STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Editorial Office **ISSN 1234-7655**
Marek Cierpiał-Wolan, Ph.D., Scientific Secretary
m.wolan@stat.gov.pl
Secretary:
Beata Witek, b.witek@stat.gov.pl
Agata Bara, a.bara@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, June 2016
Vol. 17, No. 2, pp. 157−160

FROM THE EDITOR

A set of twelve articles included in this issue is structured conventionally in three major sections. It is appended by short reports from two international conferences – a selection of papers from these conferences constitutes the last section of this issue.

The first section – sampling methods and estimation – contains three articles. It starts with **G. N. Singh's**, **Mukti Khetan's**, **Shweta Maurya's** paper ***Some Effective Estimation Procedures Under Non-Response in Two-Phase Successive Sampling.*** The authors discuss an import issue of how to assess the effect of non-response in estimation of the current population mean in two-phase successive sampling on two occasions. They use the sub-sampling technique of non-respondents and propose exponential methods of estimation under two-phase successive sampling arrangement. Properties of the proposed estimation procedures have been examined along with some suggestions on estimation procedures for survey practitioners concerning the use of auxiliary information (in the form of exponential methods of estimation).

The next paper, ***Transmuted Kumaraswamy Distribution*** by **Muhammad Shuaib Khan**, **Robert King**, **Irene Lena Hudson** is devoted to one of the most widely applied statistical distribution in hydrological problems and many natural phenomena. The authors propose a generalization of the Kumaraswamy distribution referred to as the transmuted Kumaraswamy (TKw) distribution using the quadratic rank transmutation map (studied by Shaw et al., 2009). They provide a comprehensive account of the mathematical properties of the new distribution with specific expressions for the moments, moment generating function, entropy, mean deviation, Bonferroni and Lorenz curves, and formulated moments for order statistics. The TKw distribution parameters are estimated by using the method of maximum likelihood, and Monte Carlo simulation is performed in order to investigate the performance of MLEs – the usefulness of the proposed model is illustrated using the flood and HIV/AIDS data. In conclusions, the authors indicate on the better performance of the model (in terms of a better fit) than the Kw distribution.

Ranjita Pandey's and **Anoop Chaturvedi's** paper ***Bayesian Inference for State Space Model with Panel Data*** explores panel data set-up in a Bayesian state space model. The authors apply the conditional posterior densities of parameters to determine the marginal posterior densities using the Gibbs sampler. They believe that the theoretical framework they developed is generally more effective and useful for applied researchers and practitioners, especially in terms of a more precise panel data-based prediction.

The research paper section starts with an article by **Mauro Mussini**, ***On Measuring Income Polarization: An Approach Based on Regression Trees*** in which the application of regression trees for analysing income polarization is proposed. Using an approach to polarization based on the analysis of variance, the author shows how the regression trees can uncover groups of homogeneous income receivers in a data-driven way. Since the regression tree can deal with nonlinear relationships between income and the characteristics of income receivers, it can detect which characteristics and their interactions actually play a role in explaining income polarization. In consequence, the author believes the regression tree is a useful flexible statistical tool to explore whether income receivers concentrate around local poles. Some interesting partition of income receivers is demonstrated for the case of Italy. For instance, an empirical analysis of Italian income data shows that the interactions among employment status, educational qualification and age form well-identified groups of income receivers, whereas the other characteristics do not play a clear role in explaining income polarization.

Oleksandr Osaulenko's article ***Quality of Life and Poverty in Ukraine – Preliminary Assessment Based on the Subjective Well-Being Indicators*** provides a first-hand account of the research on the topic in Ukraine. It starts with an overview of the database and methodology, followed by main results of the quality of life and poverty research conducted by the national statistical office. A system of subjective well-being indicators is based on self-evaluation of the attained level of well-being, of the level of meeting the basic living needs and the levels of deprivation of consumption. In addition, methodological approaches to analyzing economic and infrastructure deprivation (due to the geographic limitations of services accessibility) are briefly described. The paper reviews the factors that underlie the Ukrainian list of deprivations and define the percentage of population that is particularly affected by multiple deprivation. The data covers a period of several years allowing for exploration of the distribution of deprivations by different population groups.

In the last article of this section, ***On the Relationships Between Smart Growth and Cohesion Indicators in the EU Countries*** by **Beata Bal-Domańska**, **Elżbieta Sobczak,** the problem of evaluation of the relationships between smart growth and economic and social cohesion factors is discussed from the perspective of the Europe 2020 strategy's objectives toward improving the situation in education, digital society and research and innovation. The authors employ aggregate measures for these three phenomena based on panel data models. Social cohesion is described by the level of employment rate as one of the conditions essential to the well-being and prosperity of individuals, and economic cohesion is defined by the level of GDP per capita in PPS. The study covered the group of 27 European Union countries during the period of 2002-2011. Some of the conclusions have policy implications. For instance, it was found that an increase in the employment rate was related to the increasing role of employment in smart specialization sectors.

The conference papers' section is opened by ***Heteroscedastic Discriminant Analysis Combined with Feature Selection for Credit Scoring*** in which **Katarzyna Stąpor**, **Tomasz Smolarczyk**, **Piotr Fabian** propose an approach for building a credit scoring model based on the combination of heteroscedastic extension of classical Fisher Linear Discriminant Analysis and a feature selection algorithm that retains sufficient information for classification purpose. Starting with an observation that credit granting is a fundamental and one of the most complex question being faced by credit institutions, the authors attempt to develop an effective classification model that would be helpful for managers. To this aim, they focused on the feature selection algorithm that retains sufficient information for classification purpose and tested five feature subset selection algorithms: two filters and three wrappers. In order to evaluate the accuracy of the proposed credit scoring model and to compare it with the existing approaches, the German credit data set was used (Chen, Li, 2010). In the conclusions they found that the proposed hybrid approach is an effective and promising method for building credit scoring models. The analysis results in better prediction accuracy, also due to applying variable importance analysis for identifying the most relevant variables for the classification purpose.

In **Iwona Skrodzka's** article ***Knowledge-Based Economy in the European Union – Cross-Country Analysis*** spatial differences in the level of development of the knowledge-based economy in the European Union countries are discussed using a soft modelling approach. The estimation of a synthetic measure of KBE, as well as the arrangement and classification of the UE-27 countries into typological groups for the years 2000 and 2013 are provided. For instance, the highest level of development of the knowledge-based economy was observed for Sweden, Denmark, Finland and Luxembourg, whereas the lowest one for Greece, Bulgaria and Romania. Eleven countries, including Poland, improved their ranking in 2013 compared to 2000, while nine countries reduced their positions (the highest increase was in Hungary, while the largest fall in Italy in 2013).

In the next article, ***New Method of Variable Selection for Binary Data Cluster Analysis*** by **Jerzy Korzeniewski** the problem of the level of measurement in the variable selection procedures is discussed, with intention of improving the efficiency of the existing methods, with special reference to the marketing type data. The author proposes that a variable selection method be based on connecting the filtering of the input set of all variables with grouping of sets of variables similar with respect to analogous groupings of objects. The new method allows for linking good features of two entirely different approaches to variable selection in cluster analysis, i.e. *filtering* methods and *wrapper* methods. The proposed method of variable selection yields best results when the classical *k*-means method of objects grouping is slightly modified.

