the impact of the predictors $X_1, ..., X_p$ on the variable Y. Owing to its simplicity, a linear model is widely used:

$$f(\boldsymbol{x}) = b_0 + b_1 x_1 + ... + b_p x_p + \varepsilon, \quad (2)$$

where ε is the random component, which is assumed to have a normal distribution with mean equal to 0 and constant variance $\forall i \in \{1, ..., N\}$. Moreover, ε_i are independent of each other and independent of predictors. The classical approach to the estimation of the linear model parameters is the ordinary least squares method (OLS). Estimators obtained in this way, under the Gauss-Markov theorem, are unbiased and have the smallest variance in the class of linear and unbiased estimators (Maddala 2008, pp. 228–229). A problem arises when there are irrelevant variables in the data, which do not affect the variable Y. Then the model does not guarantee an accurate prediction for new objects (out of training sample), which is the primary objective of modelling. It has a theoretical foundation in the bias-variance trade-off and formally it is described in (Hastie, Tibshirani, Friedman 2009, pp. 219–224). Generally, too complex models, which are well fitted to the training set, are characterized by a low bias and a high variance of the prediction error. On the other hand, too simple models, which do not extract all information from data, are characterized by a high bias and a low variance of the prediction error. The number of parameters to estimate (which is strongly related to the number of variables) is usually adopted as a measure of the linear model complexity. Thus, elimination of the irrelevant variables should improve the quality of the model. The market features in real estate valuation task are carefully selected by experts, and it would seem that they undoubtedly have an impact on the transaction prices. However, many authors indicate the need for formal, statistical feature selection in the econometric models of valuation (Lis 2005; Zeliaś 2006; Bitner 2007).

The effect of feature selection can be achieved by regularization. A penalty component $P(\boldsymbol{b})$ for large absolute values of the parameters is imposed on the criterion used in the estimation task:

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \left(\sum_{i=1}^{N} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} \right)^2 + \lambda \cdot P(\boldsymbol{b}) \right). \tag{3}$$

Regularization decreases the absolute values of the estimates, and some of them are equal to zero. The core idea is to control the complexity of the model. The objective is to reach a compromise between the bias and the variance, and consequently to get a model with optimal generalization ability, which means the accuracy of prediction for new objects. Various forms of penalty component and the references are given in the Table 1.

Table 1. Penalty components in regularized linear regression

Ridge regression	LASSO	Elastic net
$P(\boldsymbol{b}) = \sum_{j=1}^{p} b_j^2$	$P(\boldsymbol{b}) = \sum_{j=1}^{p} \lvert b_j \rvert$	$P_\alpha(\boldsymbol{b}) = \sum_{j=1}^{p} \left(\alpha b_j^2 + (1-\alpha) \lvert b_j \rvert \right)$
(Hoerl, Kennard 1970)	(Tibshirani 1996)	(Zou, Hastie 2005)

Source: own work.

The regularization parameter λ defines the amount of the penalty, and, as a result, it controls the complexity of the model. Determination of the appropriate value of λ is a key task for the effective application of this method. In practice, a sequence of models corresponding to different values of λ is built, and then the optimal model is selected. Information criteria or prediction error estimated via cross-validation can be used as a model selection criterion. The empirical comparison of these criteria can be found in (Kubus 2013).

The task of parameter estimation in the ridge regression has a solution in a closed form, see, i.e. (Maddala 2008; Hastie *et al.* 2009). LASSO requires quadratic programming with linear constrains but approximate solutions are more practical and more commonly used. Presently, LARS algorithm (Efron, Hastie, Johnstone, Tibshirani 2004) is the most popular because of a low computational complexity. In the case of the elastic net it has been proven that the estimation task can be reformulated on the LASSO (Zou, Hastie 2005). In the following iterations of the LARS algorithm, the coefficients are updated based on the current regression residuals, thus previously unexplained variability of the response *Y*. In every step, the updating formula takes into account some predictors most correlated with *Y*, while only one predictor per iteration is introduced into the model.

An important modelling stage is to assess the quality of th model. A widely used evaluation function is the mean square error, whose unbiased estimator has the form of:

$$MSE = \frac{1}{N-p-1}\sum_{i=1}^{N}\left(y_i - \hat{f}(\boldsymbol{x}_i)\right)^2, \tag{4}$$

or its square root called residual standard error. However, the assessment of the predictive ability of a model, measured on the training set which was used for estimation of model parameters, is too optimistic (Hastie *et al.* 2009, p.228). A model fitted perfectly to the data does not guarantee a great ability to generalize, that is an accurate prediction for new objects (out of the training sample), which is the primary objective of modelling. To estimate the prediction error, the researcher should use a separate set of objects, from the same population, that did not take part in the learning stage. Cross-validation is quite a common strategy of the error estimation. The main idea of cross-validation is to reuse the learning sample many times. In this method, the training set *U* is split into *K* disjoint and approximately equinumerous subsets $V_1, V_2, ..., V_K$. Then *K* models $(\hat{f}_1, ..., \hat{f}_K)$ are built based on training samples $U_i^{CV} = U - V_i$ $(i = 1, ..., K)$, and the prediction errors are estimated based on test samples V_i. Finally, the error is averaged.

3. Local regression models

The assumption about the linear dependency between predictors and the dependent variable is very restrictive. There are numerous propositions of more flexible regression functions in the literature. A short review of non-parametric regression models with the examples of applications in R program is given in (Trzęsiok, Trzęsiok 2009). Some of these methods utilize the idea of local fitting, for example tree-based models. We take under consideration a slightly different approach, which also focus on local fitting. Consider a point $\boldsymbol{x} \in \boldsymbol{X} = (X_1, ..., X_p)$, which represents the valued real estate. Instead of building a global model (in all the domain) we estimate the linear regression function in the neighbourhood of $\boldsymbol{x}$. Additionally, one can use information about distances between points $\boldsymbol{x_i} \in U$ and a query point $\boldsymbol{x}$. Formally, a model of local regressions can be expressed by the formula:

$$\tilde{f}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}, \hat{\boldsymbol{b}}(\boldsymbol{x})), \tag{5}$$

where $\hat{f}$ is the linear model (2) fitted locally in the neighbourhood of ***x***. The parameter vector ***b*** is estimated with the use of weighted least squares:

$$\hat{\boldsymbol{b}}(\boldsymbol{x}) = \underset{\boldsymbol{b}}{arg\,min} \sum_{i=1}^{N} K\left(\frac{\left\|\boldsymbol{x}-\boldsymbol{x}_i\right\|}{h(\boldsymbol{x})}\right)\left(y_i - \hat{f}(\boldsymbol{x_i},\boldsymbol{b})\right)^2 . \tag{6}$$

The weights are obtained by a kernel function $K(\cdot)$, where $h(\boldsymbol{x})$ is a width function that determines the width of the neighbourhood at ***x***. Various forms of kernel functions as well as various techniques of determining the width can be found in the literature (Loader 1999; Hastie *et al.* 2009). A simple solution is to use *k* nearest neighbours method, which adaptively sets the width due to the local density of objects from the training set. Still, one can use weights dependent on the distances from ***x***, or adopt equal weights for all objects from the neighbourhood of ***x***. Note that the models are fitted separately at each query point ***x***.

Assuming that the validity of the feature can differ in various regions of the domain, we propose to introduce the regularization to the criterion of estimation. The penalty component that we have chosen is the elastic net (Zou, Hastie 2005). We have also adopted the simplest system of weights, applying *k* nearest neighbours kernel function. Formally, it can be written as:

$$\hat{\boldsymbol{b}}^{\boldsymbol{l\text{-}EN}}(\boldsymbol{x}) = \underset{\boldsymbol{b}}{arg\,min} \sum_{i=1}^{N} K_{k-NN}\left(\left\|\boldsymbol{x}-\boldsymbol{x}_i\right\|\right)\left(\left(y_i - \hat{f}(\boldsymbol{x_i},\boldsymbol{b})\right)^2 + \lambda\sum_{j=1}^{p}\left(\alpha b_j^2 + (1-\alpha)\left|b_j\right|\right)\right) , \tag{7}$$

where $K_{k-NN}(\cdot)$ is the indicator function, which returns equal weights for objects from the neighbourhood of ***x***, and 0 otherwise.

4. Empirical study

To compare the effectiveness of the global and local modelling of a linear regression function we analyse three publicly available data sets from the real estate market.

The first one concerns housing values in suburbs of Boston (Harrison, Rubinfeld 1978). The market features that are taken into account are {*per capita crime rate by town; proportion of residential land zoned for lots over 25,000 sq. ft.; proportion of non-retail business acres per town; Charles River* dummy variable (= 1 if tract bounds river; 0 otherwise)*; nitric oxides concentration (parts per 10 million); average number of rooms per dwelling; proportion of owner-occupied units built prior to 1940; weighted distances to five Boston employment centres; index of accessibility to radial highways; full-value property-tax rate per $10,000; pupil-teacher ratio by town; 1000(Bk – 0.63)^2 where Bk is the* *proportion of blacks by town; % lower status of the population*}. This data set contains 506 observations (https://archive.ics.uci.edu/ml/datasets.html).

The second data set consists of aggregated data from each of 20,460 neighbourhoods in California. It is available in the StatLib repository (http://lib.stat.cmu.edu/). The dependent variable is the median house value in each neighbourhood. The predictors are {*median income, housing median age, total rooms, total bedrooms, population, households, latitude, longitude*}.

The last data set comes from (Maddala 2008, p. 234-235), and it concerns sale prices of rural land in Florida (per acre). There are 67 multidimensional observations, which are characterized by market features/predictors {*proportion of acreage that is wooded; distance from parcel to Sarasota airport; distance from parcel to highway; acreage of parcel; month in which the parcel was sold*}. Although it is not large data set we decide to take it under consideration to have a wider field for the comparisons.

We use two estimation techniques in the global regression model as well as in local regressions. The ordinary least squares method is compared with regularization, where we apply the elastic net penalty component (Zou, Hastie 2005). The optimal value of the regularization parameter λ has been chosen with the use of Bayesian information criterion BIC. The number of nearest neighbours in the model of local regressions, estimated according to formula (6) or (7), has been set arbitrarily. It is k = 30 in Florida and Boston data set, and k = 50 in the third, larger data set. The residual standard errors for these models have been estimated via 10 fold cross-validation. The comparison is shown in the Table 2.

Table 2. Residual standard errors estimated via cross-validation for global and local fitting

Global linear model

	Florida	Boston	California
OLS	2537.71	4719.15	69581.07
Elastic net	2538.45	4887.26	73979.68

Local linear model

	Florida	Boston	California
OLS	2218.81	3719.58	64044.36
Elastic net	2154.19	3475.37	66036.35

Source: own calculations.

The prediction error of local regression models is lower in all the cases. In two out of three data sets (Florida and Boston) locally regularized regression has given better results. Note that the regularization has not affected the reduction of error in global models. This allows us to suppose that the validity of variables varies in different regions of the feature space.

5. Summary

We found the idea of the local fitting of regression function natural and especially suited to the comparative approach in the real estate valuation. Empirical examples investigated in the previous section confirmed our intuition. The core idea of the comparative approach is to use the information about similar properties that were sold in a short period of time. The use of a distance-based similarity measure in the feature space enables fully automatic and objective identification of such real estate. A special meaning is assigned to it in the analysis of large data sets. Our approach improved the accuracy of the valuation in two data sets. In the third one we obtained a slightly worse model, but still competitive for a global regression. Poorer performance of California housing data set seems to confirm the correct selection of the potential predictors made by experts. When there are no irrelevant variables in the data set, the unbiased OLS estimators are recommended. Note that locally regularized linear regression can be modified in several ways. One can utilize: various weighting systems, adaptive setting of the parameter *k* with the use of a validation set, various penalty components, polynomial regression function.

REFERENCES

BITNER, A., (2007). Konstrukcja modelu regresji wielorakiej przy wycenie nieruchomości [Construction of the multiple regression model in real estate valuation], Acta Scientiarum Polonorum, Administratio Locorum, 6 (4), pp. 59–66.

DOSZYŃ, M., (2012). Ekonometryczna wycena nieruchomości [Econometric evaluation of real estate], Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania Uniwersytetu Szczecińskiego, No. 26, pp. 41–52, Szczecin.

EFRON, B., HASTIE, T., JOHNSTONE, I., TIBSHIRANI, R., (2004). Least Angle Regression, ,,Annals of Statistics" 32 (2), pp. 407–499.

FORYŚ, I., (2010). Wykorzystanie metod taksonomicznych do wyboru obiektów podobnych w procesie wyceny lokali mieszkalnych [The multivariate analysis using to the choice the similar object in the housing valuation process], Studia i Materiały Towarzystwa Naukowego Nieruchomości, Vol. 18, No. 1, pp. 95–105, TNN, Olsztyn.

HARRISON, D., RUBINFELD, D. L., (1978). Hedonic prices and the demand for clean air, J. Environ. Economics & Management, Vol. 5, 81–102.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J., (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, Springer, New York.

HOERL, A. E., KENNARD, R., (1970). Ridge regression: biased estimation for nonorthogonal problems, „Technometrics" 12, pp. 55–67.

HOZER, J., (ed.), (2008). Wycena nieruchomości [Real estate valuation], KEiS US, IADiPG w Szczecinie, Szczecin.

KUBUS, M., (2013). On model selection in some regularized linear regression methods, Acta Universitatis Lodziensis, Folia Oeconomica 285, pp. 115–223.

LIS, C., (2001). Sieci neuronowe a masowa wycena nieruchomości [Neural networks and the mass valuation of real estate], Zeszyty Naukowe US, No 318, Prace Katedry Ekonometrii i Statystyki, Szczecin.

LIS, C., (2005). Ekonometryczne modele cen transakcyjnych lokali mieszkalnych [Econometric models transaction prices of residential premises], Zeszyty Naukowe US, No. 415, Prace Katedry Ekonometrii i Statystyki, No. 16, Szczecin.

LOADER, C., (1999). Local Regression and Likelihood, Springer, New York.

MACH, Ł., (2012). Determinanty ekonomiczno-gospodarcze oraz ich wpływ na rozwój rynku nieruchomości mieszkaniowych [Economic determinants and their impact on development of residential real estate market], Ekonometria, 4 (38), pp. 106–116.

MADDALA, G. S., (2008). Ekonometria [Econometrics], PWN, Warszawa.

MORAJDA, J., (2005). Wykorzystanie perceptronowych sieci neuronowych w zagadnieniu wyceny nieruchomości [The use of perceptrons neural networks in the issue of real estate valuation], Zeszyty Naukowe Małopolskiej Wyższej Szkoły Ekonomicznej w Tarnowie, 7, pp. 101–108.

TIBSHIRANI, R., (1996). Regression shrinkage and selection via the lasso, J.Royal. Statist. Soc. B., 58, pp. 267–288.

TRZĘSIOK, J., TRZĘSIOK, M., (2009). Nieparametryczne metody regresji [Nonparametric regression methods], [in:] M. Walesiak, E. Gatnar (eds), Statystyczna analiza danych z wykorzystaniem programu R [Statistical data analysis with a use of R program], Wydawnictwo Naukowe PWN, Warszawa, pp. 156–192.

TRZĘSIOK, M., (2013). Wycena rynkowej wartości nieruchomości z wykorzystaniem wybranych metod wielowymiarowej analizy statystycznej [Real estate market value estimation based on multivariate statistical analysis], Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, No. 278, Taksonomia 20, Klasyfikacja i analiza danych – teoria i zastosowania, pp. 188–196.

ZELIAŚ, A., (2006). Kilka uwag na temat doboru zmiennych występujących na rynku nieruchomości [Several remarks about the methods of selecting variables occurring on the real estate market], Zeszyty Naukowe US, No 450, Prace Katedry Ekonometrii i Statystyki, No. 17, pp. 685–696, Szczecin.

ZOU, H., HASTIE, T., (2005). Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society Series B, 67 (2), pp. 301–320.

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 525–540

SPATIAL AUTOCORRELATION IN ASSESSMENT OF FINANCIAL SELF-SUFFICIENCY OF COMMUNES OF WIELKOPOLSKA PROVINCE

Agnieszka Kozera[1], Romana Głowicka-Wołoszyn[2]

ABSTRACT

The aim of the article was to identify the spatial effects in assessment of financial self-sufficiency of the governments of communes *(gminas)* of Wielkopolska province *(voivodship)* in 2014, employing global and local Moran I statistics. The level of the governments' self-sufficiency was examined by positional TOPSIS method. The study was based on publicly accessible databases compiled by the Ministry of Finance (*Wskaźniki do oceny sytuacji finansowej jednostek samorządu terytorialnego*) and the Central Statistical Office (*Local Data Bank*). Calculations were performed in R with packages *spdep*, *maptools* and *shapefiles*. The study demonstrated that the communes of Wielkopolska province of comparable levels of financial self-sufficiency exhibited a moderate tendency to cluster. Clusters of high levels gathered around larger urban centres, especially around Poznań, while clusters of low levels – in economically underdeveloped agricultural south-eastern and northern part of the province.

Key words: financial self-sufficiency, communes, TOPSIS method, spatial autocorrelation, Moran I statistics

1. Introduction

Communes are the lowest units of local government charged with responsibility to satisfy basic needs of local communities and to maintain favourable conditions for economic activity of local business. They are also the units that shoulder the bulk of the burden to finance local government's undertakings. They possess legal personality and can set their own financial policy. Hence, they decide on income collection, outlay proportions, budget execution and disposal of assets, described jointly as financial self-sufficiency, whose sound levels underpin sustainable local development and need satisfaction.

[1] Poznań University of Life Sciences – Faculty of Economics and Social Sciences.
E-mail: akozera@up.poznan.pl.

[2] Poznań University of Life Sciences – Faculty of Economics and Social Sciences.
E-mail: roma@up.poznan.pl.