In **Dominik Krężołek's** paper, ***The GlueVar Risk Measure and Investor's Attitudes to Risk – An Application to the Non-Ferrous Metals Market,*** the issue of risk in economic investment decisions is discussed and a new risk measure is proposed – the GlueVaR risk measure. It can be defined as a linear combination of VaR and GlueVaR and is aimed at helping to deal with uncertainty and

volatility that is characteristic to the economic investment decisions. The most commonly used risk measure, Value-at-Risk, suffers from a significant drawback, which is the lack of subadditivity, but is crucial in terms of portfolio diversification. The proposed GlueVaR measure allows for calculating the level of investment loss depending on investment's attitudes to risk while meeting the needed requirement, therefore it may be used in portfolio risk assessment. The application of the GlueVaR risk measure is presented for the non-ferrous metals market. Compared to classical measures, the most useful feature of the proposed new risk measures is that for a particular investor it is possible to implicitly define the set of adverse events and determine the importance of such events.

Marta Hozer-Koćmiel, **Christian Lis** present the results of ***Examining Similarities in Time Allocation Amongst European Countries.*** Time allocation has been defined as the daily distribution of time to various activities. Professional work time, domestic work time and leisure time appear to be the most important for the economic approach. It has been proved that there are coherent groups of countries with similar structure of time allocation. The taxonomic methods used in order to verify the thesis included: cluster analysis, k-means method, generalised distance measure GDM and interval taxonomic method TMI. The analysis was performed on the basis of HETUS survey data. In conclusions, two groups of countries that show strong similarity in setting the time budget of the population have emerged from the analysis: (i) the 'new' European countries (that has undergone economic transformation), characterized by a distinctly longer professional work time and shorter leisure time; and (ii) the Scandinavian countries and the more developed countries of Western Europe, where the basic variables of time allocation were opposite, i.e., relatively short professional work time and long leisure time.

Joanna M. Landmesser's paper, ***Decomposition of Differences in Income Distributions Using Quantile Regression,*** deals with microeconometric techniques useful for the study of differences between groups of objects. Using the Machado-Mata quantile regression approach, the empirical decomposition of the inequalities in income distributions of one-person households in urban and rural areas was performed using data from the Household Budget Survey for Poland in 2012. It was found that the tendency towards increased income inequalities between urban and rural residents when moving to the right of the income distribution can be observed. The rural residents are at a disadvantage. The decomposition of the inequalities revealed a growing share of the part explained by different characteristics of people (especially in educational level), and a declining share of the unexplained part, associated with the evaluation of those characteristics.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, June 2016
Vol. 12, No. 2, pp. 161

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, June 2016
Vol. 17, No. 2, pp. 163–182

SOME EFFECTIVE ESTIMATION PROCEDURES UNDER NON-RESPONSE IN TWO-PHASE SUCCESSIVE SAMPLING

G. N. Singh[1], **M. Khetan**[2], **S. Maurya**[3]

ABSTRACT

This work is designed to assess the effect of non-response in estimation of the current population mean in two-phase successive sampling on two occasions. Sub-sampling technique of non-respondents has been used and exponential methods of estimation under two-phase successive sampling arrangement have been proposed. Properties of the proposed estimation procedures have been examined. Empirical studies are carried out to justify the suggested estimation procedures and suitable recommendations have been made to the survey practitioners.

Key words: non-response, successive sampling, two-phase sampling, mean square error, optimum replacement strategy.

1. Introduction

In collecting information through sample surveys, there may arise numerous problems; one of them is non-response. It frequently occurs in mail surveys, where some of the selected units may refuse to return back the filled in questionnaires. An estimate obtained from such an incomplete survey may be misleading, especially when the respondents differ significantly from the non-respondents, because the estimate may be a biased one. Hansen and Hurwitz (1946) suggested a technique of sub-sampling of non-respondents to handle the problem of non-response. Cochran (1977) and Fabian and Hyunshik (2000) extended the Hansen and Hurwitz (1946) technique for the situation when besides the information on the character under study, information on auxiliary character is also available. Recently, Choudhary *et al*. (2004) Singh and Priyanka (2007), Singh and Kumar

[1] Department of Applied Mathematics, Indian School of Mines, Dhanbad 826004. E-mail: gnsingh_ism@yahoo.com.
[2] Department of Applied Mathematics, Indian School of Mines, Dhanbad 826004. E-mail: mukti.khetan11@gmail.com.
[3] Department of Applied Mathematics, Indian School of Mines, Dhanbad 826004.

(2009, 2010), Singh *et al.* (2011) and Garcia Luengo (2013) used the Hansen and Hurwitz (1946) technique for the estimation of population mean on the current occasion in two-occasion successive sampling.

If the study character of a finite population is subject to change over time, a single occasion survey is insufficient. For such a situation successive sampling provides a strong tool for generating reliable estimates over different occasions. Sampling on successive occasions was first considered by Jessen (1942) in the analysis of farm data. The theory of successive (rotation) sampling was further extended by Patterson (1950), Eckler (1955), Rao and Graham (1964), Sen (1971, 1972, 1973), Gupta (1979), Das (1982) and Singh and Singh (2001) among others.

In sample surveys, the use of auxiliary information has shown its significance in improving the precision of estimates of unknown population parameters. When the population parameters of auxiliary variable are unknown before start of the survey we go for two-phase (double) sampling structure to provide the reliable estimates of the unknown population parameters. Singh and Singh (1965) used two-phase (double) sampling for stratification on successive occasions. Recently, Singh and Prasad (2011) and Singh and Homa (2014) applied two-phase sampling scheme with success in the estimation of the current population mean in two-occasion successive sampling.

The aim of the present work is to study the effect of non-response when it occurs on various occasions in two-occasion successive (rotation) sampling. Recently, Bahl and Tuteja (1991), Singh and Vishwakarma (2007) and Singh and Homa (2013) suggested exponential type estimators of population mean under different realistic situations. Motivated with the dominating nature of these estimators and utilizing the information on a stable auxiliary variable with unknown population mean over both occasions, some new exponential methods of estimation have been proposed to estimate the current population mean in two-phase successive (rotation) sampling arrangement.. The Hansen and Hurwitz (1946) technique of sub-sampling of non-respondents has been used to reduce the negative effects of non-response. Properties of the proposed estimators are examined and their empirical comparisons are made with the similar estimator and with the natural successive sampling estimator when complete response is observed on both occasions. Results are interpreted and followed by suitable recommendations.

2. Sample structures and symbols

Let $U = (U_1, U_2, -, -, -, U_N)$ be the finite population of N units, which has been sampled over two occasions. The character under study is denoted by x(y) on the first (second) occasion respectively. It is assumed that the non-response occurs only in study variable x(y) and information on an auxiliary variable z (stable over occasion), whose population mean is unknown on both occasions, is available and positively correlated with study variable. Since we have assumed that non-

response occurs on both occasions, the population can be divided into two classes – those who will respond at the first attempt and those who will not on both occasions. Let the sizes of these two classes be N_1^* and N_2^* respectively on the first occasion and the corresponding sizes on the current (second) occasion be N_1 and N_2, respectively. To furnish a good estimate of the population mean of the auxiliary variable z on the first occasion, a preliminary sample of size n' is drawn from the population by the simple random sampling without replacement (SRSWOR) method, and information on z is collected. Further, a second-phase sample of size n ($n' > n$) is drawn from the first-phase (preliminary) sample by the SRSWOR method and henceforth the information on the study character x is gathered. We assume that out of selected n units, n_1 units respond and n_2 unit do not respond. Let n_{2h} denote the size of sub-sample drawn from the non-responding units in the sample on first occasion. A random sub-sample s_m of $m = n\lambda$ units is retained (matched) from the responding units on the first occasion for its use on the second occasion under the assumption that these units will give complete response on the second occasion as well. Once again, to furnish a fresh estimate of the population mean of the auxiliary variable z on the second occasion, a preliminary (first-phase) sample of size u' is drawn from the non-sampled units of the population by the SRSWOR method and information on z is collected. A second-phase sample of size $u = (n-m) = n\mu$ ($u' > u$) is drawn from the first-phase (preliminary) sample by the SRSWOR method and the information on study variable y is gathered. It is obvious that the sample size on the second occasion is also n. Here λ and μ ($\lambda+\mu =1$) are the fractions of the matched and fresh samples, respectively, on the second (current) occasion. We assume that in the unmatched portion of the sample on the current (second) occasion u_1 units respond and u_2 units do not respond. Let u_{2h} denote the size of the sub-sample drawn from the non-responding units in the fresh sample (s_u) on the current (second) occasion. Hence, onwards, we use the following notations:

$\bar{X}, \bar{Y}, \bar{Z}$: The population means of the variables x, y and z respectively.