Financial self-sufficiency is not only determined by demographic, social and economic factors, but also by the geographic location of a commune with its natural resources and neighbourhood interactions. The analysis of spatial effects in assessment of financial self-sufficiency may accommodate decisions of higher level administrative units regarding financial support for the communes of clustered low values because spatial autocorrelation statistics give a fuller picture of the direction and strength of spatial interactions and structures than more traditional methods [Janc 2006].

The aim of the article was to identify the spatial effects in assessment of financial self-sufficiency of communal governments of Wielkopolska province in 2014. The study was based on publicly accessible databases compiled by the Ministry of Finance (*Wskaźniki do oceny sytuacji finansowej jednostek samorządu terytorialnego / Indicators for assessment of the financial situation of self-governing territorial units)* and the Central Statistical Office'*sLocal Data Bank*). Calculations were performed in R with packages *spdep, maptools* and *shapefiles*.

2. Research methods

In order to identify spatial effects in assessment of financial self-sufficiency of communes of Wielkopolska province ($N = 226$), first, the synthetic evaluation of the self-sufficiency was performed using the TOPSIS positional method[3]. The procedure to create a synthetic feature is a multi-stage process with six distinctive steps. Step I consisted of selecting simple features defining analysed objects and determining their preference toward a general benchmark. Material and statistical considerations given to the selection of 6 indicators describing self-sufficiency of the communes: WDWM – own income per population (PLN *per capita*), WFIP – share of general and targeted subsidies in total income, WAP – ratio of tax income to current income, WBF – tax income per population (PLN *per capita*), and WIWO – share of investment expenditures in total expenditures [cf. Kozera, Wysocki 2015]. WFIP was considered the only destimulant in the adopted set of features and turned into a stimulant by the following transformation [Wysocki 2010]:

$$x_{ik} = a - b \cdot x_{ik}^{D}, \tag{1}$$

where: x_{ik}^{D} – value of the k^{th} feature, a destimulant $(k \in I_D)$ in the i^{th} object (i = 1,…, N),

x_{ik} – value of the k^{th} feature (k = 1, …, K) transformed into a stimulant,

a, b – arbitrary constants, here a = 0 and b = 1.

[3] It is a robust modification of the ideal method proposed by Hellwig [1968] where the synthetic index values are calculated with respect to one ideal. In TOPSIS (Hwang, Yoon 1981) a positive and negative ideals are considered. Using methods based on just one ideal can lead to erroneous results (cf. Binderman 2006, 2011).

In Step II the values of simple features were normalized with L1-median standardization[4] [Lira i in. 2002, Młodak 2006]:

$$z_{ik} = \frac{x_{ik} - m\tilde{e}d_k}{1.4826 \cdot m\tilde{a}d_k}, \tag{2}$$

where: x_{ik} – value of the k^{th} feature in the i^{th} object,

$m\tilde{e}d_k$ – L1-median vector component corresponding to the k^{th} feature,

$m\tilde{a}d_k = med_i \left| x_{ik} - m\tilde{e}d_k \right|$ – median absolute deviation of k^{th} feature values from the median component of the k^{th} feature,

1.4826 – a constant scale factor corresponding to normally distributed data, $\sigma \approx E(1.4826 \cdot m\tilde{a}d_k(X_1, X_2, ..., X_K))$,

σ – standard deviation.

The distribution of feature values standardized in this way is considered "close to the normal distribution of zero expectation and unitary standard deviation" [Młodak 2009, s. 3-21].

In Step III the coordinates of ideal and negative ideals were computed according to the following formulae [Wysocki 2010]:

$$A^{+} = \left(\max_i (z_{i1}), \max_i (z_{i2}), ..., \max_i (z_{iK}) \right) = \left(z_1^{+}, z_2^{+}, ..., z_K^{+} \right) \tag{3}$$

for the ideal, and:

$$A^{-} = \left(\min_i (z_{i1}), \min_i (z_{i2}), ..., \min_i (z_{iK}) \right) = \left(z_1^{-}, z_2^{-}, ..., z_K^{-} \right). \tag{4}$$

for the negative ideal. These coordinate values in Step IV yielded the distance of each object to the ideal (A^{+}) and negative ideal (A^{-}) using the following formula [Wysocki 2010]:

$$d_i^{+} = med_k \left(\left| z_{ik} - z_k^{+} \right| \right),\ d_i^{-} = med_k \left(\left| z_{ik} - z_k^{-} \right| \right),\ (i = 1, \ldots, N), \tag{5}$$

where: med_k – marginal median for the k^{th} feature.

The construction of the synthetic index in Step V followed the TOPSIS method [Hwang, Yoon 1981]:

$$S_i = \frac{d_i^{-}}{d_i^{-} + d_i^{+}},\ (i = 1, \ldots, N), \tag{6}$$

where $0 \le S_i \le 1$ can be easily verified.

[4] As opposed to classical approach, using spatial (L1) median is robust against the undue influence of outliers (Młodak 2006).

The values of the synthetic index were used in Step VI to linearly arrange analysed communes in non-decreasing order. Then four typological classes were established, with cut-offs depending on the mean $(\bar{S})$ and standard deviation (s_S) of the index [Wysocki 2010]:

$$\begin{aligned}
&\text{class I (high level of financial self-sufficiency): } S_i \geq \bar{S} + s_S,\\
&\text{class II (medium high level): } \bar{S} \leq S_i < \bar{S} + s_S,\\
&\text{class III (medium low level): } \bar{S} - s_S \leq S_i < \bar{S},\\
&\text{class IV (low level): } S_i < \bar{S} - s_S.
\end{aligned} \tag{7}$$

In the next phase of the study the strength and character of spatial autocorrelation of financial self-sufficiency were examined. The global Moran I statistic was used to find the autocorrelation overall estimate within the whole province. The local Moran I statistics were employed to identify the spatial layout of communes of Wielkopolska province, detecting clusters of similar values of financial self-sufficiency, as well as outliers.

Spatial analysis is primarily interested in the type of the effect, i.e. spatial homo- or heterogeneity. Spatial autocorrelation is usually understood as a correlation between the values of the same variable for different objects in space, and, therefore, it measures the degree of dependence of these objects set within a geographic framework [Kossowski et al. 2013]. To get a global gauge of this dependence, the global Moran I statistic is commonly employed [Bivand et al. 2008, Kopczewska 2006]:

$$I = \frac{N}{S_0} \cdot \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_j (x_j - \bar{x})^2}, \tag{8}$$

where: w_{ij} – spatial weight of the link between objects *i* and *j* (and element of 0-1 neighbourhood weight matrix **W** based on common border criterion[5]), where:

$$w_{ij} = \begin{cases} 1 & \text{when the } i^{\text{th}} \text{ object is a neighbour of the } j^{\text{th}} \text{ object,} \\ 0 & \text{when the } i^{\text{th}} \text{ object is not a neighbour of the } j^{\text{th}}, \\ 0 & \text{when } i = j \text{ (diagonal elements of the matrix),} \end{cases} \tag{9}$$

$$S_0 = \sum_i \sum_j w_{ij},$$

[5] This matrix is often used in social and economic studies and theoretical explorations of spatial economics [cf. Sikora et al. 2014, Pietrzykowski 2011, Kopczewska 2006, Bivand, and Portnov 2004].

x_i – feature value for the i^{th} object,

$\bar{x}$ – feature value averaged over all objects,

N – number of all studied objects.

The statistic is a concise measure of the spatial dependence structure of studied objects (communes of Wielkopolska province) and it ranges over the [-1, 1] interval. Positive values of the statistic signal the existence of clustering effects among the objects, while negative values – checkerboard patterns [Kossowski et al., 2013].

Separately, the correlation between the feature value for a given object with the values of adjacent objects can be studied using local Moran I statistic [Kopczewska 2006]:

$$I_i = \frac{(x_i - \bar{x})\sum_{j=1}^{N} w_{ij}(x_j - \bar{x})}{\sum_{j=1}^{N}(x_j - \bar{x})^2 / N}, \qquad (10)$$

where w_{ij} are the elements of row-standardized spatial weight matrix **W**.

The local Moran I statistic is positive for an object surrounded by those of similar feature values. The autocorrelation is then said to be positive. On the other hand, negative values of local Moran I correspond to a situation when the object has diametrically different feature values to its neighbours, and to the autocorrelation being negative. More broadly, local Moran I statistic can be used to identify agglomeration effects – clusters of similar feature values and outliers, or objects that stand out from its neighbourhood.

Figure 1. Spatial dependencies between an object and its neighbourhood

	Low values in the object	High values in the object
High values in the neighbourhood	**LH** (negative spatial autocorrelation)	**HH** (positive spatial autocorrelation)
Low values in the neighbourhood	**LL** (positive spatial autocorrelation)	**HL** (negative spatial autocorrelation)

Source: own elaboration of Kopczewska [2006].

Two types of clusters can be defined (Fig.1): HH – objects of high feature values surrounded by objects of likewise high values, and LL – objects of low feature values bordering similar neighbours. Similarly, there can be two types of outliers: LH and HL. The map of agglomeration effects is based on statistically significant local Moran I values.

The calculation of global and local Moran I statistics and the maps presented in the article were executed in R using packages *spdep*, *maptools* and *shapefiles*.

Results of Phase I – Synthetic evaluation of financial self-sufficiency of communes of Wielkopolska province

The analysis was carried out by taxonomic methods described above; four typological classes of financial self-sufficiency of the communes were distinguished (Table 1), their spatial delimitation presented in Figure 2.

Class I, comprised of 23 or 10.2% of all province's communes, showed a high level of financial self-sufficiency. Nine of those were from the Poznań metropolitan area (i.e. Suchy Las, Tarnowo Podgórne, Czerwonak, Dopiewo, and Komorniki), attesting to the beneficial influence of large urban centres. On the other hand, high self-sufficiency of such communes as Przykona was a direct consequence of the existing mining industry exploiting its natural resources, or in the case of Powidz – of tourist functional type of the commune coupled with the location of an air base. Communal governments of Class I stood out for the highest level of own income per capita (WDWM – PLN 2,394), the lowest level of state intervention (WFIP 29,2%), and the highest level of fiscal wealth (WAP – PLN 1,911) (Table 1).

Class II of medium high self-sufficiency included 80 communes, or 35.4% of the total, and had above average level of own income per capita (PLN 1,715) and fiscal wealth (PLN 1,310). They were found mainly in the central west part of the province and relatively close to its capital or other large urban centres, which bolstered their residential and service-oriented character (Table 1).

Table 1. Intraclass values of the financial self-sufficiency indicators of communes of Wielkopolska province in 2014 (median for average values)

Specification	Typological class of financial self-sufficiency				Total
	I high	II medium high	III medium low	IV low	
Financial self-sufficiency indicators					
Percentage of all communes (%)	10.2	35.4	42.0	12.4	100.0
WDWM (*PLN per capita*)	2,394.1	1,715.7	1,194.3	843.4	1,399.7
WFIP (%)	29.2	43.6	59.4	69.8	52.9
WBF (*PLN per capita*)	1,910.8	1,310.4	868.2	594.2	1,020.1
WAP (%)	32.7	23.5	16.0	10.0	18.7
WIWO (%)	19.7	15.1	13.5	12.6	14.9

Table 1. Intraclass values of the financial self-sufficiency indicators of communes of Wielkopolska province in 2014 (median for average values) (cont.)

Specification	Typological class of financial self-sufficiency				Total
	I high	II medium high	III medium low	IV low	
Social and economic determinants					
Distance to urban centres (km)	29.7	31.4	36.7	37.5	34.2
Population density (inhabitant/km^2)	122.0	91.5	62.0	57.0	69.0
Net migration per 1000 inhabitants [a)]	2.6	-0.9	-1.9	-1.7	-1.1
Employed on individual farmsteads per 100 working age persons [b)]	11.3	15.9	29.0	44.4	24.4
Number of economic entities in REGON registry per 100 inhabitants	1,682.2	1,538.9	1,196.7	1,005.8	1,248.4

a) calculated for 2012-2014.

b) own calculation.

Source: own elaboration based on data from the Central Statistical Office (Local Data Bank) and the Ministry of Finance (Wskaźniki do oceny sytuacji finansowej….).

Class III of medium low level self-sufficiency was formed by 95 communes, chiefly from the north and south of the province, which represented 42.0% of the total. Their financial sufficiency indicators were all below average, and they were marked also by below average population density and below average level of economic activity (Table 1).

Class IV of financially least self-sufficient group of 28 communes (12.4% of all), located mainly in the southeast of the province, was characterized by the lowest, PLN 843.4 levels of own income per capita (which was 40% lower than the average). At the same time, these communes had the highest share of transfers in total income (WFIP = 69.8%). Low level of own income and its low shares in the communes' budgets lead to low levels of investment expenditure shares in total expenditures. These communes were agriculture-oriented with high numbers of employed on individual farmsteads (44.4) and a low economic activity, owing largely to remoteness from urban centres, especially from Poznań, the economic centre of the province.

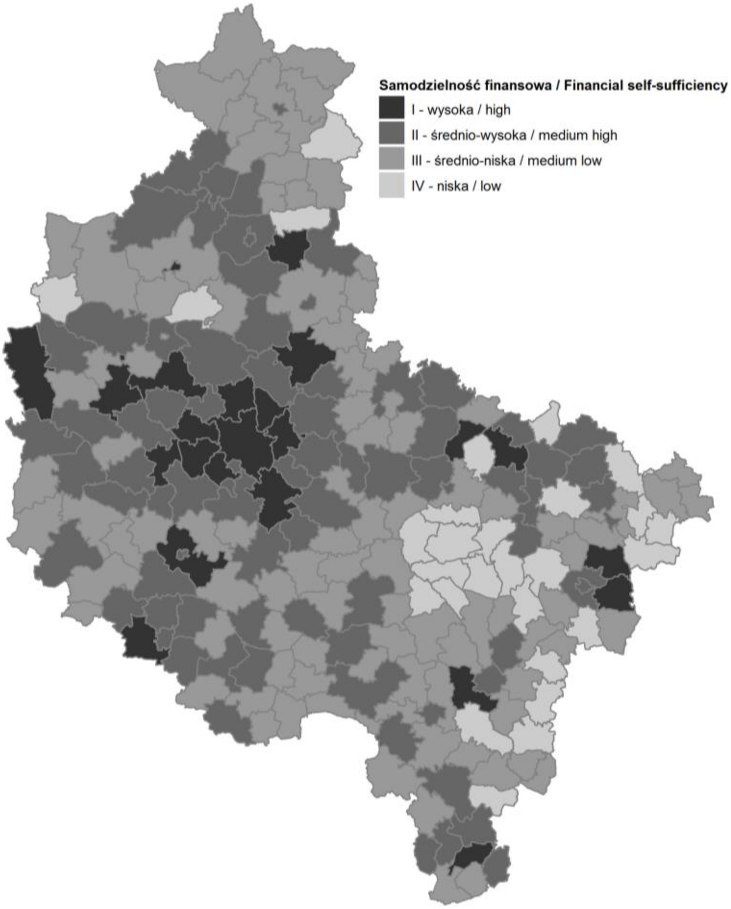

Samodzielność finansowa / Financial self-sufficiency
I - wysoka / high
II - średnio-wysoka / medium high
III - średnio-niska / medium low
IV - niska / low

Figure 2. Spatial delimitation of financial self-sufficiency typological classes of communes of Wielkopolska province in 2014

Source: ibid.

Results of Phase I – Identification of the strength and character of financial self-sufficiency spatial autocorrelation in communes of Wielkopolska province

The strength and character of financial self-sufficiency spatial autocorrelation in communes of Wielkopolska province were identified using the global Moran I statistic, whose value reached 0.289 and proved statistically significant ($p < 0.01$). The result indicated a moderately positive spatial autocorrelation and a tendency of communes with comparable levels of financial self-sufficiency to cluster together. It should be stressed, however, that the measure does not offer an insight into the local deviations from the globally depicted pattern. In fact, in certain areas, such as the Poznań metro[6], communes display stronger connections than would have been suspected from global autocorrelation calculations only.

Thus, a more exhaustive inquiry called for employment of the local Moran I statistic, computed for every commune, which allowed for studying departures from global autocorrelation. Figure 3 shows the dispersion of local Moran I values, with standardized values of financial self-sufficiency synthetic index on the horizontal axis, and the spatial lag[7] in the index on the vertical one. Values of local Moran I statistic for selected communes are presented in Table 2. In general, they proved statistically significant for 40 communes (Figure 5), 36 of which were clustered, i.e., adjacent to communes with similar index values. Figure 4 presents the membership of the communes in Moran I scatterplot quadrants. The darkest colour denotes clusters of high financial self-sufficiency (HH), while the lightest one denotes clusters of low financial self-sufficiency (LL).

The study indicates that the communes with high financial self-sufficiency concentrate mainly in Poznań metropolitan area, producing a cluster to encompass not just the first ring around the city, but also the second and the third ones. Smaller clusters of high financial self-sufficiency were found in the north around the cities of Chodzież and Piła, and in the south around Leszno, but also in tourist districts (Powidz, Ślesin) and in mineral rich belts of eastern Wielkopolska.

The research also found low financial self-sufficiency clusters (LL): in the north, of communes heavily forested and sparsely populated, and in the southeast, of communes with agricultural character, sizable employment on individual farms, and low economic activity [cf. Kozera, Wysocki 2015].

[6] Poznań metropolitan area is the economic powerhouse of the province, situated in its center and encompassing the City of Poznań and 17 communes of the Poznań county.

[7] Spatial lag is taken for each object to be the weighted average of feature values in adjacent objects, $lag(x_i) = \sum_{j \in N_i} w_{ij} x_j$ [Kopczewska 2006].