$\bar{y}_m, \bar{y}_u, \bar{y}_{u_1}, \bar{y}_{u_{2h}}, \bar{x}_n, \bar{x}_{n_1}, \bar{x}_{n_{2h}}, \bar{x}_m, \bar{z}_m, \bar{z}_u$: The sample means of the respective variables based on the sample sizes shown in suffices.

$\bar{z}_n', \bar{z}_u'$: The sample means of the auxiliary variable z and based on the first-phase samples of sizes u' and n' respectively.

$\rho_{yx}, \rho_{xz}, \rho_{yz}$: The population correlation coefficients between the variables shown in suffices.

S_x^2, S_y^2, S_z^2: The population variances of the variables x, y and z respectively.

S_{2x}^2, S_{2y}^2: The population variances of the variables x and y respectively in the non-responding units of the population.

C_x, C_y, C_z: The coefficients of variation of the variables x, y and z respectively.

C_{2x}, C_{2y} : The coefficients of variation of the variables x and y in the non-responding units of the population.

$W^* = \frac{N_2^*}{N}$: The proportion of non-responding units in the population at first occasion. $W = \frac{N_2}{N}$: The proportion of non-responding units in the population on the current (second) occasion.

$f_1 = \frac{n_2}{n_{2h}}$ and $f_2 = \frac{u_2}{u_{2h}}$.

3. Formulation of estimation strategy

To estimate the population mean $\bar{Y}$ on the current (second) occasion, two different estimators are considered – one estimator T_u based on sample s_u of size u drawn afresh on the second occasion and the second estimator T_m based on the sample s_m of size m, which is common to both occasions. Since the non-response occurs in the samples s_n and s_u, we have used the Hansen and Hurwitz (1946) technique to propose the estimators T_u and T_m. Hence, the estimators T_u and T_m for estimating the current population mean $\bar{Y}$ are formulated as

$$T_u = \bar{y}_u^* \exp\left(\frac{\bar{z}_u' - \bar{z}_u}{\bar{z}_u' + \bar{z}_u}\right) \text{ and } T_m = \bar{y}_m \exp\left(\frac{\bar{x}_n^* - \bar{x}_m}{\bar{x}_n^* + \bar{x}_m}\right)\left(\frac{\bar{z}_n'}{\bar{z}_m}\right)$$

where

$$\bar{x}_n^* = \frac{n_1 \bar{x}_{n_1} + n_2 \bar{x}_{n_{2h}}}{n} \text{ and } \bar{y}_u^* = \frac{u_1 \bar{y}_{u_1} + u_2 \bar{y}_{u_{2h}}}{u}.$$

Combining the estimators T_u and T_m, finally we have the following estimator of population mean $\bar{Y}$ on the current (second) occasion

$$T = \varphi T_u + (1-\varphi) T_m \tag{3.1}$$

where $\varphi\ (0 \le \varphi \le 1)$ is the unknown constant (scalar) to be determined under certain criterions.

4. Properties of the estimator T

Since the estimators T_u and T_m are exponential type estimators, the population mean $\bar{Y}$ are biased, therefore the resulting estimator T defined in

equation (3.1) is also a biased estimator of $\bar{Y}$. The bias B (.) and the mean square error M (.) of the estimator T are derived up to the first order of approximations using the following transformations:

$\bar{y}_m = (1+e_1)\bar{Y}$, $\bar{y}_u = (1+e_2)\bar{Y}$, $\bar{y}_u^* = (1+e_3)\bar{Y}$, $\bar{x}_m = (1+e_4)\bar{X}$, $\bar{x}_n = (1+e_5)\bar{X}$, $\bar{x}_n' = (1+e_6)\bar{X}$, $\bar{x}_n^* = (1+e_7)\bar{X}$, $\bar{z}_m = (1+e_8)\bar{Z}$, $\bar{z}_u = (1+e_9)\bar{Z}$, $\bar{z}_u' = (1+e_{10})\bar{Z}$, $\bar{z}_n' = (1+e_{11})\bar{Z}$, such that $E(e_i) = 0$, $|e_i| < 1 \ \forall$ i = 1, 2, 3, - - -, 11. Under the above transformations, the estimators T_u and T_m take the following forms:

$$T_u = \bar{Y}(1+e_3)\exp\left[\frac{1}{2}(e_{10}-e_9)\left(1+\frac{1}{2}(e_{10}+e_9)\right)^{-1}\right] \tag{4.1}$$

and

$$T_m = \bar{Y}(1+e_1)(1+e_{11})(1+e_8)^{-1}\exp\left[\frac{1}{2}(e_7-e_4)\left(1+\frac{1}{2}(e_7+e_4)\right)^{-1}\right] \tag{4.2}$$

Thus, we have the following theorems:

Theorem 4.1.

Bias of the estimator T to the first order of approximations is obtained as

$$B(T) = \varphi B(T_u) + (1-\varphi)B(T_m) \tag{4.3}$$

where

$$B(T_u) = \bar{Y}\left(\frac{1}{u}-\frac{1}{u'}\right)\left(\frac{3}{8}C_z^2 - \frac{1}{2}\rho_{yz}C_yC_z\right)$$

and

$$B(T_m) = \bar{Y}\left\{\begin{array}{l}\left(\frac{1}{m}-\frac{1}{n}\right)\left(\frac{3}{8}C_x^2 + \frac{1}{2}\rho_{xz}C_xC_z - \frac{1}{2}\rho_{xy}C_yC_x\right) \\ -\frac{1}{8}\frac{(f_1-1)}{n}W^*C_{2x}^2 + \left(\frac{1}{m}-\frac{1}{n'}\right)\left(C_z^2 - \rho_{yz}C_yC_z\right)\end{array}\right\}$$

Proof

The bias of the estimator T is given by

$$\begin{aligned} B(T) = E\left[T-\bar{Y}\right] &= \varphi E(T_u-\bar{Y}) + (1-\varphi)\,E(T_m-\bar{Y}) \\ &= \varphi B(T_u) + (1-\varphi)\,B(T_m) \end{aligned} \tag{4.4}$$

where

$$B(T_u) = E\left[T_u-\bar{Y}\right] \text{ and } B(T_m) = E\left[T_m-\bar{Y}\right].$$

Substituting the expressions of T_u, and T_m from equations (4.1) and (4.2) in equation (4.4), expanding the terms binomially and exponentially, taking expectations and retaining the terms up to the first order of sample sizes, we have the expressions for the bias of the estimator T as described in equation (4.3).

Theorem 4.2.