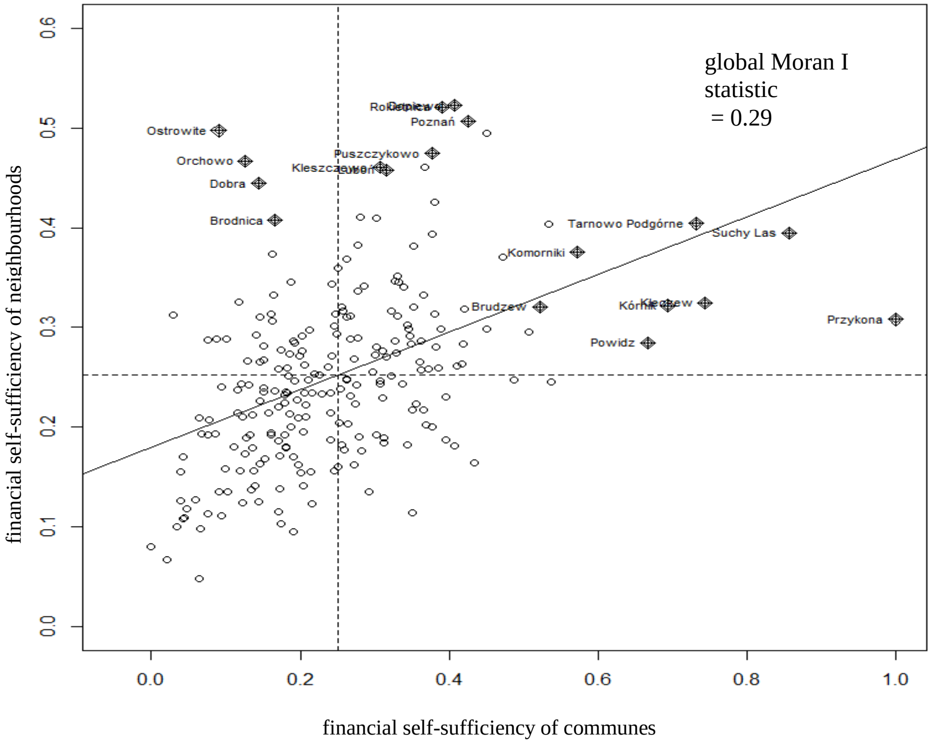


Figure 3. Dispersion of local Moran I values for the levels of communes of Wielkopolska province financial self-sufficiency in 2014

Source: ibid.

Table 2 presents examples of communes of Wielkopolska province whose local Moran I statistics proved significant: some of the HH and some of the LL clusters, as well as all outliers, i.e. communes surrounded by neighbours of diametrically different levels of financial self-sufficiency. The latter included LH (low surrounded by high) communes: Dobra of Turek county, and Orchowo and Ostrowite of Słupca county. Dobra, of medium low financial self-sufficiency, is a typically agricultural commune that borders Przykona and Turek communes with their coal mines and power plants. Similarly, Orchowo and Ostrowite, of low financial self-sufficiency and bordering tourism-oriented communes, can have their low self-sufficiency explained by a considerable distance to urban centres, bleak demographics (sparse population, low net migration), and economic relative inactivity. On the other side of the Moran diagram there was an HL outlier, Stare Miasto of Konin county. It is a residential neighbourhood of the city of Konin with double the average population density and a high number of registered economic entities (Table 2).

Table 2. Values of local Moran I statistic with respect to the level of financial self-sufficiency for selected communes of Wielkopolska province

Direction of auto-correlation		Positive						Negative			Total
Commune		Poznań	Suchy Las	Tarnowo Podgórne	Grodziec	Gizałki	Zagórów	Stare Miasto	Orchowo	Ostrowite	
Class of financial self-sufficiency		I (high)			IV (low)			II (medium high)	III (medium low)	IV (low)	
Local Moran I statistic		7.451	9.693	8.985	5.571	5.571	4.132	-1.652	-2.951	-3.880	
p-value		0.000	0.000	0.000	0.000	0.000	0.000	0.951	0.998	1.000	
Type of neighbour-hood dependency		HH			LL			HL	LH		
Socio-economic determinants	Distance to Poznań (km)	×	10.8	19.5	107	92.3	84.0	103.0	95.7	137.0	×
	Population density (km^2)	2,083	140	236	45	43	57	120	40	50	69
	Net migration per 1000 inhabitants	-4.2	17.5	18.7	-1.7	-2.8	-3.3	8.5	-5.6	0.2	-1.1
	Employed on individual farmsteads per 100 working age persons	0.7	5.7	10.7	57.2	46.3	41.2	40.2	29.0	43.4	24.4
	Number of economic entities in REGON registry per 100 inhabitants	3,153.6	3,487.0	3,251.8	1,047.4	1,291.5	1,092.8	1,628.6	1,117.4	935.1	1,248.4

Source: ibid.

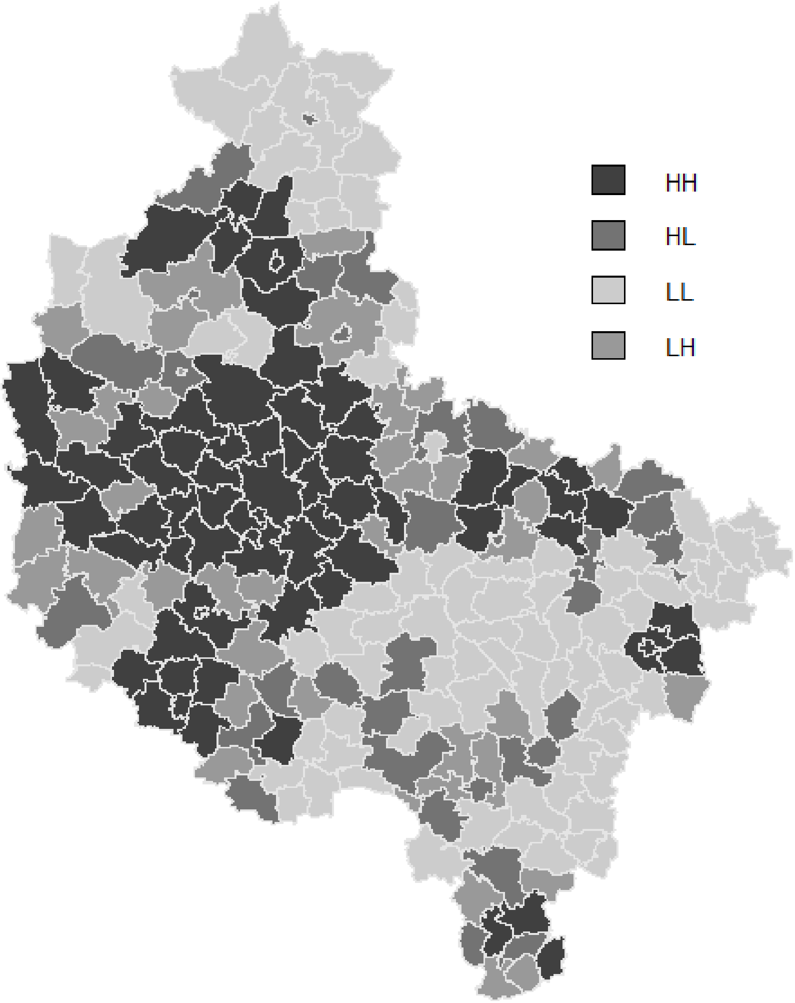


Figure 4. Membership of communes of Wielkopolska province in Moran scatterplot quadrants

Source: ibid.

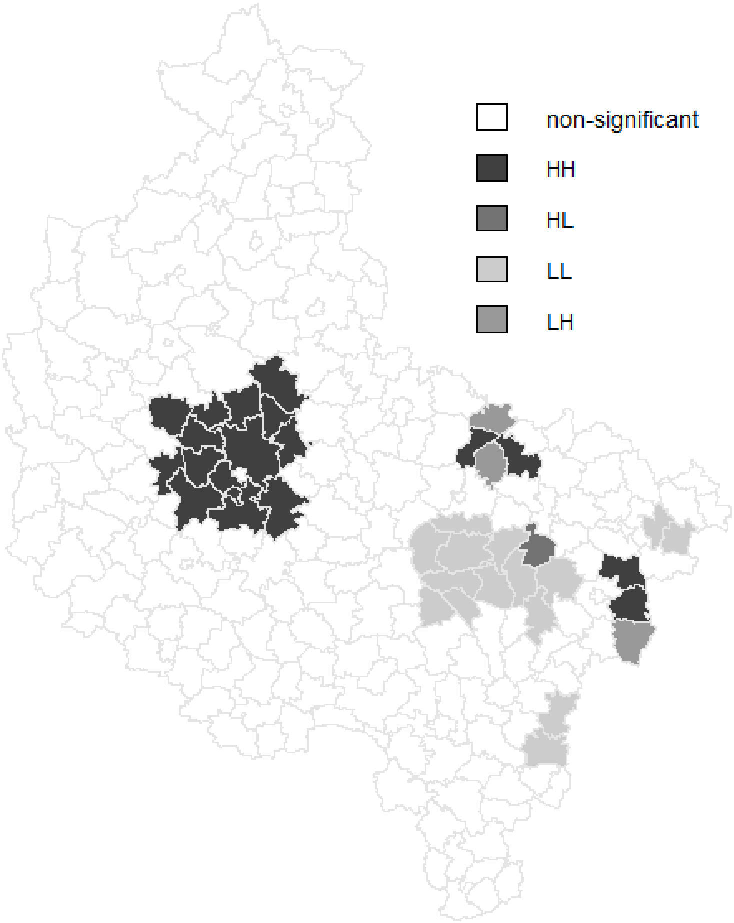


Figure 5. Significance of local Moran I statistics for communes of Wielkopolska province

Source: ibid.

3. Conclusions

Spatial methods are becoming increasingly popular in analysing not just economic, but also financial data. Autocorrelation tools may facilitate this analysis and simplify the process of revealing the underlying spatial structure of connections between local government units, also in the scope of financial self-sufficiency.

This paper studied the spatial effects in assessment of financial self-sufficiency of communes of Wielkopolska province. Computations of global

Moran I statistic showed positive spatial autocorrelation, which signified the existence of clusters of communes with similar levels of financial self-sufficiency. Local Moran I statistic assisted in finding those clusters: a high level in the Poznań metropolitan area, the product of the ongoing suburbanization of the provincial capital, and a low level in economically underdeveloped agricultural communes of northern and south-eastern parts of the province. It is difficult to speculate how the results for communes of Wielkopolska province extend to the whole country, or what is the strength that other urban centres exert on their neighbourhoods' financial self-sufficiency. Certainly, these questions warrant further research.

REFERENCES

BINDERMAN, A., (2006). Klasyfikacja danych na podstawie dwóch wzorców[On a classification of objects basing on two models]. Ekonomika i Organizacja Gospodarki Żywnościowej [Economics and Organization of Agri-Food Sector] SGGW Warszawa, No. 60, pp. 25–34.

BINDERMAN, A., (2011). Wielokryterialne metody analizy zróżnicowania polskiego rolnictwa w 2009 roku [On Multi-Criteria Decision Methods for Study of a Level of Differentiation of Polish Agriculture in 2009]. Metody ilościowe w badaniach Ekonomicznych [Quantitative Methods in Economics], Tom XII/2, 2011, pp. 58–68.

BIVAND, R. S., PEBESMA, E. J., GOMEZ-RUBIO, V., (2008). Applied spatial data analysis with R. New York: Springer.

BIVAND, R. S., PORTNOV, B. A., (2004). Exploring spatial data analysis techniques using R: The case of observations with no neighbors. In: Anselin, L., Florax, R. J., and Rey, S. J., editors, Advances in Spatial Econometrics: Methodology, Tools and Applications, pp. 121–142. Springer-Verlag, Berlin.

HELLWIG, Z., (1968). Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr [Procedure of evaluating high level manpower data and typology of countries by means of the taxonomic method]. Przegląd Statystyczny [Statistical Review], No. 4, pp. 307–327.

HWANG, C. L., YOON, K., (1981). Multiple Attribute Decision Making Methods and Applications. Springer-Verlag, Berlin.

JANC, K., (2006). Zjawisko autokorelacji przestrzennej na przykładzie statystyki I Morana oraz lokalnych wskaźników zależności przestrzennej (LISA) – wybrane zagadnienia metodyczne [Spatial autocorrelation as exemplified by Moran's I statistic and local indicators of spatial association (LISA)] [w:] Komornicki T., Podgórski T. (red.): Idee i praktyczny uniwersalizm geografii. Dokumentacja geograficzna [Ideas and practical universalism in geography], 33, pp. 76–83.

KOPCZEWSKA, K., (2006). Ekonometria i statystyka przestrzenna z wykorzystaniem programu R Cran [Econometrics and spatial statistics with R Cran]. Wydawnictwo CeDeWu.pl.

KOSSOWSKI, T., PERDAŁ, R., HAUKE, J., (2013). Identyfikacja efektów przestrzennych w badaniu obszarów wzrostu i stagnacji w Polsce w zakresie infrastruktury technicznej [Identification of spatial effects in the study of growth and stagnation areas in Poland in terms of technical infrastructure] [w:] Gulczyński W. (red.): Lokalne i regionalne problemy gospodarki przestrzennej [Local and regional problems of spatial economy]. Wydawnictwo Wyższej Szkoły Biznesu w Gorzowie Wielkopolskim, pp. 79–97.

KOZERA, A., WYSOCKI, F., (2015). Typ funkcjonalny a samodzielność finansowa gmin wiejskich województwa wielkopolskiego [The functional type and financial self-sufficiency of rural communes of the Wielkopolska province]. Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu, tom XVII, No. 6.

MORAN, P. A. P., (1950). Notes on continuous stochastic phenomena. Biometrika, 37, pp. 17–23.

MŁODAK, A., (2006). Analiza taksonomiczna w statystyce regionalnej [The taxonomic analysis of regional statistics]. Wyd. Difin. Warszawa.

MŁODAK, A., (2009). Historia problemu Webera [History of the Weber problem]. Matematyka Stosowana: matematyka dla społeczeństwa [Applied mathematics: mathematics for society], Polskie Towarzystwo Matematyczne, tom 10/51, pp. 3–21.

PIETRZAK, M., (2010). Problem identyfikacji struktury danych przestrzennych [The problem of identification of the structure of spatial data]. Acta Universitatis Nicolai Copernici, Ekonomia XLI, Nauki Humanistyczno-Społeczne, No. 397, pp. 83–98.

PIETRZYKOWSKI, R., (2011). Wykorzystanie metod statystycznej analizy przestrzennej w badaniach ekonomicznych [The use of statistical methods for spatial analysis in the study of economic] . Roczniki Ekonomiczne Kujawsko-Pomorskiej Szkoły Wyższej w Bydgoszczy, No. 4, pp. 97–112.

SIKORA, J., SZELĄG-SIKORA, A., CUPIAŁ, M., (2014). Autokorelacja przestrzenna wykorzystania pozabudżetowych środków w gminach województwa wielkopolskiego [Systems IT for agricultural businesses]. Infrastruktura i Ekologia Terenów Wiejskich, Polska Akademia Nauk, Komisja Technicznej Infrastruktury Wsi, No. IV, pp. 1317–1326.

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 541–556

KERNEL ESTIMATION OF CUMULATIVE DISTRIBUTION FUNCTION OF A RANDOM VARIABLE WITH BOUNDED SUPPORT

Aleksandra Baszczyńska[1]

ABSTRACT

In the paper methods of reducing the so-called boundary effects, which appear in the estimation of certain functional characteristics of a random variable with bounded support, are discussed. The methods of the cumulative distribution function estimation, in particular the kernel method, as well as the phenomenon of increased bias estimation in boundary region are presented. Using simulation methods, the properties of the modified kernel estimator of the distribution function are investigated and an attempt to compare the classical and the modified estimators is made.

Key words: boundary effects, cumulative distribution function, kernel method, bounded support.

1. Introduction

Nonparametric methods are becoming increasingly popular in statistical analysis of economic problems. In most cases, this is caused by the lack of information, especially historical data, about the economic variable being analysed. Smoothing methods concerning functions, such as density or cumulative distribution, play a special role in a nonparametric analysis of economic phenomena. Knowledge of density function or cumulative distribution function, or their estimates, allows one to characterize the random variable more completely.

Estimation of functional characteristics of random variables can be carried out using kernel methods. The properties of the classical kernel methods are satisfactory, but when the support of the variable is bounded, kernel estimates may suffer from boundary effects. Therefore, the so-called boundary correction is needed in kernel estimation.

[1] Department of Statistical Methods, University of Łódź. E-mail: albasz@uni.lodz.pl.

Kernel estimator of cumulative distribution function has to be modified when the support of the variable is defined as $[a,\infty)$, $(-\infty,b]$ or $[a,b]$. Such a situation is frequently observed in an economic analysis, for example, when data are considered only on the positive real line (e.g.: arable land, energy use, CO_2 emission, external debts stocks, current account balance, total reserves, etc.). Near zero, the classical kernel distribution function estimator is poor because of its considerable bias. The bias comes from the behaviour of kernel estimator, which has no knowledge of the boundary and assigns probability on the negative real line. A range of boundary correction methods for kernel distribution function estimator is present in the literature. They are addressed mainly to boundary kernels (Tenreiro, 2013; Tenreiro, 2015) and reflection method (Koláček, Karunamuni, 2009; Koláček, Karunamuni, 2012).

In Section 2 we introduce the kernel method, which for the first time was implemented in density estimation in the late 1950s. The properties of the kernel density estimator, as well as the modifications, are presented taking into account the boundary effects reduction of classical kernel density estimator. In Section 3 some selected methods of distribution function estimation are presented, including the kernel method. Some methods of choosing the smoothing parameter of kernel method and properties of estimator are shown, and methods of boundary correction are used in the case of cumulative distribution function estimation. In Section 4 the results of a simulation study are given and an attempt to compare the considered estimators is made. In addition, the comparison of the values of smoothing parameters is presented. The simulations and the plots were carried out using MATLAB software.

The aim of the paper is to give a detailed presentation of methods of the modified kernel distribution function estimation and to compare the considered methods. The simulation shows that when boundary correction is used in kernel estimation of distribution function, the estimator has better properties.

2. Kernel method

The kernel method originated from the idea of Rosenblatt and Parzen dedicated to density estimation. The Rosenblatt-Parzen kernel density estimator is as follows (cf.: Härdle, 1994; Wand, Jones, 1995; Silverman 1996; Domański et al., 2014):

$$\hat{f}_n(x)=\frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right), \tag{1}$$

where: $X_1, X_2,..., X_n$ is the random sample from the population with unknown density function $f(x)$; n is the sample size; h_n is the smoothing parameter, which controls the smoothness of the estimator ($\lim_{n\to\infty} h=0$, $\lim_{n\to\infty} nh=\infty$).

Throughout this paper the notation $h = h_n$ will be used. $K(u)$ is the weighting function called the kernel function. When $K(u)$ is symmetric and unimodal function and the following conditions are fulfilled:

$$\begin{cases} \int\limits_{-\infty}^{\infty} K(u)du = 1, \\ \int\limits_{-\infty}^{\infty} uK(u)du = 0, \\ \int\limits_{-\infty}^{\infty} u^2 K(u)du = \kappa_2 > 0, \end{cases} \tag{2}$$

the kernel function is called the second order kernel function (or classical kernel function).

The most frequently used Gaussian kernel function $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$ is a function belonging to this group, although its support is unbounded. It stands in contrast to other kernel functions fulfilling conditions (2), like functions presented in Table 1, for which support is bounded. The indicator function $I(|u| \le 1)$ is defined as follows: $I(|u| \le 1) = 1$ for $|u| \le 1$, $I(|u| \le 1) = 0$ for $|u| > 1$.

Table 1. Kernel functions

Kernel function	$K(u)$
Uniform	$\frac{1}{2} I(\lvert u\rvert \le 1)$
Triangle	$(1-\lvert u\rvert) I(\lvert u\rvert \le 1)$
Epanechnikov	$\frac{3}{4}(1-u^2) I(\lvert u\rvert \le 1)$
Quartic	$\frac{15}{16}(1-u^2)^2 I(\lvert u\rvert \le 1)$
Triweight	$\frac{35}{32}(1-u^2)^3 I(\lvert u\rvert \le 1)$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) I(\lvert u\rvert \le 1)$

Higher order kernel functions (the order of the kernel is the order of the first nonzero moment) can be used, especially in reducing the mean squared error of the estimator. But the higher order kernels properties may sometimes be unacceptable because they may result in taking negative values for the density function estimators.

When the support of random variable is, for example, left-bounded (support of random variable is $[0,\infty)$), the properties of the estimator (1) may differ in boundary region $[0,h]$ and in inner region (h,∞)(cf.: Jones, 1993; Jones, Foster, 1996; Li, Racine, 2007). The estimator (1) is not consistent in boundary region. As a result, the support of the kernel density estimator may differ from the support of the random variable and the estimator may be non-zero for negative values of random variable. Moreover, this situation may appear when the kernel function has unbounded as well as bounded support. Removing boundary effects can be done in various ways. The best known and most often used method is the reflection method, which is characterized by both simplicity and best properties.

Assuming that the support of random variable is $[0,\infty)$, the reflection kernel density estimator, based on reflecting data about zero, has the following form (cf. Kulczycki, 2005):

$$\hat{f}_{nR}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{x-X_i}{h}\right)+K\left(\frac{x+X_i}{h}\right)\right]. \tag{3}$$

The estimator (3) is a consistent estimator of unknown density function f. Moreover, it integrates to unity and for x close to zero the bias is of order $O(h)$. The analysis of the properties of this estimator is presented in Baszczyńska (2015), among others.

3. Distribution function estimation

Let $X_1, X_2, ..., X_n$ denote independent random variables with a density function f and a cumulative distribution function F. One can estimate the cumulative distribution function (CDF) by:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i), \tag{4}$$

where I_A is the indicator function of the set A: $I_A(x)=1$ for $x \in A$, $I_A(x)=0$ for $x \notin A$.

The empirical distribution function defined by (4) is not smooth, at each point $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ it jumps by $\frac{1}{n}$.

The smoothed version of the empirical distribution estimator is the Nadaraya kernel estimator of CDF:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-X_i}{h}} K(y)dy = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x-X_i}{h}\right), \tag{5}$$

where h is a smoothing parameter such as $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh = \infty$ and $W(x) = \int_{-1}^{x} K(t)dt$. Assuming that function $K(x) \geq 0$ is a unimodal, symmetric kernel function of the second order with support $[-1,1]$ (examples of these kernels are presented in Table 1), the properties of function $W(x)$ are the following:

$$W(x) = 0 \text{ for } x \in (-\infty,-1],$$

$$W(x) = 1 \text{ for } x \in [1,\infty),$$

$$\int_{-1}^{1} W^2(x)dx \leq \int_{-1}^{1} W(x)dx = 1, \tag{6}$$

$$\int_{-1}^{1} W(x)K(x)dx = \frac{1}{2},$$

$$\int_{-1}^{1} xW(x)K(x)dx = \frac{1}{2}\left[1 - \int_{-1}^{1} W^2(x)dx\right].$$

Function $W(x)$ is a cumulative distribution function because $K(x)$ is a probability density function. For example, when the kernel function is Epanechnikov kernel, the function $W(x)$ has the form:

$$W(x) = \begin{cases} 0 \quad \text{for} \quad x \leq -1, \\ -\frac{1}{4}x^3 + \frac{3}{4}x + \frac{1}{2} \quad \text{for} \quad |x| \leq 1, \\ 1 \quad \text{for} \quad x \geq 1. \end{cases}$$

Assuming additionally that $F(x)$ is twice continuously differentiable, the mean integrated squared error (MISE) of kernel distribution estimator (5) is as follows:

$$MISE\left[\hat{F}(x)\right]=E\int_{-\infty}^{+\infty}\left[\hat{F}(x)-F(x)\right]^2dx=$$

$$=\frac{1}{n}\int_{-\infty}^{+\infty}F(x)(1-F(x))dx-c_1\frac{h}{n}+c_2h^4+o\left(\frac{h}{n}+h^4\right), \tag{7}$$

where:

$$c_1=1-\int_{-1}^{1}W^2(t)dt\,, \tag{8}$$

$$c_2=\frac{\kappa_2^2}{4}\int_{-\infty}^{+\infty}\left[F^{(2)}(t)\right]^2dt\,, \tag{9}$$

$F^{(s)}$ denotes the *s*th derivative of the cumulative distribution function.

Kernel distribution estimator (5) is a consistent estimator of the distribution function. The expectation value, bias and variance are, respectively:

$$E\left[\hat{F}(x)\right]=F(x)+\frac{1}{2}F^{(2)}(x)h^2\kappa_2+o\left(h^2\right),$$

$$B\left[\hat{F}(x)\right]=\frac{1}{2}F^{(2)}(x)h^2\kappa_2+o\left(h^2\right),$$

$$D^2\left[\hat{F}(x)\right]=\frac{1}{n}F(x)(1-F(x))-\frac{1}{n}hf(x)\left[1-\int_{-1}^{1}W^2(t)dt\right]+o\left(\frac{h}{n}\right).$$

The method of choosing the value of the smoothing parameter in kernel estimation of the cumulative distribution function is of crucial interest, as it is in kernel estimation of the density function. Some procedures used frequently in CDF estimation are presented in Table 2.

Table 2. Methods of choosing the smoothing parameter in kernel estimation of the cumulative distribution function

Method	Smoothing parameter
Cross-validation, CV	$\hat{h}_{CV} = \underset{h\in H_n}{\arg\min}\, CV(h)$, $CV(h)=\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\left[I_{(-\infty,x]}(X_i)-\hat{F}_{-i}(x,h)\right]^2 dx$, $\hat{F}_{-i}(x,h)$ is a kernel estimator based on the sample with X_i deleted
Maximal smoothing principle, MSP	$h_{MS} = \left(\frac{7c_1}{15\kappa_2^2}\right)^{1/3} n^{-1/3}\sqrt{7}\hat{\sigma}^2$
Plug-in, PI	$h_{PI} = \left(\frac{c_1}{-\hat{\psi}_1\kappa_2^2}\right)^{1/3} n^{-1/3}$, $\hat{\psi}_k(g)=\frac{1}{n^2 g}\sum_{i,j=1}^{n}L^{(2k)}\left(\frac{X_i-X_j}{g}\right)$, g is an initial smoothing parameter, $L^{(2k)}$is the *2k*th derivative of the initial kernel function L
Iteration, IM	$h_{IT,j+1}=\frac{4h_{IT,j}}{c_1 n}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\Omega\left(\frac{X_i-X_j}{h_{IT,j}}\right)$, $j=0,1,\ldots,$ $\Omega(u)=(K*K*W*W-2K*W*W+W*W)(u)$, $f*g$ denotes convolution

When the random variable has bounded support (without loss of generality one can take $[0,\infty)$), as in the case of the kernel density estimation, the properties of the kernel distribution function get poorer, in comparison with the situation when the support is unbounded.

For x in boundary region $x\in[0,h]$, let $c=x/h$, $0\le c\le 1$, the expectation value and the variance of estimator (5) are the following:

$$
{}_{B}E\left[\hat{F}(x)\right] = F(x) + hf(0)\int_{-1}^{-c} W(t)dt +
$$

$$
+ h^2 f^{(1)}(0)\left\{\frac{c^2}{2} + c\int_{-1}^{-c} W(t)dt - \int_{-1}^{-c} tW(t)dt\right\} + o\left(h^2\right),
$$

$$
{}_{B}D^2\left[\hat{F}(x)\right] = \frac{1}{n}F(x)(1 - F(x)) - \frac{1}{n}hf(0)\left[\int_{-1}^{c} W^2(t)dt - c\right] + o(h).
$$

It is to note that in boundary region the estimator is not consistent, but variance is of the same order.

The reflection kernel distribution estimator has the form (cf. Horovà et al., 2012):

$$
\hat{F}_R(x) = \frac{1}{n}\sum_{i=1}^{n}\left[W\left(\frac{x - X_i}{h}\right) - W\left(\frac{x + X_i}{h}\right)\right]. \tag{10}
$$

The generalized reflection kernel distribution estimator, improving the bias of the estimator and holding onto low variance, is the following (cf. Karunamuni, Alberts, 2005; Karunamuni, Zhang, 2008):

$$
\hat{F}_{GR}(x) = \frac{1}{n}\sum_{i=1}^{n}\left[W\left(\frac{x - g_1(X_i)}{h}\right) - W\left(\frac{x + g_2(X_i)}{h}\right)\right], \tag{11}
$$

where g_1 and g_2 are cubic polynomials with such coefficients that the bias of the estimator is of order $O\left(h^2\right)$.

In boundary region the expectation value and variance of the estimator (11) are, respectively:

$$
E\left[\hat{F}_{GR}(x)\right] =
$$

$$
= F(x) +
$$

$$
+ h^2\left\{f^{(1)}(0)\left[\frac{c^2}{2} + 2c\int_{-1}^{-c} W(t)dt - \int_{-c}^{c} tW(t)dt\right] - f(0)g_1^{(2)}(0)\int_{-1}^{c}(c - t)W(t)dt - f(0)g_2^{(2)}(0)\int_{-1}^{-c}(c + t)W(t)dt\right\}
$$

$$
+ o\left(h^2\right),
$$

$$
D^2\left[\hat{F}_{GR}(x)\right] = \frac{1}{n}F(x)(1 - F(x)) - \frac{1}{n}hf(0)\left[\int_{-1}^{c} W^2(t)dt - 2\int_{-1}^{c} W(t)W(t - 2c)dt + \int_{-1}^{-c} W^2(t)dt\right] + o(h)\cdot
$$

4. Results of the simulation study

The objective of the simulation study was to compare properties of chosen estimators of the distribution function. The estimators were considered in a special situation when the support of the random variable is bounded. The comparison was made through the graphical representation of the results of the estimation. This form of presenting the estimator is of crucial importance, especially from the user's point of view. The graph provides a fast, comprehensive and readable form of presenting the functional characteristic of the random variable, even for inexperienced users.

In the simulation study the following populations with Weibull distribution $W(0,\delta,\gamma)$ with different scale and shape parameters were examined:

$W1(0,1,0.1)$,

$W2(0,1,0.5)$,

$W3(0,1,1)$,

$W4(0,1,2)$,

$W5(0,1,3.4)$,

$W6(0,1,5)$,

$W7(0,4,1)$,

$W8(0,4,2)$.

The use of a wide range of distribution parameters ensures that varied populations are considered in the study. The difference between populations can be seen, for example, in location, dispersion, asymmetry and kurtosis.

The samples $X_1, X_2, ..., X_n$ of size $n = 10,20,...,100$ were drawn from each population and the following estimators of the distribution function were calculated: empirical distribution function (4), kernel distribution function (5) and reflection kernel distribution function (11). For kernel estimators, Gaussian, Epanechnikov and quartic kernels were used, with Silverman's practical rule (RR), maximal smoothing principle (MSP), plug-in method (PI) and iteration method (IM) used for choosing the smoothing parameter.

Some results for medium size sample $n = 50$ drawn from selected populations, where Epanechnikov kernel and Silverman's rule were used in kernel estimators, are presented in Figures 1-8.

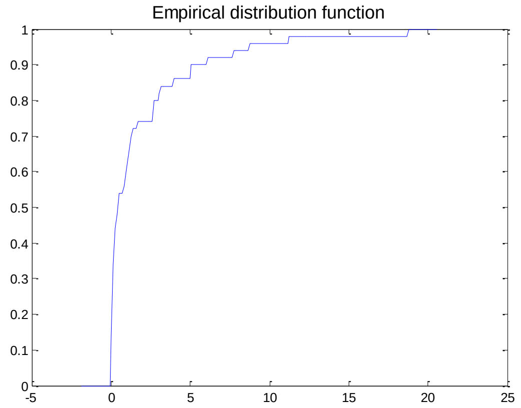
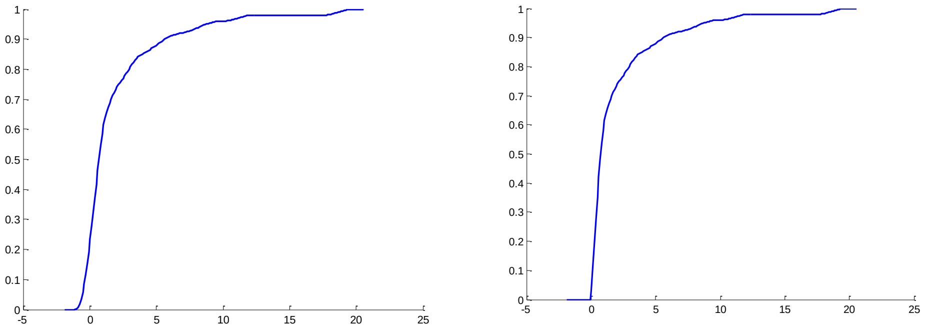


Figure 1. Empirical distribution function, sample size $n = 50$ from $W2(0,1,0.5)$ population



Kernel estimator

Reflection kernel estimator

Figure 2. Kernel distri bution function estimators, sample size $n = 50$ from $W2(0,1,0.5)$ population

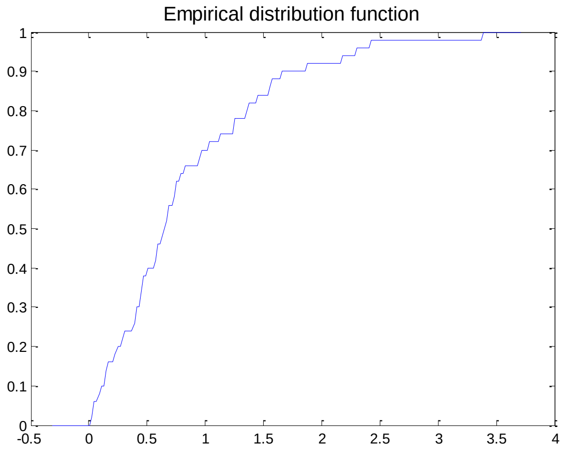
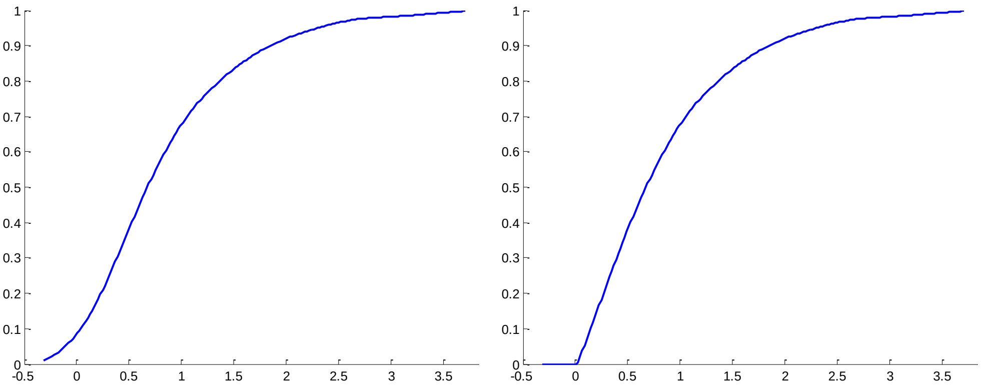


Figure 3. Empirical distribution function, sample size $n = 50$ from $W3(0,1,1)$ population



Kernel estimator

Reflection kernel estimator

Figure 4. Kernel distribution function estimators, sample size $n = 50$ from $W3(0,1,1)$ population