The mean square error of the estimator T to the first order of approximations is obtained as

$$M(T) = \varphi^2 M(T_u) + (1-\varphi)^2 M(T_m) + 2\varphi(1-\varphi)C \tag{4.5}$$

where

$$M(T_u) = E\left(T_u - \bar{Y}\right)^2 = \left[\left(\frac{1}{u} - \frac{1}{u'}\right)\left(\frac{1}{4} - \rho_{yz}\right) + \left(\frac{1}{u} - \frac{1}{N}\right) + \frac{W(f_2-1)}{u}\right] S_y^2 \tag{4.6}$$

$$M(T_m) = E\left(T_m - \bar{Y}\right)^2 = \left[\begin{array}{l} \left\{\left(\frac{1}{m} - \frac{1}{n}\right)\left(\frac{1}{4} + \rho_{xz} - \rho_{yx}\right)\right\} + \left(\frac{1}{m} - \frac{1}{N}\right) \\ + \left\{\left(\frac{1}{m} - \frac{1}{n'}\right)\left(1 - 2\rho_{yz}\right)\right\} + \frac{1}{4}\frac{(f_1-1)}{n} W^* \end{array}\right] S_y^2 \tag{4.7}$$

and

$$C = E\left[\left(T_u - \bar{Y}\right)\left(T_m - \bar{Y}\right)\right] = -\frac{S_y^2}{N}. \tag{4.8}$$

Proof

It is obvious that the mean square error of the estimator T is given by

$$M(T) = E\left[T - \bar{Y}\right]^2 = E\left[\varphi\left(T_u - \bar{Y}\right) + (1-\varphi)\left(T_m - \bar{Y}\right)\right]^2$$

$$= \varphi^2 E\left(T_u - \bar{Y}\right)^2 + (1-\varphi)^2 E\left(T_m - \bar{Y}\right)^2 + 2\varphi(1-\varphi) E\left[\left(T_u - \bar{Y}\right)\left(T_m - \bar{Y}\right)\right]$$

$$= \varphi^2 M(T_u) + (1-\varphi)^2 M(T_m) + 2\varphi(1-\varphi)C \tag{4.9}$$

Substituting the expressions of T_u, and T_m from equations (4.1)-(4.2) in equation (4.9), expanding the terms binomially and exponentially, taking expectations and retaining the terms up to the first order of sample sizes, we have the expression of the mean square error of the estimator T as it is given in equation (4.5).

Remark 4.1.

The expression of the mean square error in the equation (4.5) is derived under the assumptions (i) that the coefficients of variation of non-response class are similar to that of the population, i.e. $C_{2x} = C_x$ and $C_{2y} = C_y$, and (ii) since x and y are the same study variable over two occasions and z is the auxiliary variable correlated to x and y, looking at the stability nature of the coefficients of variation, viz. Reddy (1978), the coefficients of variation of the variables x, y and z in the population are considered equal, i.e. $C_x = C_y = C_z$.

5. Minimum mean square error of the estimator T

Since the mean square error of the estimator T in equation (4.5) is the function of unknown constant φ, it is minimized with respect to φ, and subsequently the optimum value of φ is obtained as

$$\varphi_{opt} = \frac{M(T_m)-C}{M(T_u)+M(T_m)-2C} . \tag{5.1}$$

Now, substituting the value of φ_{opt} in equation (4.5), we get the optimum mean square error of T as

$$M(T)_{opt} = \frac{M(T_u).M(T_m)-C^2}{M(T_u)+M(T_m)-2C}. \tag{5.2}$$

Further, substituting the values from equations (4.6)-(4.8) in equation (5.2), we get the simplified value of $M(T)_{opt}$ which is given below:

$$M(T)_{opt} = \frac{a_3+\mu a_2+\mu^2 a_1}{a_6+\mu a_5+\mu^2 a_4}\frac{S_y^2}{n}. \tag{5.3}$$

where

$a_1 = ac+k^2f^2$, $a_2 = ad+cb-k^2f^2$, $a_3 = bd$, $a_4 = c-a+2kf$, $a_5 = a-b+d-2kf$, $a_6 = b$,

$a = -(f+t_1a_0)$, $b= a_0+1+(f_2-1)W$, $c = t_2d_1+c_1+f-\frac{1}{4}(f_1-1)W^*$,

$d = 1-f+(1-t_2)d_1+\frac{1}{4}(f_1-1)W^*$, $k =-1$, $a_0= \frac{1}{4}-\rho_{yz}$, $c_1= \frac{1}{4}+\rho_{xz}-\rho_{xy}$, $d_1= 1-2\rho_{yz}$,

$f = \frac{n}{N}$, $f_1 = \frac{n_2}{n_{2h}}$, $f_2 = \frac{u_2}{u_{2h}}$, $t_1 = \frac{n}{u'}$ and $t_2 = \frac{n}{n'}$.

6. Optimum replacement strategy

Since the mean square error of the estimator T given in equation (5.3) is the function of μ (fractions of the sample to be drawn afresh at the second occasion), the optimum value of μ is determined to estimate the population mean $\bar{Y}$ with maximum precision and lowest cost. To determine the optimum value of μ, we minimized the mean square error of the estimator T given in equation (5.3) with respect to μ, which results in quadratic equation in μ and the respective solutions of μ, say $\hat{\mu}$, are given below:

$$p_1\mu^2+2p_2\mu+p_3=0 \tag{6.1}$$

$$\hat{\mu} = \frac{-p_2 \pm \sqrt{p_2^2-p_1p_3}}{p_1} \tag{6.2}$$

where

$$p_1 = a_1a_5-a_2a_4,\ p_2 = a_1a_6-a_3a_4 \text{ and } p_3 = a_2a_6-a_3a_5.$$

From equation (6.2) it is obvious that real values of $\hat{\mu}$ exist iff the quantities under square root are greater than or equal to zero. For any combinations of correlations ρ_{yx}, ρ_{xz} and ρ_{yz}, which satisfy the conditions of real solutions, two real values of $\hat{\mu}$ are possible. Hence, while choosing the values of $\hat{\mu}$, it should be remembered that $0 \le \hat{\mu} \le 1$. If both the values of $\hat{\mu}$ satisfy the stated condition, we chose the smaller value of $\hat{\mu}$ as it will help in reducing the cost of the survey. All other values of μ are inadmissible. Substituting the admissible value of $\hat{\mu}$, say $\mu^{(0)}$, from equation (6.2) into equation (5.3), we have the optimum value of the mean square error of T, which is shown below:

$$M(T^0)_{opt} = \frac{a_3+\mu^{(0)}a_2+\mu^{(0)2}a_1}{a_6+\mu^{(0)}a_5+\mu^{(0)2}a_4}\frac{S_y^2}{n}. \tag{6.3}$$

7. Some special cases

Case 1: When non-response occurs only at first occasion

When non-response occurs only at first occasion, the estimator for the mean $\bar{Y}$ on the current occasion may be obtained as

$$T^* = \varphi^* \xi_{1u} + \left(1-\varphi^*\right) T_m \tag{7.1}$$

where

$\xi_{1u} = \bar{y}_u \exp\left(\frac{\bar{z}_u' - \bar{z}_u}{\bar{z}_u' + \bar{z}_u}\right)$ and T_m is defined in section 3, where φ^* $(0 \le \varphi^* \le 1)$ is the unknown constant (scalar) to be determined under certain criterions.

7.1. properties of the estimator T^*

Since the estimator T^* is exponential type estimator, it is biased for the population mean $\bar{Y}$. The bias B (.) and the mean square error M (.) of the estimator T^* are derived up to the first order of approximations similar to that of the estimator T.

Theorem 7.1.

The bias of the estimator T^* to the first order of approximations is obtained as

$$B(T^*) = \varphi^* B(\xi_{1u}) + (1-\varphi^*) B(T_m) \tag{7.2}$$

where

$$B(\xi_{1u}) = \bar{Y}\left(\frac{1}{u} - \frac{1}{u'}\right)\left(\frac{3}{8}C_z^2 - \frac{1}{2}\rho_{yz}C_yC_z\right)$$

and $B(T_m)$ is defined in section 4.

Theorem 7.2.