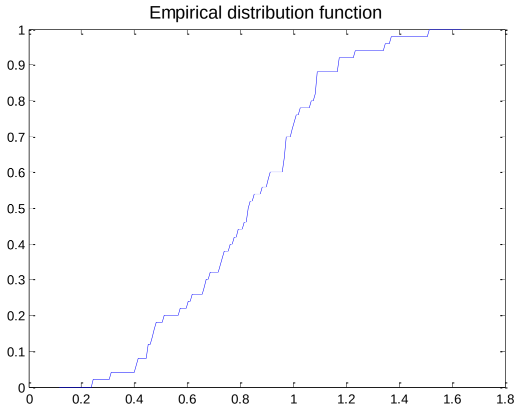
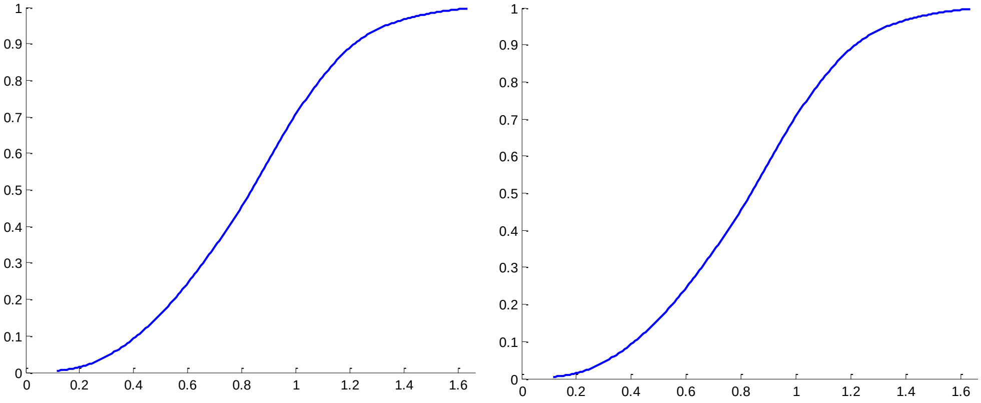


Figure 5. Empirical distribution function, sample size $n = 50$ from $W5(0{,}1{,}3.4)$ population



Kernel estimator

Reflection kernel estimator

Figure 6. Kernel distribution function estimators, sample size $n = 50$ from $W5(0{,}1{,}3.4)$ population

When the group of samples from populations with the same scale parameter but with shape parameter differences (populations $W1(0{,}1{,}0.1)$–$W6(0{,}1{,}5)$) is considered, it can be noticed that the lower the value of shape parameter, the bigger the incompatibility between the support of the random variable and the support of the kernel distribution function estimator. For high values of shape parameters (for example, in populations $W4(0,1,2) - W6(0,1,5)$) the influence of boundary effects is almost imperceptible. When samples are drawn from populations with high values of shape parameters, kernel functions, used in constructing the distribution function estimator in observations near zero, do not extend beyond the support of the random variable.

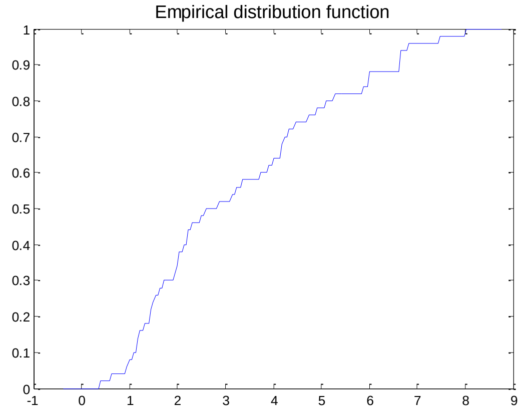
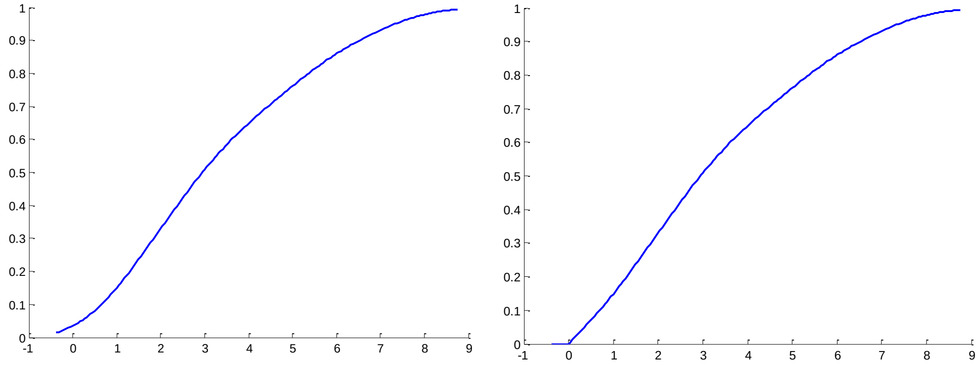


Figure 7. Empirical distribution function, sample size $n = 50$ from $W8(0,4,2)$ population



Kernel estimator

Reflection kernel estimator

Figure 8. Kernel distribution function estimators, sample size $n = 50$ from $W8(0,4,2)$ population

It is worth stressing that when Gaussian kernel function was used in kernel estimator, the boundary effects were bigger, in comparison with other kernel functions. This results directly from the properties of Gaussian kernel which is the only one among the studied kernel functions that has unbounded support.

The kernel distribution function estimators behave in a very similar way, even for very small samples (n=10, n=20). Hence, the sample size is not the essential factor in the occurrence of boundary effects. Taking into account the values of shape parameters and sample sizes, the same results were observed when for the same populations with bounded random variable, the kernel density function estimators were constructed (cf. Baszczyńska, 2015). However, it must be indicated that boundary effects influence the shape of estimators more strongly in the case of density estimator, in some cases even giving the wrong impression of multimodality.

To extend the study, the dependence between kernel function and smoothing parameter was observed. The results are presented in Table 3.

Table 3. Values of smoothing parameters in kernel distribution function estimation for samples from Weibull distribution populations

Population	Kernel function	Method of smoothing parameter choice			
		RR	MSP	PI	IM
$W1(0,1,0.1)$	Epanechnikov	0.8635	0.9224	0.2655	
	quartic	1.0203	1.0899	0.3137	
$W2(0,1,0.5)$	Epanechnikov	1.0958	1.1706	0.3878	0.1872
	quartic	1.2948	1.3832	0.4582	0.2638
$W3(0,1,1)$	Epanechnikov	0.5852	0.6251	0.4817	0.4974
	quartic	0.6914	0.7386	0.5692	0.5699
$W4(0,1,2)$	Epanechnikov	0.3792	0.4051	0.4105	0.4199
	quartic	0.4481	0.4787	0.4851	0.4885
$W5(0,1,3.4)$	Epanechnikov	0.2771	0.2961	0.3523	0.3267
	quartic	0.3275	0.3498	0.4162	0.3858
$W6(0,1,5)$	Epanechnikov	0.2265	0.2419	0.3178	0.2645
	quartic	0.2676	0.2859	0.3756	0.1986
$W7(0,4,1)$	Epanechnikov	3.0632	3.2722	0.7626	1.8332
	quartic	3.6195	3.8666	0.9011	2.1697
$W8(0,4,2)$	Epanechnikov	1.9802	2.1154	0.7755	1.6573
	quartic	2.3399	2.4996	0.9163	2.0212

In the procedure of kernel estimation of cumulative distribution function, two kernel functions: Gaussian and Epanechnikov functions, influence the kernel estimator in a very similar way. The application of these kernel functions is connected with almost the same values of smoothing parameters. It can indicate that Gaussian and Epanechnikov kernels have similar smoothing properties, although they are characterized by different support. When quartic kernel is used, the smoothing parameter is smaller in comparison with other kernel functions.

For samples from populations with small shape parameter, the kernel distribution estimator with smaller smoothing parameters was used. The bigger the shape parameter, the bigger the smoothing parameter in kernel estimation. In general, using Silverman's reference rule ensures smaller values of smoothing parameter. When the shape parameter of population distribution is small, the iterative method is rather poor, the smoothing parameter is unacceptably big, which is denoted by a grey spot in Table 3.

5. Conclusion

The kernel method is an intuitive, simple and useful procedure, especially in density and distribution function estimation. When the support of the random variable is bounded, this procedure needs modification. The modified kernel distribution function estimator ensures that the estimator is consistent, even in boundary region, and the support of the estimator is the same as the support of the random variable being analysed. In kernel method two parameters should be predetermined: kernel function and smoothing parameter. Quartic kernel function indicates higher values of smoothing parameter. Silverman's reference rule, though based on the assumption that the population distribution is normal, gives smaller values of the smoothing parameter.

REFERENCES

BASZCZYŃSKA, A., (2015). Bias Reduction in Kernel Estimator of Density Function in Boundary Region, Quantitative Methods in Economics, XVI, 1.

DOMAŃSKI, C., PEKASIEWICZ, D., BASZCZYŃSKA, A., WITASZCZYK, A., (2014). Testy statystyczne w procesie podejmowania decyzji [Statistical Tests in the Decision Making Process], Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

HÄRDLE, W., (1994). Applied Nonparametric Regression, Cambridge University Press, Cambridge.

LI, Q., RACINE, J. S., (2007). Nonparametric Econometrics. Theory and Practice, Princeton University Press, Princeton and Oxford.

JONES, M. C., (1993). Simple Boundary Correction for Kernel Density Estimation, Statistics and Computing, 3, pp. 135–146.

JONES, M. C., FOSTER, P. J., (1996). A Simple Nonnegative Boundary Correction Method for Kernel Density Estimation, Statistica Sinica, 6, pp. 1005–1013.

KARUNAMUNI, R. J., ALBERTS, T., (2005). On Boundary Correction in Kernel Density Estimation, Statistical Methodology, 2, pp. 191–212.

KARUNAMUNI, R. J., ZHANG, S., (2008). Some Improvements on a Boundary Corrected Kernel Density Estimator, Statistics and Probability Letters, 78, pp. 497–507.

KOLÁČEK, J., KARUNAMUNI, R. J., (2009). On Boundary Correction in Kernel Estimation of ROC Curves, Australian Journal of Statistics, 38, pp. 17–32.

KOLÁČEK, J., KARUNAMUNI, R. J., (2012). A Generalized Reflection Method for Kernel Distribution and Hazard Function Estimation, Journal of Applied Probability and Statistics, 6, pp. 73–85.

KULCZYCKI, P., (2005). Estymatory jądrowe w analizie systemowej [Kernel Estimators in Systems Analysis], Wydawnictwa Naukowo-Techniczne, Warszawa.

HOROVÀ, I., KOLÁČEK, J., ZELINKA, J., (2012). Kernel Smoothing in MATLAB. Theory and Practice of Kernel Smoothing, World Scientific, New Jersey.

SILVERMAN, B.W., (1996). Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

TENREIRO, C., (2013). Boundary Kernels for Distribution Function Estimation, REVSTAT Statistical Journal, 11, 2, pp. 169–190.

TENREIRO, C., (2015). A Note on Boundary Kernels for Distribution Function Estimation, http://arxiv.org/abs/1501.04206 [14.11.2015].

WAND, M. P., JONES, M.C., (1995). Kernel Smoothing, Chapman and Hall, London.

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 557–574

IN SEARCH OF HEDGES AND SAFE HAVENS IN GLOBAL FINANCIAL MARKETS

Stanisław Wanat[1], Sławomir Śmiech[2], Monika Papież[3]

ABSTRACT

The aim of the paper is to search for hedges and safe havens within three instrument classes: assets (represented by the S&P500 index), gold and oil prices, and dollar exchange rates. Weekly series of returns of all the instruments from the period January 1995 – June 2015 are analysed. The study is based on conditional correlations between the instruments in different market regimes obtained with the use of copula-DCC GARCH models. It is assumed that different market regimes will be identified by statistical clustering techniques; however, only conditional variances (without conditional covariances) will be taken into account. The reason for this assumption is connected with the fact that variances can be understood as market risk, and, as such, are a good indicator of market conditions. A considerable advantage of such an approach is the lack of need to determine the number of market regimes, as it is established by clustering quality measures. What is more, the methodology used in the paper makes it possible to treat the relations between instruments symmetrically. The results obtained in the study reveal that only dollar exchange rates can be treated as a (strong) hedge and a (strong) safe haven for other instruments, while gold and oil are a hedge for assets.

Key words: market regimes, clustering methods, copula, DCC-GARCH.

1. Introduction

The global financial crisis caused a dramatic decline in stock prices almost simultaneously in the stock markets worldwide. Portfolios based on public equity and other financial instruments depreciated at a rapid rate. The economic downturn, which was the result (or the symptom) of the crisis, hampered the demand for commodities. The prices of oil, other energy sources, food and most metals (excluding gold) dropped. Some experts claim that, in anticipation of

[1] Cracow University of Economics. E-mail: wanats@uek.krakow.pl.
[2] Cracow University of Economics. E-mail: smiechs@uek.krakow.pl.
[3] Cracow University of Economics. E-mail: papiezm@uek.krakow.pl.

decreases in prices, investors purchased gold, which seemed an attractive instrument during a crisis due to various reasons (a liquid instrument, a means of payment in the past).

Baur and Lucey (2010) were the first to formalize and measure the role of gold as an instrument used as a substitute for assets. They defined two categories of instruments: a hedge (a strong hedge) and a safe haven (a strong safe haven). An instrument is defined as a hedge (a strong hedge) for assets if its rates of return are uncorrelated or negatively correlated with their rates of return. An instrument is defined as a safe haven (a strong safe haven) if its returns are uncorrelated or negatively correlated with the returns of other assets in times of market stress or turmoil. Baur and Lucey (2010) study the role of the price of gold (its rate of return) with reference to stock market indexes and bonds in the USA, Great Britain and Germany. On the basis of the analyses of daily returns in the period between 30 November 1995 and 30 November 2005, they conclude that gold is a hedge for assets in the USA and Great Britain and for bonds in Germany, and a safe haven for assets in these three countries.

The aim of the paper is to investigate the possibilities of a hedge and a safe haven within three instrument classes: assets (represented by the S&P500 index), gold and oil prices, and effective dollar exchange. As it is done in studies by Joy (2011), Reboredo (2013a), and Reboredo (2013b), in this paper weekly series of returns from the period between January 1995 and June 2015 are analysed with the use of conditional correlations between instruments for different market regimes obtained from copula-DCC GARCH models. In the study it is assumed that the identification of different markets regimes will be conducted with the use of statistical clustering methods, although only conditional variances will be taken into account. The reason for such an assumption is connected with the fact that variances can be understood as market risk, thus the change (growth) of risk (variance) is a good (and classic) indicator of the condition of financial markets. One of the advantages of our approach is the lack of need to a priori determine the number of market regimes, as it is established by clustering quality measures. What is more, the methodology used in the paper allows for treating the relations between the instruments symmetrically. As a result, it is possible to establish the role of particular instruments in different market regimes. To the best of our knowledge, such an approach has not been used in the analyses of hedges and safe havens so far.

The paper is organised as follows. The second chapter reviews the subject literature devoted to identification of financial instruments which could be used as hedges and safe havens, the third chapter presents the methodology and the empirical strategy used in the paper, the fourth one presents the data and discusses the results obtained, while the fifth chapter contains the conclusions.

2. Literature review

Numerous papers published in recent years address the question whether gold is actually a good haven for the investment portfolio and whether there are other instruments which could play this role. Most authors adopt Baur and Lucey's (2010) definition of a hedge and a safe haven, who verify the role of gold with reference to stock market indexes and bonds in the USA, Great Britain and Germany. In subsequent papers the set of analysed instruments and research methodology have been expanded, thus they can be divided into four distinct groups.

Studies from the first group consider possibilities of using gold to hedge portfolios consisting of assets and bonds. Apart from the already mentioned work by Baur and Lucey (2010), the following papers use this approach:

(i) Baur and McDermott's (2010) paper, in which the analysis of daily, weekly and monthly rates of return in the period between 1979 and do 2009 lead to the conclusion that gold is a hedge and a safe haven for assets in European countries and in the USA, but is neither a hedge nor a safe haven for assets in developing countries and in Australia, Canada and Japan,

(ii) Beckmann, Berger and Czudaj's (2015) paper, in which it is concluded, on the basis of the analysis of monthly rates of return in the period between 1970 and 2012, that: gold is a strong hedge for assets in the euro area, Indonesia, Russia and Turkey; gold is not a hedge for assets in China and Germany; gold is a hedge for assets in the remaining economies; gold is a strong safe haven for assets in India and Great Britain and is not a hedge in the euro area, Indonesia, and Russia.

The second group includes studies which investigate the role of gold as a hedge for foreign currency portfolios. The following two papers can serve as an example here:

- Joy's (2011) paper, which reveals, on the basis of weekly data of 16 exchange rates in the period between 10 January 1986 and 29 August 2008, that gold is a hedge and a weak hedge only for US dollar exchange rate (it is not a hedge for exchange rates of other currencies),
- Reboredo's (2013b) paper, which confirms, on the basis of weekly data of 8 exchange rates in the period between 7 January 2000 and 21 September 2012, the results obtained by Joy (2011) that gold is a hedge only for US dollar exchange rate.

The papers from the third group examine gold as a hedge for commodity markets. Reboredo's (2013a) paper may serve as an example here. On the basis of weekly data from the period between 7 January 2000 and 30 September 2011, this paper concluded that gold is not a hedge for oil prices, but a safe haven for them.

The fourth group comprises papers which search for other hedges than gold, e.g.:

- Hood and Malik's (2013) paper, in which daily rates of return are analysed in the period between November 1995 and November 2010, and the findings reveal that gold and VIX are a hedge and a safe haven for assets in the USA, while other precious metals are neither a hedge nor a safe haven.
- Ciner, Gurdgiev and Lucey's (2013) paper, which concludes, on the basis of daily data from the period 1990-2010, that: gold is a hedge and a safe haven for dollar exchange rates and for British pound exchange rates (since 2000); gold is a safe haven for assets and bonds in the USA; bonds are a safe haven for assets in the USA; oil is a safe haven for bonds in the USA, assets listed on the British stock exchange are a safe haven for pound exchange rate and for oil; British bonds are a safe haven for assets in Great Britain; pound exchange rate is a safe haven for assets and bonds in Great Britain and for gold.

The analysis of gold as a hedge is conducted with the use of various instruments, which allow for identifying 'normal' and 'turmoil' market regimes. In their seminal paper, Baur and Lucey (2010) use an autoregressive distributed lag model including different dummies for indicating lower quantiles of any instruments of interest. Joy (2011), Ciner, Gurdgiev and Lucey (2013) use a DCC-GARCH model, while Reboredo (2013a), and Reboredo (2013b) use the copula function and dependencies in the tails of distribution to define the relations in turmoil. Beckmann, Berger and Czudaj (2015) use two-regime threshold model (smooth transition regression), in which one regime corresponds to normal market conditions, while the other to a market in crisis.

3. Methodology

The aim of the empirical strategy used in the study is to investigate the possibility of using three categories of financial instruments – assets (represented by the S&P500 index), commodities (gold and oil) and US dollar rate as hedges and safe havens. It consists of two stages:

- identification of market regimes,
- the analysis of correlations between the instruments in different market regimes.

During the first stage it is assumed that market regimes will be identified with the use of statistical clustering methods of weekly periods *t* according to conditional variances of returns of all analysed instruments. This assumption is based on another assumption that variance growth is a good (and classic) indicator of financial market conditions. The market regime with the highest variance is used to identify instruments which can be treated as safe havens. Conditional variances are obtained on the basis of four-dimensional copula DCC-GARCH model.

A copula-based multivariate GARCH model used in this study allows for modelling the conditional dependence structure when standardized innovations are non-elliptically distributed. Thus, it makes it possible to model the volatility of particular financial instruments using univariate GARCH models with different standardized residual distribution. Generally, copulas allow the researcher to specify the models for the marginal distributions separately from the dependence structure that links these distributions to form a joint distribution. They offer a greater flexibility in modelling and estimating margins than in the case of using parametric multivariate distributions (see, e.g. Nelsen, 1999; Joe, 1997). Secondly, at present the copula-GARCH methodology is widely used in the analysis of financial time series (see, e.g. Patton, 2006; Serban et al., 2007; Lee and Long, 2009; Doman, 2011; Wu et al., 2012; Aloui et al., 2013; Li and Yang, 2013; Zolotko and Okhrin, 2014; and for a review Patton, 2012).

In the copula-GARCH model, multivariate joint distributions of the return vector $r_t = (r_{1,t},...,r_{k,t})'$, $t = 1,...,T$ conditional on the information set available at time $t-1$ (Ω_{t-1}) is modelled using conditional copulas introduced by Patton (2006). This model takes the following form:

$$r_{1,t} \mid \Omega_{t-1} \sim F_{1,t}(\cdot \mid \Omega_{t-1}),...,r_{k,t} \mid \Omega_{t-1} \sim F_{k,t}(\cdot \mid \Omega_{t-1}) \tag{1}$$

$$r_t \mid \Omega_{t-1} \sim F_t(\cdot \mid \Omega_{t-1}) \tag{2}$$

$$F_t(r_t \mid \Omega_{t-1}) = C_t\left(F_{1,t}(r_{1,t} \mid \Omega_{t-1}),...,F_{k,t}(r_{k,t} \mid \Omega_{t-1}) \mid \Omega_{t-1}\right) \tag{3}$$

where C_t denotes the copula, while F_t and $F_{i,t}$ denote the joint cumulative distribution function and the cumulative distribution function of the marginal distributions at time *t*, respectively.

In a general case, univariate rates of return $r_{i,t}$ can be modelled by various specifications of the mean model by using the ARIMA process and various specifications of the variance model (e.g. sGARCH, fGARCH, eGARCH, gjrGARCH, apARCH, iGARCH and csGARCH). In the study, for all series of returns, the following ARIMA process is applied:

$$r_{i,t} = \mu_{i,t} + y_{i,t}, \tag{4}$$

$$\mu_{i,t} = E\left(r_{i,t} \mid \Omega_{t-1}\right), \quad \mu_{i,t} = \mu_{i0} + \sum_{j=1}^{P_i} \varphi_{ij} r_{i,t-j} + \sum_{j=1}^{Q_i} \theta_{ij} y_{i,t-j}, \tag{5}$$

$$y_{i,t} = \sqrt{h_{i,t}}\, z_{i,t}, \tag{6}$$

variance for one series is modelled with the use of a standard GARCH model (sGARCH):

$$h_{i,t} = Var(r_{i,t} \mid \Omega_{t-1}), \quad h_{i,t} = \omega_i + \sum_{j=1}^{p_i} \alpha_{ij} y_{i,t-j}^2 + \sum_{j=1}^{q_i} \beta_{ij} h_{i,t-j}, \tag{7}$$

and for the remaining three series variance is modelled with the use of an exponential GARCH model (eGARCH) (Nelson, 1991):

$$\log(h_{i,t}) = \omega_i + \sum_{j=1}^{p_i} \left(\alpha_{ij}\varepsilon_{i,t-j} + \gamma_{ij}\left(\left|\varepsilon_{i,t-j}\right| - E\left|\varepsilon_{i,t-j}\right|\right)\right) + \sum_{j=1}^{q_i} \beta_{ij} \log(h_{i,t-j}), \; \varepsilon_{i,t} = \frac{y_{i,t}}{\sqrt{h_{i,t}}}, \tag{8}$$

where $z_{i,t}$ are i.i.d. random variables which conditionally follow some distributions with the required properties (in the empirical analysis the following distributions are considered: normal distribution, skew-normal distribution, student-t, skew-student, generalized error distribution).

The dependence structure of the margins is then assumed to follow an elliptical copula with conditional correlations R_t. The dynamics of R_t is modelled with the use of the dynamic conditional correlation model DCC(*m*, *n*):

$$H_t = D_t R_t D_t, \tag{9}$$

$$D_t = diag(\sqrt{h_{1,t}}, \dots, \sqrt{h_{k,t}}), \tag{10}$$

$$R_t = \left(diag(Q_t)\right)^{-1/2} Q_t \left(diag(Q_t)\right)^{-1/2}, \tag{11}$$

$$Q_t = \left(1 - \sum_{j=1}^{m} c_j - \sum_{j=1}^{n} d_j\right)\overline{Q} + \sum_{k=1}^{m} c_j(\varepsilon_{t-j}\varepsilon'_{t-j}) + \sum_{k=1}^{n} d_j Q_{t-j}, \tag{12}$$

where conditional variances $h_{i,t}$ are modelled with the use of one-dimensional GARCH(*p*,*q*) processes (7) or (8), $\varepsilon_t = D_t^{-1} y_t$ ($y_t = (y_{1,t}, \dots, y_{k,t})'$) and $\overline{Q}$ is unconditional covariance matrix of standardized residuals ε_t. In specification (12) $c_j\ (j = 1,\dots,m)$, $d_j\ (j = 1,\dots,n)$ are scalars which capture the effect of previous shocks and previous dynamic correlation on the current conditional correlation respectively.

The parameters of the above copula-DCC-GARCH model are assessed using the inference function for margins (*IFM*) approach (this method is described in detail in the works by (e.g.): (Joe, 1997, pp. 299–307; Doman, 2011, pp. 35–37; Wanat, 2012, pp. 98-99)). Calculations have been done in the R package (”rmgarch” ,version 1.2-6) developed by Alexios Ghalanos.

To identify financial market regimes, statistical methods of unsupervised classification are used. The groups obtained are expected to be periods with a similar level of risk (i.e. similar conditional variance). Although the number of groups is not known a priori, it is assumed that it should be neither too small nor too large. In fact, clustering results are assessed taking into account both statistical criteria and economic interpretation of financial market regimes obtained. Clustering is conducted by means of hierarchical methods in which groups are created recursively by linking together the most similar objects (Ward's method is applied here). Other methods of division, i.e. the k-means method and the partitioning among medoids (PAM) method proposed by Kaufman and Rousseeuw (1990) are also used. In both cases, after making the initial decision about the desired number of groups, objects are allocated in such a way that the relevant criterion is met. For the k-means method the allocation of objects should minimize a within-group variance. In the PAM method the representatives of groups (medoids) are selected at each step of the analysis, and then the remaining objects are allocated to the group which includes the closest medoid. The former method is more robust to outliers than the k-means method, because it minimizes the sum of dissimilarities instead of the sum of squared Euclidean distance. In order to evaluate the optimal number of clusters in the data, we use internal validity indexes: Calinski Harabasz pseudo F statistics (Calinski and Harabasz, 1974), the average silhouette width (Kaufman and Rousseeuw, 1990), the Dunn index (Dunn, 1974), and Xie and Beni's (1991) index. The final classification of objects is, therefore, the result of the comparison of the results of respective grouping algorithms.

During the second stage of the study conditional correlations between different markets regimes obtained from the copula-DCC-GARCH method described above are analysed. In accordance with the definition adopted, the negative correlation (the lack of correlation) in the regime with 'normal volatility' signifies a strong hedge (a hedge), and the negative correlation (the lack of correlation) in the regime with considerably higher volatility signifies a strong safe haven (a safe haven).

4. Data and empirical results

The dataset consists of variables which represent equity, currency, gold and oil markets. Equity is represented by the S&P500 index (SP500), exchange rates are represented by the Federal Reserve Bank's Nominal Trade Weighted Effective Index (USD_B), the price of gold is represented by the gold futures contracts traded on the New York Mercantile Exchange (NYMEX) and it based in US dollars per troy ounce (GOLD), the price of crude oil is represented by contracts of crude oil futures traded on the NYMEX and it based in US dollars per barrel (WTI). The study is based on weekly data, the sample period ranges from January

1995 to June 2015 and comprises 1069 observations per variable. Weekly logarithmic rates of return are analysed, and the descriptive statistics for index returns, exchange rate returns, crude oil returns and gold returns are reported in Table 1. In the analysed period all instruments increase their values, which is visible in positive means of returns and positive medians of returns. WTI rates of return are characterised by the greatest volatility, and USD_B rates of return are characterised by the smallest volatility. Dollar exchange rate is the only instrument with positive asymmetry, while WTI and SP500 have a considerable negative asymmetry.

Table 1. Descriptive statistics for returns

	S&P500	USD_B	GOLD	WTI
Mean	0.141	0.019	0.106	0.110
Median	0.365	0.001	0.116	0.244
Max	8.308	3.618	12.808	15.054
Min	-15.279	-2.841	-11.827	-19.561
Std. Dev.	1.962	0.563	1.963	3.800
Skewness	-0.890	0.348	-0.075	-0.523
Kurtosis	8.709	6.277	7.768	4.980

Source: Author's own calculation.

During the first stage of the study market regimes are identified with the use of the copula-DCC GARCH model which yields conditional variances of returns of the analysed instruments. In the empirical study different variants of the ARMA-GARCH specifications are considered for individual returns. On the basis of information criteria and tests of model adequacy (the results can be obtained from the authors on request), the following models have been selected: ARMA(1,1)-eGARCH(2,2) for S&P500 and gold, ARMA(1,1)-eGARCH(1,1) for US dollar exchange rate and ARMA(1,1)-sGARCH(1,1) for oil. In all models for standardized residuals the skewed Student's t distributions (with skew and shape parameters ξ and υ respectively) are assumed. On the other hand, Gauss and Student's t copulas have been considered in the analysis of the dynamics of dependencies between the rates of return, and, also on the basis of information criteria, Student's t with conditional correlation and constant shape parameter η have been chosen. Table 2 presents the results of estimation, and Figure 1 the estimated conditional variances.

Table 2. Copula-DCC–GARCH model estimation results

	SP500	USD_B	GOLD	WTI
GARCH Model	eGARCH(2,2)	eGARCH(1,1)	eGARCH(2,2)	sGARCH(1,1)
Mean Model	ARMA(1,1)	ARMA(1,1)	ARMA(1,1)	ARMA(11)
Distribution	Skewed Student's	Skewed Student's	Skewed Student's	Skewed Student's
Parameters of univariate models				
μ	0.12185 (0.00991)	0.02937 (0.16007)	0.04532 (0.41810)	0.09392 (0.44854)
φ_1	-0.17052 (0.01358)	0.20266 (0.11964)	-0.11796 (0.08054)	0.10080 (0.69841)
θ_1	0.32150 (0.00000)	0.08406 (0.53949)	0.34913 (0.00000)	0.09081 (0.72951)
ω	0.05288 (0.00037)	-0.05176 (0.00700)	0.04854 (0.02876)	0.24883 (0.05896)
α_1	-0.31312 (0.00000)	0.04231 (0.03079)	0.08058 (0.01535)	0.07009 (0.00001)
α_2	0.13897 (0.00375)	-	-0.01038 (0.76281)	-
β_1	1.00000 (0.00000)	0.96169 (0.00000)	0.13882 (0.00000)	0.91452 (0.00000)
β_2	-0.05326 (0.00163)	-	0.81050 (0.00000)	-
γ_1	0.13848 (0.07462)	0.20927 (0.00000)	0.26303 (0.00000)	-
γ_2	0.04129 (0.60239)	-	0.14724 (0.00549)	-
ξ (skew)	0.73460 (0.00000)	1.08361 (0.00000)	0.94755 (0.00000)	0.86033 (0.00000)
υ (shape)	18.48111 (0.04565)	15.36894 (0.03426)	7.50841 (0.00000)	9.46780 (0.00005)
Copula-DCC parameters				
Distribution	Four-dimensional t-Student			
DCC Order	DCC(1.1)			
	Parametry			
c_1	0.021553 (0.00069)			
d_1	0.970817 (0.00000)			
η (mshape)	15.096596 (0.00001)			

Probability values (p-values) are in parentheses.

Source: Author's own calculation.

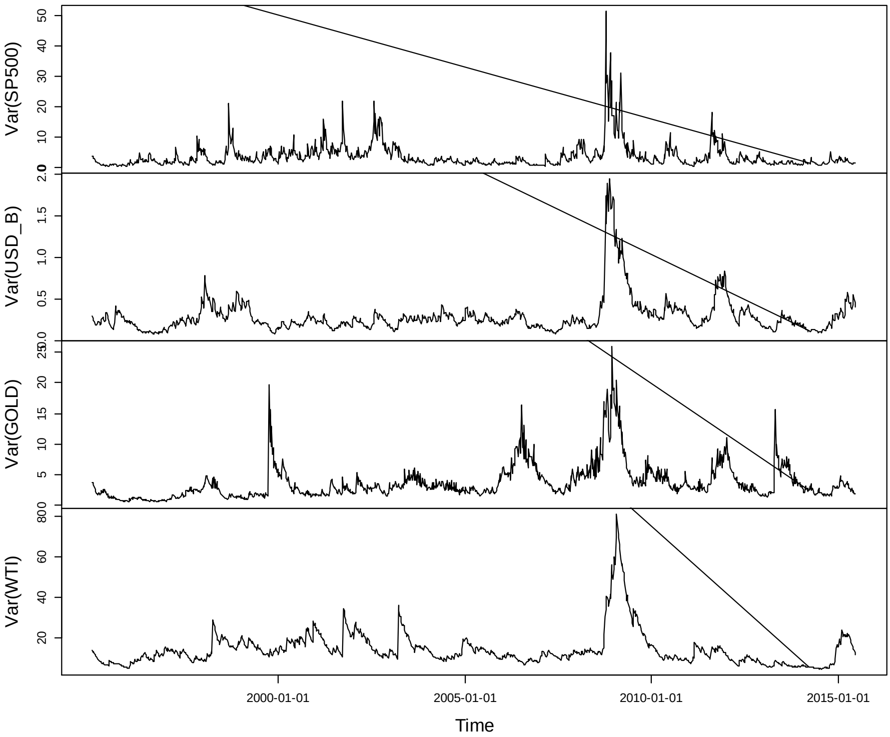


Figure 1. Conditional variances.

During the second stage conditional variances obtained with the use of Copula-DCC–GARCH model are clustered. Precisely, the moments of time characterised by four-dimensional vectors of conditional variances obtained for particular instruments are clustered. This procedure is supposed to indicate the periods in which financial markets are characterised by similar levels of risk.

Clustering is conducted with the use of three methods: Ward's method, the k-means method and the partitioning among medoids (PAM). The assessment of the quality of clustering is presented in Table 3, and index validation clustering is calculated with the assumption that the number of groups in not smaller than 2 and not larger than 6. The majority of measures, regardless of the clustering method applied, reveal that the division into two clusters is optimal (the highest values of the Silhouette and Dunn index, the lowest value of the Xie-Beni index). The Silhouette index clearly indicates that the best division is obtained for the k-means method (the average silhouette width 0.7782 for 2 clusters, and 0.4741 for three clusters). The Calinski Harabasz index indicates that the optimal division consists of three clusters when Ward's method and the k-means method are used, and of four clusters when PAM is used. The Dunn index and Xie-Beni index indicate the division into five groups when PAM is used. Taking into consideration a possibility of multi-faceted interpretation of clustering results obtained with the use of the Silhouette index, that is a possibility of assessing the objects (here moments) which belong to a given group (regime), it has been decided that in further analysis clusters which have been obtained by the k-means methods will be used. The division into two and three groups is analysed in this case[4]. After comparing the elements of clusters created for the division into two and three groups, it turns out that all elements from a less numerous (consisting of 29 elements) group obtained after the division into two clusters constitute a single cluster obtained after the division into three clusters. What is interesting is that this cluster includes only the periods at the beginning of 2009, that is the moment of the collapse of the commodity market. The more numerous group (obtained after the division into two clusters) is divided into two separate groups.

Table 3. Validation indices for data partitions.