The mean square error of the estimator T^* to the first order of approximations is obtained as

$$M(T^*) = \varphi^{*2} M(\xi_{1u}) + (1-\varphi^*)^2 M(T_m) + 2\varphi^*(1-\varphi^*)C^* \tag{7.3}$$

where

$$M(\xi_{1u}) = E(\xi_{1u} - \bar{Y})^2 = \left[\left(\frac{1}{u} - \frac{1}{u'}\right)\left(\frac{1}{4} - \rho_{yz}\right) + \left(\frac{1}{u} - \frac{1}{N}\right)\right]S_y^2 \tag{7.4}$$

$$C^* = E\left[(\xi_{1u} - \bar{Y})(T_m - \bar{Y})\right] = -\frac{S_y^2}{N} \tag{7.5}$$

and $M(T_m)$ is defined in section 4.

Since the mean square error of the estimator T^* in equation (7.3) is the function of unknown constant φ^*, it is minimized with respect to φ^*, and subsequently the optimum value of φ^* is obtained as

$$\varphi^*_{opt} = \frac{M(T_m) - C^*}{M(\xi_{1u}) + M(T_m) - 2C^*}. \tag{7.6}$$

Now substituting the value of φ^*_{opt} in equation (7.6), we get the optimum mean square error of the estimator T^* as

$$M(T^*)_{opt} = \frac{M(\xi_{1u}).M(T_m)-C^{*2}}{M(\xi_{1u})+M(T_m)-2C^*}, \tag{7.7}$$

Further, substituting the values from equations (4.7), (7.4) and (7.5) in equation (7.7), we get the simplified value of $M(T^*)_{opt}$, which is given below:

$$M(T^*)_{opt} = \frac{a_3^*+\mu^* a_2^*+\mu^{*2}a_1}{a_6^*+\mu^* a_5^*+\mu^{*2}a_4}\frac{S_y^2}{n} \tag{7.8}$$

where

$a_2^* = ad+cb^*-k^2f^2$, $a_3^* = b^*d$, $a_5^* = a-b^*+d-2kf$, $a_6 = b^*$, $b^* = a_0+1$, a_1 and a_4 are defined in section 5.

To determine the optimum values of μ^*, we minimized the mean square error of the estimator T^* given in equation (7.8) with respect to μ^*, which results in quadratic equation in μ^* and the respective solutions of μ^*, say $\hat{\mu}^*$, are given below:

$$p_1^*\mu^{*2}+2p_2^*\mu^*+p_3^*=0 \tag{7.9}$$

$$\hat{\mu}^* = \frac{-p_2^* \pm \sqrt{p_2^{*2}-p_1^*p_3^*}}{p_1^*} \tag{7.10}$$

where

$$p_1^* = a_1a_5^*-a_2^*a_4,\ p_2^* = a_1a_6^*-a_3^*a_4 \text{ and } p_3^* = a_2^*a_6^*-a_3^*a_5^*.$$

Substituting the admissible value of $\hat{\mu}^*$, say $\mu^{*(0)}$, from equation (7.10) into equation (7.8), we have the optimum value of the mean square error of the estimator T^*, which is shown below:

$$M(T^{*0})_{opt} = \frac{a_3^*+\mu^{*(0)}a_2^*+\mu^{*(0)2}a_1}{a_6^*+\mu^{*(0)}a_5^*+\mu^{*(0)2}a_4}\frac{S_y^2}{n}\ . \tag{7.11}$$

Case 2: When non-response occurs only at second occasion

When non-response occurs only at current (second) occasion, the estimator for the mean $\bar{Y}$ at current occasion may be obtained as

$$T^{**} = \varphi^{**}T_u+\left(1-\varphi^{**}\right)\xi_{1m} \tag{7.12}$$

where

$$\xi_{1m} = \bar{y}_m \exp\left(\frac{\bar{x}_n-\bar{x}_m}{\bar{x}_n+\bar{x}_m}\right)\left(\frac{\bar{z}_n'}{\bar{z}_m}\right) \text{ and } T_u \text{ is defined in section 3,}$$

where φ^{**} $(0 \le \varphi^{**} \le 1)$ is the unknown constant (scalar) to be determined under certain criterions.

7.2. Properties of the estimator T**

Since the estimator T^{**} is exponential type estimator, it is biased for the population mean $\bar{Y}$. The bias B (.) and the mean square error M (.) of the estimator T^{**} are derived up to the first order of approximations similar to that of the estimator T.

Theorem 7.3.

The bias of the estimator T^{**} to the first order of approximations is obtained as

$$B(T^{**}) = \varphi^{**}B(T_u)+(1-\varphi^{**})B(\xi_{1m}) \tag{7.13}$$

where

$$B(\xi_{1m}) = \bar{Y}\left\{\left(\frac{1}{m}-\frac{1}{n}\right)\left(\frac{3}{8}C_x^2+\frac{1}{2}\rho_{xz}C_xC_z-\frac{1}{2}\rho_{xy}C_yC_x\right)+\left(\frac{1}{m}-\frac{1}{n'}\right)\left(C_z^2-\rho_{yz}C_yC_z\right)\right\}$$

and $B(T_u)$ is defined in section 4.

Theorem 7.4.

The mean square error of the estimator T^{**} to the first order of approximations is obtained as

$$M(T^{**}) = \varphi^{**2}M(T_u)+\left(1-\varphi^{**}\right)^2 M(\xi_{1m})+2\varphi^{**}\left(1-\varphi^{**}\right)C^{**} \tag{7.14}$$

where

$$M(\xi_{1m}) = E\left(\xi_{1m}-\bar{Y}\right)^2 = \left[\left\{\left(\frac{1}{m}-\frac{1}{n}\right)\left(\frac{1}{4}+\rho_{xz}-\rho_{yx}\right)\right\}+\left(\frac{1}{m}-\frac{1}{N}\right)+\left(\frac{1}{m}-\frac{1}{n'}\right)\left(1-2\rho_{yz}\right)\right]S_y^2 \tag{7.15}$$

$$C^{**} = E\left[\left(T_u-\bar{Y}\right)\left(\xi_{1m}-\bar{Y}\right)\right] = -\frac{S_y^2}{N} \tag{7.16}$$

and $M(T_u)$ is defined in section 4.

Since the mean square error of the estimator T^{**} in equation (7.14) is the function of unknown constant φ^{**}, it is minimized with respect to φ^{**} and subsequently the optimum value of φ^{**} is obtained as

$$\varphi^{**}_{opt} = \frac{M(\xi_{1m})-C^{**}}{M(T_u)+M(\xi_{1m})-2C^{**}}. \tag{7.17}$$

Now substituting the value of φ_{opt}^{**} in equation (7.17), we get the optimum mean square error of T^{**} as

$$M(T^{**})_{opt} = \frac{M(T_u).M(\xi_{1m})-C^{**2}}{M(T_u)+M(\xi_{1m})-2C^{**}} \tag{7.18}$$

Further, substituting the values from equations (4.6), (7.15) and (7.16) in equation (7.18), we get the simplified value of $M(T^{**})_{opt}$ which is given below:

$$M(T^{**})_{opt} = \frac{a_3^{**}+\mu^{**}a_2^{**}+\mu^{**2}a_1^{**}}{a_6+\mu^{**}a_5^{**}+\mu^{**2}a_4^{**}}\frac{S_y^2}{n} \tag{7.19}$$

where

$a_1^{**} = ac^{*}+k^2f^2$, $a_2^{**} = ad^{*}+c^{*}b-k^2f^2$, $a_3^{**} = bd^{*}$, $a_4^{**} = c^{*}-a+2kf$, $a_5^{**} = a-b+d^{*}-2kf$, $a_6 = b$, $c^{*} = f+c_1+t_2d_1$, $d^{*} = (1-f)+d_1(1-t_2)$.