Validation criterion	Number of clusters				
	2	3	4	5	6
			Ward's method		
Silhouette	**0.4144**	0.4132	0.2420	0.2468	0.2760
Calinski Harabasz index	525.5172	**1128.4202**	891.4865	888.8947	845.2170
Dunn index	0.0074	**0.0092**	0.0070	0.0070	0.0070
Xie-Beni index	**247.1301**	118.3295	181.6336	146.8908	128.1764
			k-means		
Silhuette	**0.7782**	0.4741	0.3431	0.3232	0.3574
Calinski Harabasz index	1112.4198	**1205.3381**	1101.2847	933.7506	919.3722
Dunn index	**0.0229**	0.0065	0.0027	0.0050	0.0050
Xie-Beni index	**28.9347**	227.7294	1026.4980	274.5502	232.5727
			PAM		
Silhuette	**0.4637**	0.2568	0.3236	0.3285	0.2934
Calinski Harabasz index	609.9561	463.9710	**1075.3220**	962.8447	909.4470
Dunn index	0.0038	0.0042	0.0049	**0.0051**	0.0034
Xie-Beni index	938.7358	684.6099	317.8286	**261.2086**	505.2739

Source: *Own calculations performed with the use of the 'clusterSim' package developed by M. Walesiak and A. Dudek (the Silhouette and Calinski Harabasz index) and the 'clusterCrit' package developed by Bernard Desgraupes (the Dunn and Xie-Beni index).*

Note: numbers in bold indicate the optimal number of groups with reference to a given criterion.

[4] The division into 3 groups is attractive as it offers a possibility of a sensible economic interpretation. Market regimes identified in the study could be described as: regimes of a low, moderate and high volatility.

In further analysis it is assumed that different market regimes correspond to different classes. Conditional variances in different market regimes are demonstrated in Figure 4. It reveals that the first regime, which occurs only during the greatest turbulences in global markets at the beginning of 2009, is characterised by a high volatility (a high risk level), the second regime – by a heightened volatility (a moderate risk level[5]) and the third regime – by a low volatility of instruments (a low risk level). Additional information on the clustering quality from Figure 2 leads to the conclusion that cohesion and separation are quite similar for the three clusters considered. What is more, if period *t* belongs to the first regime, its neighbour belongs to the second regime, and if period *t* belongs to the second regime, its neighbour belongs to the first regime. During a crisis regimes with moderate and high risks are neighbours, while regimes with low and high risks are never neighbours. These results strongly support the decision to apply the definition of a safe haven to the third regime and the definition of a hedge to the first and second regimes.

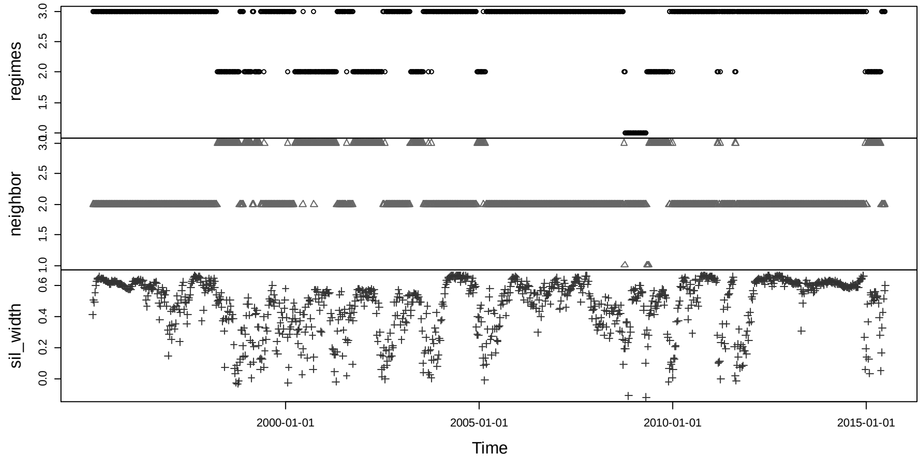


Figure 2. Market regimes

During the second stage of the study conditional correlations between instruments in the whole period between January 1995 and June 2015 and in different market regimes are analysed. It should be mentioned here that the results of testing for parameter constancy indicate strong evidence against the assumption of constant conditional correlations: the test, developed by Engle and Sheppard (2001), uses a χ^2-statistic to test the null of $R_t = R$. The resulting test statistics, 50.022 (p-value =0.0000), is highly significant, rejecting the null hypothesis of constant conditional correlations. Fig. 5 demonstrates the dynamic conditional

[5] The only exception is gold for which conditional variances in the first and in the second regime seem to be similar.

correlations between instruments. Generally, negative correlations between US dollar exchange rate and other instruments and positive correlations between gold and oil can be observed. Correlations between the S&P500 index and commodities (gold, oil) in certain periods are positive and in other periods are negative.

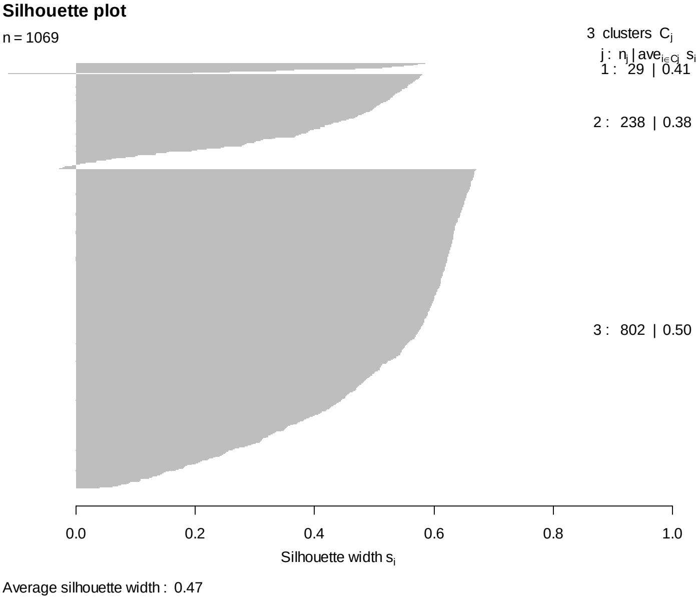


Figure 3. Silhouette plot

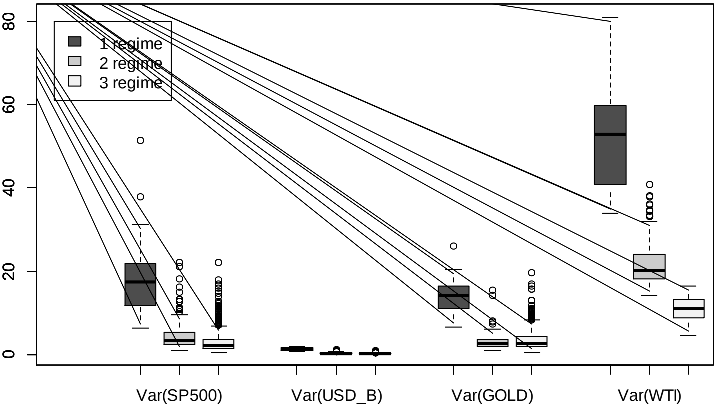


Figure 4. Distribution of variance in different regimes

The analysis of correlations in different market regimes allows for identification of the instruments which can serve as a hedge and a safe haven for other financial instruments. It has been assumed that the definition of a safe haven refers to the first regime, characterised by the highest volatility (the greatest risk), and the definition of a hedge refers to two other regimes (the second and third regime). The distribution of correlations in different regimes is demonstrated in Figure 6, and, on its basis, it can be concluded that:

1. Correlations between instruments are not the same in all market regimes.
2. Differences in correlations in the first regime are much smaller than in the remaining two regimes, which means that in the turmoil periods, relations are more stable than in other two regimes (correlations remain at a similar level).
3. The greatest differences between correlations for particular pairs can be found in the first regime.
4. The level of correlation in the first regime considerably differs from the level in the second and third regimes.
5. Correlations in the first regime do not change the sign in comparison with the second and third regimes and are considerably stronger, except for correlations between the S&P500 index and gold.
6. The level of correlation is similar between particular instruments in the second and third regime, that is the ones with a moderate and low risk levels respectively.

Taking into consideration definitions of a hedge and a safe haven adopted in this study, the following conclusions can be drawn:

(i) Dollar exchange rate is negatively correlated with other instruments in all market regimes, throughout the majority of the analysed period, thus it can be treated as a (strong) hedge and a (strong) safe haven for the remaining instruments.

– Oil is weakly correlated with the S&P500 index in the second and third regimes, so it can be a hedge for assets.

– Gold is weakly correlated with the S&P500 index in all regimes, so it can be treated as a hedge and a safe haven for assets.

– Oil is only a hedge for assets in 'normal' conditions.

– Both commodities (gold and oil) are weakly correlated with assets in the regimes with low and moderate volatility. However, only gold remains uncorrelated with assets in the regime with the highest volatility. It means that gold and oil are a hedge for assets, but only gold is a safe haven for assets.

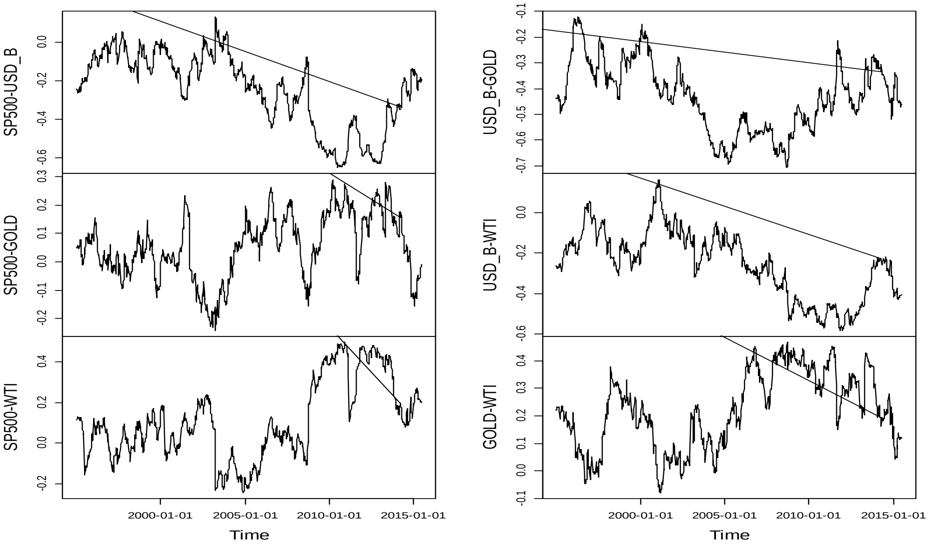


Figure 5. Dynamic conditional correlations

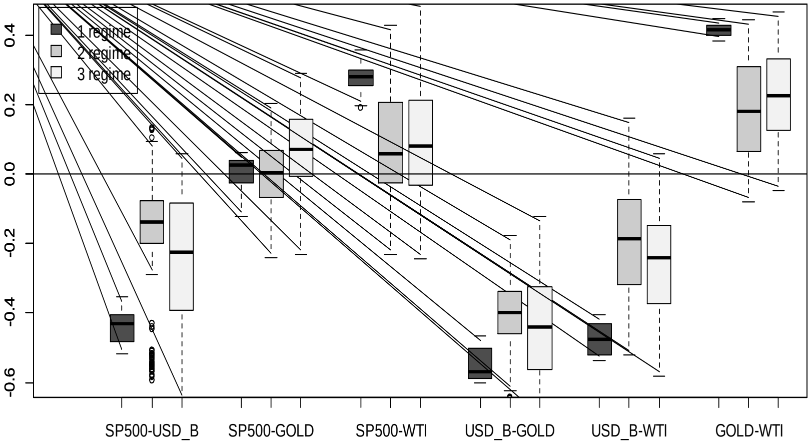


Figure 6. Distribution of correlation in different market regimes

5. Conclusions

The aim of the study was to identify instruments which can serve as a hedge and a safe haven for other financial instruments. Three classes of instruments were taken into consideration: US dollar exchange rate, the S&P500 index, and the prices of two commodities, gold and oil. The empirical strategy applied in the study consisted of two steps: in the first one market regimes were identified and in the second one correlations between instruments in different market regimes were analysed. Market regimes were identified by analysing the risk (volatility) of the instruments. Both statistical criteria of clustering and a possibility of a sensible economic interpretation indicate three different market regimes in the analysed period: a regime of low volatility of instruments, a regime of heightened volatility of instruments and a regime of the highest volatility of financial instruments (which occurred only in the period of the greatest turbulences in global markets at the beginning of 2009). In the second step it was decided that the definition of a safe haven would refer to the regime with the highest volatility, and the definition of a hedge would refer to the two remaining regimes. This allowed for differentiating mutual correlations between instruments in market regimes with low and moderate volatility, which had not been done in the subject literature before. The distribution of correlations obtained for different market regimes justifies drawing the following conclusions.

Firstly, correlations between instruments are not the same in all market regimes. The greatest differences are observed when comparing correlations in the regime with the highest volatility and in two remaining regimes. Mutual correlations in the regimes with low and moderate volatility are similar. What is interesting is that in the period of the highest volatility correlations are usually (with the exception of the pair S&P500-GOLD) higher (per module), but have the same sign as in the two remaining market regimes. Secondly, only dollar exchange rate is negatively correlated with other instruments, thus it can be treated as a (strong) hedge and as a (strong) safe haven for other instruments. Thirdly, both commodities (gold and oil) are weakly correlated with assets in the regimes of low and moderate volatility. However, only gold remains uncorrelated with assets in the regime with the highest volatility. Similar results for gold were obtained by Baur and Lucey (2010) and Baur and McDermott (2010).

Acknowledgements

Supported by the grant No. 2012/07/B/HS4/00700 of the Polish National Science Centre.

REFERENCES

ALOUI, R., BEN AÏSSA, M. S., NGUYEN, D. K., (2013). Conditional dependence structure between oil prices and exchange rates: A copula-GARCH approach. Journal of International Money and Finance, 32, pp. 719–738.

BAUR, D. G., MCDERMOTT, T. K., (2010). Is gold a safe haven? International evidence. Journal of Banking & Finance, 34(8), pp. 1886–1898.

BAUR, D. G., LUCEY, B. M., (2010). Is gold a hedge or a safe haven? An analysis of stocks, bonds and gold. Financial Review, 45(2), pp. 217–229.

BECKMANN, J., BERGER, T., CZUDAJ, R., (2015). Does gold act as a hedge or a safe haven for stocks? A smooth transition approach. Economic Modelling, 48, pp. 16–24.

CALINSKI, T., HARABASZ, J., (1974). A dendrite method for cluster analysis. Communications in Statistics, 3 (1), pp. 1–27.

CINER, C., GURDGIEV, C., LUCEY, B. M., (2013). Hedges and safe havens: An examination of stocks, bonds, gold, oil and exchange rates. International Review of Financial Analysis, 29, pp. 202–211.

DESGRAUPES, B., (2015). Clustering Indices, package ‘clusterCrit’, https://cran.r-project.org/web/packages/clusterCrit.

DOMAN, R., (2011). Zastosowanie kopuli w modelowaniu dynamiki zależności na rynkach finansowych [The use of copulas in modelling the dynamic dependence on the financial markets], Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań.

DUNN, C. J., (1974). Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4, pp. 95–104.

ENGLE, R.F. AND SHEPPARD, K., (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH, NBER Working Paper.

HOOD, M., MALIK, F., (2013). Is gold the best hedge and a safe haven under changing stock market volatility? Review of Financial Economics, 22 (2), pp. 47–52.

JOE, H., (1997). Multivariate models and dependence concepts. Chapman-Hall, London.

JOY, M., (2011). Gold and the US dollar: Hedge or haven? Finance Research Letters, 8 (3), pp. 120–131.

KAUFMAN, L., ROUSSEEUW, P. J., (1990). Finding Groups in Data: An Introduction to Cluster Analysis, New York: Wiley & Sons.

LEE, T.-H., LONG, X., (2009). Copula-Based Multivariate GARCH Model with Uncorrelated Dependent Errors. Journal of Econometrics, 150, pp. 207–218.

LI, M., YANG, L., (2013). Modeling the volatility of futures return in rubber and oil – A Copula-based GARCH model approach. Economic Modelling, 35, pp. 576–581.

NELSEN, R. B., (1999). An Introduction to Copulas. Springer-Verlag, New York.

NELSON, D. B., (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. Econometrica, 59, pp. 347–370.

PATTON, A. J., (2006). Modelling asymmetric exchange rate. International Economic Review, 47, pp. 527–556.

PATTON, A. J., (2012). A review of copula models for economic time series. Journal of Multivariate Analysis, 110, pp. 4–18.

REBOREDO, J. C., (2013a). Is gold a hedge or safe haven against oil price movements? Resources Policy, 38(2), pp. 130–137.

REBOREDO, J. C., (2013b). Is gold a safe haven or a hedge for the US dollar? Implications for risk management. Journal of Banking & Finance, 37(8), pp. 2665–2676.

SERBAN, M., BROCKWELL, A., LEHOCZKY, J., SRIVASTAVA, S., (2007). Modelling the Dynamic Dependence Structure in Multivariate Financial Time Series. Journal of Time Series Analysis, 28, pp. 763–782.

WALESIAK, M., DUDEK A., (2015). Searching for Optimal Clustering Procedure for a Data Set, package 'clusterSim', https://cran.r-project.org/web/packages/clusterSim.

WANAT, S., (2012). Modele zależności w agregacji ryzyka ubezpieczyciela. [Dependence models in the aggregating of insurer risk], Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

WU, C. C., CHUNG, H., CHANG, Y. H., (2012). The economic value of co-movement between oil price and exchange rate using copula-based GARCH models. Energy Economics, 34, pp. 270–282.