To determine the optimum values of μ^{**}, we minimized the mean square error of the estimator T^{*} given in equation (7.19) with respect to μ^{**}, which results in quadratic equation in μ^{**}, and the respective solutions of μ^{**}, say $\hat{\mu}^{**}$, are given below:

$$p_1^{**}\mu^{**2}+2p_2^{**}\mu^{**}+p_3^{**}=0 \tag{7.20}$$

$$\hat{\mu}^{**} = \frac{-p_2^{**} \pm \sqrt{p_2^{**2}-p_1^{**}p_3^{**}}}{p_1^{**}} \tag{7.21}$$

where

$p_1^{**} = a_1^{**}a_5^{**}-a_2^{**}a_4^{**}$, $p_2^{**} = a_1^{**}a_6-a_3^{**}a_4^{**}$ and $p_3^{**} = a_2^{**}a_6-a_3^{**}a_5^{**}$.

Substituting the admissible values of $\hat{\mu}^{**}$, say $\mu^{**(0)}$, from equation (7.21) into equation (7.19), we have the optimum value of the mean square error of T^{**}, which is shown below:

$$M(T^{**0})_{opt} = \frac{a_3^{**}+\mu^{**(0)}a_2^{**}+\mu^{**(0)2}a_1^{**}}{a_6+\mu^{**(0)}a_5^{**}+\mu^{**(0)2}a_4^{**}}\frac{S_y^2}{n}. \tag{7.22}$$

8. Comparison of efficiencies

The percentage relative loss in efficiencies of the estimator T, T^{*} and T^{**} is obtained with respect to the similar estimator and natural successive sampling estimator when the non-response is not observed on any occasion. The estimator ξ_1 is defined under similar circumstances as the estimator T but under complete response, whereas the estimator ξ_2 is the natural successive sampling estimator, and they are given as

$$\xi_j = \psi_j\xi_{ju}+\left(1-\psi_j\right)\xi_{jm} \quad ; (j= 1, 2) \tag{8.1}$$

where

$$\xi_{1u} = \bar{y}_u \exp\left(\frac{\bar{z}_u' - \bar{z}_u}{\bar{z}_u' + \bar{z}_u}\right),\ \xi_{2u} = \bar{y}_u,\ \xi_{1m} = \bar{y}_m \exp\left(\frac{\bar{x}_n - \bar{x}_m}{\bar{x}_n + \bar{x}_m}\right)\left(\frac{\bar{z}_n'}{\bar{z}_m}\right),\ \xi_{2m} = \bar{y}_m + \beta_{yx}(\bar{x}_n - \bar{x}_m)$$

Proceeding on a similar line as discussed for the estimator T the optimum mean square errors of the estimators ξ_j (j=1,2) are derived as

$$M(\xi_1^0)_{opt} = \left[\frac{b_3 + \mu' b_2 + \mu'^2 b_1}{b_6 + \mu' b_5 + \mu'^2 b_4}\right]\frac{S_y^2}{n} \tag{8.2}$$

and

$$M(\xi_2^0)_{opt} = \left[\frac{1}{2}\left\{1 + \sqrt{(1-\rho_{xy}^2)}\right\} - f\right]\frac{S_y^2}{n}. \tag{8.3}$$

where

$\mu' = \dfrac{-q_2 \pm \sqrt{q_2^2 - q_1 q_3}}{q_1}$ (fraction of the fresh sample for the estimator ξ_1),

$b_1 = ac^* + k^2 f^2$, $b_2 = ad^* + c^* b^* - k^2 f^2$, $b_3 = b^* d^*$, $b_4 = c^* - a + 2kf$, $b_5 = a - b^* + d^* - 2kf$, $b_6 = b^*$, $q_1 = b_1 b_5 - b_2 b_4$, $q_2 = b_1 b_6 - b_3 b_4$ and $q_3 = b_2 b_6 - b_3 b_5$.

Remark 8.1.

To compare the performances of the estimators T, T^* and T^{**} with respect to the estimators ξ_j (j=1, 2), we introduce the following assumptions:

(i) $\rho_{xz} = \rho_{yz}$, which is an intuitive assumption, also considered by Cochran (1977) and Feng and Zou (1997), (ii) $W = W^*$ (iii) $f_1 = f_2$.

The percentage relative losses in the precision of the estimators T, T^* and T^{**} with respect to ξ_j (j=1, 2) under their respective optimality conditions are given by

$$L_j = \frac{M\left(T^{(0)}\right)_{opt} - M\left(\xi_j\right)_{opt}}{M\left(T^{(0)}\right)_{opt}} \times 100,\ L_j^* = \frac{M\left(T^{*(0)}\right)_{opt} - M\left(\xi_j\right)_{opt}}{M\left(T^{*(0)}\right)_{opt}} \times 100$$

$$\text{and } L_j^{**} = \frac{M\left(T^{**(0)}\right)_{opt} - M\left(\xi_j\right)_{opt}}{M\left(T^{**(0)}\right)_{opt}} \times 100;\ (j=1, 2)$$

For N = 5000, n' = 1000, u' = 1000, n = 500, t_1=0.50 ,t_2=0.50, f=0.1 and different choices of f_1, ρ_{yx} and ρ_{yz} , Tables 1-6 give the optimum values of

$\mu^{(0)}, \mu^{*(0)}, \mu^{**(0)}$ and percentage relative losses L_j, L_j^* and L_j^{**} (j=1, 2) in the precision of the estimators T ,T^* and T^{**} with respect to estimators ξ_j(j=1, 2).

Table 1. Percentage relative loss L_1 in the precision of the estimator T with respect to ξ_1

W			0.05		0.10		0.15		0.20	
ρ_{yx}	f_2	ρ_{yz}	$\mu^{(0)}$	L_1	$\mu^{(0)}$	L_1	$\mu^{(0)}$	L_1	$\mu^{(0)}$	L_1
0.6	1.5	0.6	0.8640	3.2181	0.7182	5.9332	0.5829	8.2004	0.4580	10.0748
		0.7	0.4617	2.5435	0.3950	4.7480	0.3298	6.6502	0.2666	8.2850
		0.8	0.3161	2.5098	0.2740	4.7029	0.2320	6.6173	0.1903	8.2870
		0.9	0.2404	2.7911	0.2106	5.2223	0.1803	7.3438	0.1495	9.1988
	2.0	0.6	0.7182	5.9332	0.4580	10.0748	0.2385	12.8584	0.0558	14.6753
		0.7	0.3950	4.7480	0.2666	8.2850	0.1472	10.8815	0.0385	12.7754
		0.8	0.2740	4.7029	0.1903	8.2870	0.1087	11.0137	0.0312	13.0986
		0.9	0.2106	5.2223	0.1495	9.1988	0.0880	12.2538	0.0277	14.6331
0.8	1.5	0.6	0.5890	2.4511	0.5527	4.6664	0.5179	6.6733	0.4844	8.4959
		0.7	0.4775	2.4663	0.4511	4.7018	0.4254	6.7339	0.4004	8.5863
		0.8	0.3985	2.6439	0.3787	5.0344	0.3591	7.2037	0.3398	9.1796
		0.9	0.3381	3.0190	0.3234	5.7243	0.3085	8.1616	0.2936	10.3684
	2.0	0.6	0.5527	4.6664	0.4844	8.4959	0.4214	11.6712	0.3636	14.3348
		0.7	0.4511	4.7018	0.4004	8.5863	0.3525	11.8338	0.3077	14.5833
		0.8	0.3787	5.0344	0.3398	9.1796	0.3024	12.6435	0.2667	15.5811
		0.9	0.3234	5.7243	0.2936	10.3684	0.2640	14.2120	0.2353	17.4529

Table 2. Percentage relative loss L_2 in the precision of the estimator T with respect to ξ_2

W			0.05		0.10		0.15		0.20	
ρ_{yx}	f_2	ρ_{yz}	$\mu^{(0)}$	L_2	$\mu^{(0)}$	L_2	$\mu^{(0)}$	L_2	$\mu^{(0)}$	L_2
0.6	1.5	0.6	0.8640	-6.7964	0.7182	-3.8003	0.5829	-1.2986	0.4580	0.7698
		0.7	0.4617	-19.2197	0.3950	-16.5229	0.3298	-14.1959	0.2666	-12.1961
		0.8	0.3161	-38.1243	0.2740	-35.0170	0.2320	-32.3048	0.1903	-29.9392
		0.9	0.2404	-65.7067	0.2106	-61.5624	0.1803	-57.9460	0.1495	-54.7839
	2.0	0.6	0.7182	-3.8003	0.4580	0.7698	0.2385	3.8414	0.0558	5.8463
		0.7	0.3950	-16.5229	0.2666	-12.1961	0.1472	-9.0197	0.0385	-6.7029
		0.8	0.2740	-35.0170	0.1903	-29.9392	0.1087	-26.0759	0.0312	-23.1220
		0.9	0.2106	-61.5624	0.1495	-54.7839	0.0880	-49.5762	0.0277	-45.5204
0.8	1.5	0.6	0.5890	2.0316	0.5527	4.2565	0.5179	6.2720	0.4844	8.1025
		0.7	0.4775	-11.1328	0.4511	-8.5855	0.4254	-6.2701	0.4004	-4.1593
		0.8	0.3985	-29.5093	0.3787	-26.3293	0.3591	-23.4436	0.3398	-20.8151
		0.9	0.3381	-56.2099	0.3234	-51.8524	0.3085	-47.9266	0.2936	-44.3720
	2.0	0.6	0.5527	4.2565	0.4844	8.1025	0.4214	11.2914	0.3636	13.9665
		0.7	0.4511	-8.5855	0.4004	-4.1593	0.3525	-0.4591	0.3077	2.6738
		0.8	0.3787	-26.3293	0.3398	-20.8151	0.3024	-16.2073	0.2667	-12.2995
		0.9	0.3234	-51.8524	0.2936	-44.3720	0.2640	-38.1811	0.2353	-32.9609

Table 3. Percentage relative loss L_1^* in the precision of the estimator T^* with respect to ξ_1

W			0.05		0.10		0.15		0.20	
ρ_{yx}	f_2	ρ_{yz}	$\mu^{*(0)}$	L_1^*	$\mu^{*(0)}$	L_1^*	$\mu^{*(0)}$	L_1^*	$\mu^{*(0)}$	L_1^*
0.6	1.5	0.6	*	-	*	-	*	-	*	-
		0.7	0.5489	0.1481	0.5667	0.2827	0.5832	0.4057	0.5984	0.5185
		0.8	0.3748	0.3426	0.3911	0.6617	0.4065	0.9596	0.4212	1.2385
		0.9	0.2828	0.5457	0.2961	1.0599	0.3089	1.5452	0.3212	2.0041
	2.0	0.6	*	-	*	-	*	-	*	-
		0.7	0.5667	0.2827	0.5984	0.5185	0.6258	0.7180	0.6497	0.8890
		0.8	0.3911	0.6617	0.4212	1.2385	0.4484	1.7458	0.4732	2.1954
		0.9	0.2961	1.0599	0.3212	2.0041	0.3445	2.8508	0.3663	3.6142
0.8	1.5	0.6	0.6356	0.1025	0.6443	0.1997	0.6525	0.2920	0.6604	0.3798
		0.7	0.5141	0.2075	0.5234	0.4057	0.5324	0.5951	0.5411	0.7764
		0.8	0.4277	0.3345	0.4368	0.6552	0.4456	0.9628	0.4542	1.2582
		0.9	0.3613	0.4977	0.3698	0.9754	0.3780	1.4345	0.3861	1.8759
	2.0	0.6	0.6443	0.1997	0.6604	0.3798	0.6752	0.5430	0.6887	0.6916
		0.7	0.5234	0.4057	0.5411	0.7764	0.5574	1.1163	0.5727	1.4292
		0.8	0.4368	0.6552	0.4542	1.2582	0.4705	1.8150	0.4859	2.3308
		0.9	0.3698	0.9754	0.3861	1.8759	0.4016	2.7098	0.4163	3.4842

Note: '*' indicates $\mu^{*(0)}$ does not exist.

Table 4. Percentage relative loss L_2^* in the precision of the estimator T^* with respect to ξ_2

W			0.05		0.10		0.15		0.20	
ρ_{yx}	f_2	ρ_{yz}	$\mu^{*(0)}$	L_2^*	$\mu^{*(0)}$	L_2^*	$\mu^{*(0)}$	L_2^*	$\mu^{*(0)}$	L_2^*
0.6	1.5	0.6	*	-	*	-	*	-	*	-
		0.7	0.5489	-22.1500	0.5667	-21.9853	0.5832	-21.8349	0.5984	-21.6969
		0.8	0.3748	-41.1948	0.3911	-40.7427	0.4065	-40.3205	0.4212	-39.9255
		0.9	0.2828	-69.5345	0.2961	-68.6579	0.3089	-67.8305	0.3212	-67.0482
	2.0	0.6	*	-	*	-	*	-	*	-
		0.7	0.5667	-21.9853	0.5984	-21.6969	0.6258	-21.4529	0.6497	-21.2436
		0.8	0.3911	-40.7427	0.4212	-39.9255	0.4484	-39.2068	0.4732	-38.5697
		0.9	0.2961	-68.6579	0.3212	-67.0482	0.3445	-65.6050	0.3663	-64.3036
0.8	1.5	0.6	0.6356	-0.3271	0.6443	-0.2294	0.6525	-0.1367	0.6604	-0.0486
		0.7	0.5141	-13.7064	0.5234	-13.4806	0.5324	-13.2648	0.5411	-13.0583
		0.8	0.4277	-32.5814	0.4368	-32.1549	0.4456	-31.7456	0.4542	-31.3527
		0.9	0.3613	-60.2711	0.3698	-59.5015	0.3780	-58.7621	0.3861	-58.0511
	2.0	0.6	0.6443	-0.2294	0.6604	-0.0486	0.6752	0.1153	0.6887	0.2646
		0.7	0.5234	-13.4806	0.5411	-13.0583	0.5574	-12.6709	0.5727	-12.3144
		0.8	0.4368	-32.1549	0.4542	-31.3527	0.4705	-30.6119	0.4859	-29.9258
		0.9	0.3698	-59.5015	0.3861	-58.0511	0.4016	-56.7079	0.4163	-55.4606

Note: '*' indicates $\mu^{*(0)}$ does not exist.

Table 5. Percentage relative loss L_1^{**} in the precision of the estimator T^{**} with respect to ξ_1

W			0.05		0.10		0.15		0.20	
ρ_{yx}	f_2	ρ_{yz}	$\mu^{**(0)}$	L_1^{**}	$\mu^{**(0)}$	L_1^{**}	$\mu^{**(0)}$	L_1^{**}	$\mu^{**(0)}$	L_1^{**}
0.6	1.5	0.6	0.8515	3.2070	0.6634	5.8247	0.4538	7.7922	0.2206	9.0181
		0.7	0.4376	2.3271	0.3382	4.1649	0.2312	5.5017	0.1163	6.3162
		0.8	0.2964	2.0922	0.2305	3.7242	0.1600	4.9027	0.0847	5.6273
		0.9	0.2250	2.1709	0.1772	3.8495	0.1257	5.0602	0.0707	5.8180
	2.0	0.6	0.6634	5.8247	0.2206	9.0181	*	-	*	-
		0.7	0.3382	4.1649	0.1163	6.3162	*	-	*	-
		0.8	0.2305	3.7242	0.0847	5.6273	*	-	*	-
		0.9	0.1772	3.8495	0.0707	5.8180	*	-	*	-
0.8	1.5	0.6	0.5785	2.3221	0.5297	4.3575	0.4801	6.1286	0.4298	7.6552
		0.7	0.4668	2.2289	0.4284	4.1749	0.3891	5.8623	0.3490	7.3121
		0.8	0.3883	2.2786	0.3573	4.2545	0.3254	5.9574	0.2925	7.4125
		0.9	0.3288	2.4916	0.3038	4.6280	0.2776	6.4507	0.2504	7.9947
	2.0	0.6	0.5297	4.3575	0.4298	7.6552	0.3272	10.0446	0.2222	11.6475
		0.7	0.4284	4.1749	0.3490	7.3121	0.2666	9.5722	0.1818	11.0819
		0.8	0.3573	4.2545	0.2925	7.4125	0.2245	9.6656	0.1538	11.1591
		0.9	0.3038	4.6280	0.2504	7.9947	0.1934	10.3593	0.1333	11.9082

Note: '*' indicates $\mu^{**(0)}$ does not exist.