ZOLOTKO, M., OKHRIN, O., (2014). Modelling the general dependence between commodity forward curves. Energy Economics, 43, pp. 284–296.

XIE, X., BENI, G., (1991). Validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13, pp. 841–847.

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 575–578

BOOK REVIEW

Jerzy Witold Wiśniewski. **Microeconometrics in Business Management,** John Wiley @ Sons, Ltd, 2016, ISBN9781119096801. 216 pp.

Reviewed by Jan Kordos, Warsaw Management University

This book introduces the application of micro-econometric methods for modeling various aspects of economic activity for small- to large-sized enterprises, using methods that are based on both time-series and cross-section approaches. The information obtained from using these estimated models can then be used to inform business decisions that improve the efficiency of operations and planning. Basic models used in the modeling of the business (single-equation and multiple-equation systems) are introduced whilst a wide range of economic activity including major aspects of financial management, demand for labour, administrative staff and labour productivity is also explored.

The book consists of Preface, Acknowledgments, six chapters which end with Conclusion and Bibliography.

Chapter 1. A Single-Equation Econometric Model: provides an overview of the essence of an econometric model, its specification, and estimation of the model's parameters as well as its verification, followed by multiplicative econometric models, the limited endogenous variables, an econometric forecasting, its concept and conditions of econometric forecast estimation; the forecasts based on single-equation models and an analysis of econometric forecasts' precision.

Chapter 2. Multiple-Equation Econometric Models: presents a classification of multiple equation models, their reduced forms and an identification of the model; estimation of the parameters of a multiple-equation econometric model and forecasts estimation based on multiple-equation models.

Chapter 3. Econometric Modeling of a Large- and Medium-Sized Enterprise's Economic System: covers a specification of a large- and medium-sized enterprise's econometric model, the structural form of an econometric model of a large- and medium-sized enterprise and an empirical econometric model of a medium-sized enterprise, its assumptions for an econometric empirical model and equation of the sales income, equation of employment, equation of labor productivity, equation of the average wage, equation of the fixed assets, equation of the technical labor equipment and finally application of the company’s model during a decision-making process.

Chapter 4. An Empirical Econometric Model of a Small-Sized Enterprise: describes specification of a small-sized enterprise’s econometric model, its structural form and the model’s total interdependent variables; the model’s predetermined variables; a structural-form’s equations of a small-sized enterprise’s econometric model; an equation of the cash inflows; an equation of the sales income; an equation of ready-made production; an equation of labor efficiency; an equation of the average wage; an equation of the net payroll; the employment equation; an equation of the fixed assets; an equation of wage effectiveness; an equation of the efficiency of implementing the fixed assets; and finally practical applicability of a small-sized enterprise’s model.

Chapter 5. Econometric Modeling in Management of Small-Sized Enterprise: considers the concept of financial liquidity and its measurement in a small-sized enterprise; an econometric modeling of monthly financial liquidity; an econometric modeling of quarterly financial liquidity; an econometric modeling of debt recovery efficacy; measuring the effectiveness of debt recovery in an enterprise; a statistical analysis of debt recovery efficacy in an enterprise; an econometric model describing interdependencies between the financial liquidity and the debt recovery efficacy in an enterprise; and finally an econometric forecasting of financial liquidity.

Chapter 6. Econometric Model in the Analysis of an Enterprise’s Labor Resources: provides a study of a mechanism of the demand for labor; an econometric modeling of labor intensity of production; an econometric model in the selection of an efficient worker; and at the end an econometric model in the selection of an efficient white-collar worker.

In **Conclusion** the author stresses that “The purpose of this book in to invoke awareness for the need of collecting statistical data. Having adequate statistical material at one’s disposition allows application of statistical and econometric tools for improving the decision-making processes and for increases their effectiveness in an enterprise. Free software designed for dealing with those issues is currently available on the Internet”. Next: ”A modern *economist is a specialist who must be able to prepare, to interpret, and to indicate application of econometric and statistical decision-making tools that were discussed in this work”*.

In short, the book:

- Introduces econometric methods which can be used in the modeling of economic activity and forecasting, to help improve the efficiency of business operations and planning.
- Describes econometric entities through multiple-equation and single-equation microeconometric models.

- Explores the process of building and adapting basic microeconometric tools.
- Presents numerous micro-models based on time-series data and statistical cross-sectional sequences, which can be used in any enterprise.
- Features numerous real-world applications along with examples drawn from the author’s own experience.
- Is supported by a companion website featuring practice problems and statistical data to aid students to construct and estimate micro-models.
- Features end-of-chapter exercises with examples present in free software GRETL.

This book serves as a valuable resource for students, business management practitioners and researchers in econometric micro-model construction and various decision-making processes.

It should be added that an econometric model, in the form of a single stochastic equation, is a primary tool in econometrics. The dependent variable is economic in character and represents a specific economic category.

The construction of an econometric model occurs in the following five subsequent stages:

- specification of the model,
- identification of the model,
- estimation of the model's parameters,
- verification of the model, and
- application of the model.

Estimation of the model's structural parameters and its stochastic structure parameters requires having a theoretical model as well as all necessary data collected on each variable of that model.

Application of an econometric model in managing company’s finances is an example of a new approach to solving important company issues. What is considered is the problem of financial liquidity of an enterprise, in connection with effectiveness of debt collection. Here, a simplified tool for defining financial liquidity, which is expressed in the form of time series, is introduced. It therefore allows for a dynamic analysis in confrontation with the measure of efficiency of debt collection that has been defined in the book. Such analysis allows an increase in financial security, thus making management easier, especially in a small-sized company.

Knowledge of the business conducted plays a fundamental role in management. This implies the need for identification of the most important information on the company’s inside as well as on its surroundings. The information generated in the accounting system is, to a large extent, regulated by the state and mainly serves the fiscal needs. A business should create its own system for collecting important information, which is not mandatory, but necessary for rationalization of business decisions. At the same time, it is important to remember that excess of information can be just as harmful as its deficiency. This book gives an account on how to process important statistical information in business.

I highly recommend the book under review, above all, to all persons teaching business management. Those educating others, however, must possess elementary knowledge of statistics and econometrics. They can obtain this knowledge after studying the first two chapters of the book. Those teaching business can encourage students to study the entire book or its parts, depending on the needs and interests. Finally, the book can interest those preparing managerial decisions in an enterprise. Owners of small-sized enterprises, who have adequate business education, can be interested in the solutions demonstrated. They will find the solutions proposed useful in the preparation of decisions. Another important group of readers can encompass enthusiasts of applied econometrics, both in academic institutions as well as in business practice.

It should be added that computerization of the world, universal access to the Internet, emergence of free packages, allow access, collection and processing of statistical information. Application of the solutions that are proposed in the book, using modern information and computer technologies, can improve efficiency of business management, and thereby accelerate creation of wealth.

However, one should remember an acronym *GIGO* and the aphorism "*Garbage in, garbage out*" in the field of *computer science* or *information and communications technology*, which refers to the fact that *computers,* since they operate by logical processes, will unquestioningly process unintended, even nonsensical, *input data* ("*garbage in"*) and produce undesired, often nonsensical, *output* ("*garbage out*"). The principle applies to other fields as well. It was popular in the early days of computing, but applies even more today, when powerful computers can produce large amounts of erroneous information in a short time. I would like to stress that in the book under review the quality of statistical information is properly treated. From my practice I may conclude that in many cases, simple analytic models perform well, therefore the biggest performance increase comes from the data. At the end I would like to quote B. Baesens: "*The best way to improve the performance of a scorecard is not to look for fancy tools or techniques, but to improve* ***data quality*** *first".* (B. Baesen, It's the data, you stupid! Data News, 2007).

This book is also available on the website:
www.wiley.com/go/Wisniewski/Microeconometrics

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 579

Welcome to the International Conference

QUALITY OF LIFE AND SPATIAL COHESION INTERACTION OF DEVELOPMENT AND WELL-BEING IN THE LOCAL CONTEXT

organized by

Central Statistical Office of Poland (CSO)
and
The Cardinal Stefan Wyszynski University in Warsaw (CSWU)
Warsaw, November 17-18, 2016

Honorary Committee

Dominik Rozkrut, President of the CSO
Rev. Prof. Stanisław Dziekoński, Rector of the CSWU
Abp. Henryk Hoser
Prof. Józefa Hrynkiewicz, Parliament RP
Adam Struzik, Marshal of Masovian Voivodeship
Jerzy Kwieciński, Secretary of State, Ministry of Economic Development

Keynote speakers

Graham Kalton (Westat, USA)
Filomena Maggino (University of Florence, Italy)

International Scientific Advisory Committee

Graham Kalton *(USA)*
Filomena Maggino *(Italy)*
Czesław Domański *(Poland)*
Misha Belkindas *(USA)*
Elżbieta Bojanowska *(Poland)*
Zhanjun Xing *(China)*
Sławomir Zaręba *(Poland)*
Semen Matkowski *(Ukraine)*

Scientific Organization Committee

Włodzimierz Okrasa
Dominika Rogalińska
Renata Bielak
Andrzej Ochocki
Marek Cierpiał-Wolan
Henryk Skorowski
Krzysztof Zagórski
Tomasz Korczyński
Grigoris Zarotiadis
Rafał Wiśniewski

To view the call for papers please visit: https://qualityoflifeandspatialcohesion.wordpress.com/
Deadline for abstract submission is 15 October, 2016.

Conference co-Chairs

Włodzimierz Okrasa *(CSWU)*
Dominika Rogalińska *(CSO)*

Contact persons:

Tomasz Korczyński, t.korczynski@uksw.edu.pl
Iwona Miziołek, i.miziolek@stat.gov.pl
Katarzyna Pomianowska, k.pomianowska@stat.gov.pl

The project is co-financed by the European Union from the Cohesion Fund under Technical Assistance Operational Programme 2014–2020.

STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 581–584

ABOUT THE AUTHORS

Al-Kandari Noriah is an Associate Professor in the Department of Statistics and Operations Research at Kuwait University. She has been teaching since 1998. She received her BSc (1993) from Kuwait University, MSc (1994) and PhD (1998) from Aberdeen University, Scotland. Her research interests include multivariate analysis, regression analysis, parameter estimation, environmental statistics, Bayesian statistics, record linkage, and change-point analysis. She is a Fellow of the American Statistical Association and elected member of the International Statistical Institute. Department of Statistics and Operations Research, Faculty of Science, Kuwait University. P.O. Box 5969, Safat 13060. E-mail: noriah@stat.kuniv.edu. Telephone: (+965) 24985315. Fax: (+965) 24837332.

Balcerzak Adam P. is an Assistant Professor at the Faculty of Economic Sciences and Management at Nicolaus Copernicus University. His research interests are institutional economics, government role in supporting technological potential of economy, effectiveness of national innovative systems, determinants of total factor productivity in developed countries. He is a president of Polish Economic Society Branch in Toruń and the Editor-in-Chief of two international journals: ‘Equilibrium. Quarterly Journal of Economics and Economic Policy’ and ‘Oeconomia Copernicana’.

Baszczyńska Aleksandra is an Assistant Professor at the Department of Statistical Methods, University of Lodz. Her major scientific research focuses on nonparametric methods including functional characteristic estimation. The main research areas are kernel methods in estimation and testing hypothesis emphasizing special issue of the influence of smoothing parameter and kernel function on the results of estimation and testing procedures.

Głowicka-Wołoszyn Romana is an Assistant Professor at Poznań University of Life Sciences, Faculty of Economic and Social Sciences. Her main scientific interests include quantitative methods and multivariate analysis methods, especially in their application to economics and finance. Currently her research focuses on the methods of spatial statistics and econometrics.

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. Currently, he is an Assistant Professor at this University. His research interests include machine learning, times series classification and data mining.

Jędrzejczak Alina is an Associate Professor at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. Simultaneously she holds the position of an expert in the Centre of Mathematical Statistics at the regional Statistical Office in Lodz. Her main areas of interest include income distributions, income inequality and poverty measurement as well as small area estimation. Currently, she is a member of two editorial boards: Statistica & Applicazioni and Folia Oeconomica Acta Universitatis Lodziensis.

Kordos Jan graduated from Jagiellonian University (in mathematics, 1953) and Wroclaw University (in mathematical statistics, 1955); PhD in Econometrics from the Academy of Economics, Katowice, Poland (1965), Habilitation diploma in Econometrics from the same Academy (1973), and Professorship from the President of Poland (1990). He worked as the Chief of the Methodology Section at the Division of Living Conditions, Central Statistical Office/CSO (1955-1966) and of the Laboratory of Mathematical Methods at Research Center of Statistics and Economics (CSO, 1966-74). He served as the FAO Adviser in Agricultural Statistics in Ethiopia (1974-80). He acted as Director of the Division of Demographic and Social Surveys (1981-92) and as Vice-President of the CSO Poland (1992-96). He was lecturing and training on agricultural Statistics in China in the late 80s, and also in Kathmandu, Nepal (1990, 1991). During 1994-96 he served as the World Bank Consultant in Household Budget Surveys in Latvia and Lithuania. He was President of the Polish Statistical Association (1985-94). He was the founder and the Editor-in-Chief of the *Statistics in Transitio*n (1993-2007). Now, he is a Professor of Statistics at the Warsaw Management University. His publications include four books and over three hundred articles and other papers in Polish and English.

Kozera Agnieszka is an Assistant Professor at Poznań University of Life Sciences, Faculty of Economic and Social Sciences. She received his PhD from Poznań University of Life Sciences in 2014. Her main research interests include quantitative methods and multivariate analysis methods, especially in their application to economics and finance. She is an author or co-author over 40 research papers.

Krzyśko Mirosław is a Professor Emeritus at the Department of Probability and Mathematical Statistics in Adam Mickiewicz University, Poznań, Poland. His research interests are multivariate statistical analysis, analysis of multivariate functional data, statistical inference and data analysis in particular. Professor Krzyśko has published more than 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.

Kubacki Jan is a Chief Specialist at the Centre of Mathematical Statistics, Statistical Office in Lodz. In 2009, he received his PhD from Warsaw School of Economics in the area of social statistics, in particular in small area estimation under the supervision of Prof. Jan Kordos. His interests include small area estimation, sample survey methods, classification methods, data processing and statistical software development. He is a member of the Editorial Board of “Statistical News” (Wiadomości Statystyczne).

Lahiri Partha is a Professor of the Joint Program in Survey Methodology (JPSM) at the University of Maryland at College Park, and an Adjunct Research Professor of the Institute of Social Research, University of Michigan, Ann Arbor. Prior to coming to Maryland, Dr. Lahiri was the Milton Mohr Distinguished Professor of Statistics at the University of Nebraska-Lincoln. His research interests include survey sampling, official statistics, and small-area estimation. Dr. Lahiri has served on a number of advisory committees, including the U.S. Census Advisory committee and U.S. National Academy panel. Over the years Dr. Lahiri advised various local and international organizations such as the United Nations Development Program, World Bank, Gallup Organization. Dr. Lahiri is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, and an elected member of the International Statistical Institute.

Papież Monika received her MSc in mathematics from the Jagiellonian University in Cracow. Currently, she works as an Assistant Professor at the Department of Statistics, Cracow University of Economics, where she has obtained her PhD and postdoctoral degree. Her main research interests include energy economics, energy security, time series analysis and modelling and forecasting of socio-economic phenomena.

Pietrzak Michał Bernard is an Assistant Professor at the Faculty of Economic Sciences and Management at Nicolaus Copernicus University. He specializes in applications of special econometrics tools in economics. He is a secretary of Polish Statistical Association Branch in Toruń. He is a statistical editor in the journal ‘Equilibrium. Quarterly Journal of Economics and Economic Policy’.

Shanker Rama is working as Professor and Head, Department of Statistics, Eritrea Institute of Technology, Asmara, Eritrea, NE Africa. He is a well-known researcher in the field of Distribution Theory, Statistical Inference, Reliability and Survival Analysis, Operations Research, Statistical Modelling and Biostatistics. He is the member of Advisory Boards and Editorial Boards of many international journals and reviewer of reputed international and national journals of Statistics and Biostatistics. He has published more than 50 research papers in reputable journals of Statistics, Mathematics and Operations Research. Presently, he is also the Editor-in-Chief of Eritrean Journal of Science and Engineering (EJSE).

Szulc Adam is an Associate Professor at the Institute of Statistics and Demography of the Warsaw School of Economics. His fields of interests cover poverty and inequality, consumer demand systems, equivalence scales, social policy evaluation and matching estimation.

Śmiech Sławomir received his MSc in mathematics from the Jagiellonian University in Krakow. Currently, he works as an Assistant Professor at Cracow University of Economics, Department of Statistics, where he has obtained his PhD and postdoctoral degree. His scientific interests encompass problems of energy economics and monetary policy. At present, he is a member of three research teams, within which he studies interrelations between real and financial spheres of economy and energy market, builds classifications of monetary policy regimes, and searches for determinants of energy policy decisions.

Wanat Stanisław is an Associate Professor at the Department of Statistics, Cracow University of Economics. His research activities focus on the fields of dependence and copula models, with particular emphasis on applications in financial and insurance risk management. He is also interested in actuarial methods, multivariate statistical analysis and forecasting methods.

Wołyński Waldemar is an Assistant Professor at the Faculty of Mathematics and Computer Science 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 over 40 research papers.

Wywiał Janusz L. is a Professor in the Department of Statistics, Faculty of Management, University of Economics in Katowice, Poland. Since 2000 he has been the chair of this Department. His main field of interests is survey sampling, especially: problems of optimal stratification of population, optimal determination of sample sizes, sampling designs and strategies dependent on auxiliary variables, and monetary sampling. Moreover, he focuses on statistical methods in auditing, clustering methods and testing statistical hypothesis. Professor Wywiał is an author or co-author of 105 reviewed papers, 6 monographs and 7 textbooks. He has been involved in the implementation of scientific projects. Professor Wywiał is a member of several scientific bodies.

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).