Table 6. Percentage relative loss L_2^{**} in the precision of the estimator T^{**} with respect to ξ_2

W			0.05		0.10		0.15		0.20	
ρ_{yx}	f_2	ρ_{yz}	$\mu^{**(0)}$	L_2^{**}	$\mu^{**(0)}$	L_2^{**}	$\mu^{**(0)}$	L_2^{**}	$\mu^{**(0)}$	L_2^{**}
0.6	1.5	0.6	0.8515	-6.8086	0.6634	-3.9201	0.4538	-1.7490	0.2206	-0.3963
		0.7	0.4376	-19.4844	0.3382	-17.2361	0.2312	-15.6008	0.1163	-14.6045
		0.8	0.2964	-38.7159	0.2305	-36.4038	0.1600	-34.7341	0.0847	-33.7074
		0.9	0.2250	-66.7641	0.1772	-63.9026	0.1257	-61.8388	0.0707	-60.5469
	2.0	0.6	0.6634	-3.9201	0.2206	-0.3963	*	-	*	-
		0.7	0.3382	-17.2361	0.1163	-14.6045	*	-	*	-
		0.8	0.2305	-36.4038	0.0847	-33.7074	*	-	*	-
		0.9	0.1772	-63.9026	0.0707	-60.5469	*	-	*	-
0.8	1.5	0.6	0.5785	1.9021	0.5297	3.9462	0.4801	5.7249	0.4298	7.2581
		0.7	0.4668	-11.4032	0.4284	-9.1858	0.3891	-7.2633	0.3490	-5.6113
		0.8	0.3883	-29.9953	0.3573	-27.3668	0.3254	-25.1016	0.2925	-23.1658
		0.9	0.3288	-57.0593	0.3038	-53.6183	0.2776	-50.6823	0.2504	-48.1954
	2.0	0.6	0.5297	3.9462	0.4298	7.2581	0.3272	9.6578	0.2222	11.2676
		0.7	0.4284	-9.1858	0.3490	-5.6113	0.2666	-3.0360	0.1818	-1.3158
		0.8	0.3573	-27.3668	0.2925	-23.1658	0.2245	-20.1687	0.1538	-18.1818
		0.9	0.3038	-53.6183	0.2504	-48.1954	0.1934	-44.3867	0.1333	-41.8919

Note: '*' indicates $\mu^{**(0)}$ does not exist.

9. Interpretations of results

The following conclusions may be drawn from Tables 1-6:

(1) From Tables 1 and 5 it is clear that

(a) For the fixed values of W, ρ_{yx} and f_2, the values of $\mu^{(0)}, \mu^{**(0)}$ decrease with the increasing values of ρ_{yz}. This implies that the higher the value of ρ_{yz}, the lower the fraction of a fresh sample required on the current occasion.

(b) For the fixed values of W, ρ_{yx} and ρ_{yz}, the values of $\mu^{(0)}, \mu^{**(0)}$ decrease and L_1, L_1^{**} increase with the increasing values of f_2.

(c) For the fixed values of W, ρ_{yz} and f_2, no pattern is observed with the increasing values of ρ_{yx}.

(d) For the fixed values of f_2, ρ_{yz} and ρ_{yx}, the values of $\mu^{(0)}, \mu^{**(0)}$ decrease and L_1, L_1^{**} increase with the increasing values of W. This behaviour shows that with the higher non-response rate one may require to draw the smaller sample on the current occasion, which reduces the cost of a survey.

(2) From Tables 2 and 6 it may be seen that

(a) For the fixed values of W, ρ_{yx} and f_2, the values of $\mu^{(0)}, \mu^{**(0)}$ and L_2, L_2^{**} decrease with the increasing values of $\rho_{yz..}$ This implies that if one uses the information on highly correlated auxiliary variable, there is a significant gain in the precision of estimates.

(b) For the fixed values of W, ρ_{yx} and ρ_{yz}, the values of $\mu^{(0)}, \mu^{**(0)}$ decrease and L_2, L_2^{**} increase with the increasing values of f_2.

(c) For the fixed values of W, ρ_{yz} and f_2, the values of L_2, L_2^{**} increase with the increasing values of ρ_{yx}.

(d) For the fixed values of f_2, ρ_{yz} and ρ_{yx}, the values of $\mu^{(0)}, \mu^{**(0)}$ decrease and L_2, L_2^{**} increase with the increasing values of W. This pattern shows that the higher the non-response rate, the greater the loss. This behaviour is practically justified.

(3) From Table 3 it is clear that

(a) For the fixed values of W, ρ_{yx} and f_2, the values of $\mu^{*(0)}$ decrease and L_1^* increase with the increasing values of ρ_{yz}. This behaviour indicates that if the information on highly correlated auxiliary variable is available, it plays an important role in improving the precision of estimates.

(b) For the fixed values of W, ρ_{yx} and ρ_{yz}, the values of $\mu^{*(0)}$ and L_1^* increase with the increasing values of f_2.

(c) For the fixed values of W, ρ_{yz} and f_2, no pattern is visible with the increasing values of ρ_{yx}.

(d) For the fixed values of f_2, ρ_{yz} and ρ_{yx}, the values of $\mu^{*(0)}$ and L_1^* increase with the increasing values of W.

(4) From Table 4 it may be seen that

(a) For the fixed values of W, ρ_{yx} and f_2, the values of $\mu^{*(0)}$ and L_2^* decrease with the increasing values of $\rho_{yz.}$ This implies that negative loss is observed due to the presence of high correlation between the auxiliary variables. This behaviour is highly desirable.

(b) For the fixed values of W, ρ_{yx} and ρ_{yz}, the values of $\mu^{*(0)}$ and L_2^* increase with the increasing values of $f_{2.}$ This indicates that if a smaller size of sub-sample is chosen, the loss in precision increases, as it was expected.

(c) For the fixed values of W, ρ_{yz} and f_2 no pattern is seen with the increasing values of ρ_{yx}.

(d) For the fixed values of f_2, ρ_{yz} and ρ_{yx}, the values of $\mu^{*(0)}$ and L_2^* increase with the increasing values of W.

10. Conclusions and recommendations

It may be seen from the above tables that for all cases the percentage relative loss in precisions is observed wherever the optimum value of μ exists, when non-response occurs on both occasions. From Tables 1, 3 and 5, it is seen that the loss is present due to the presence of non-response on each occasion, but the negative impact of non-response is very low, which justifies the use of Hansen and Hurwitz (1946) technique in the proposed estimation procedures. From Tables 2, 4 and 6, when the proposed estimators are compared with the natural successive sampling estimator, substantial profit is visible, which justifies the intelligible use of auxiliary information in the form of exponential methods of estimation. Finally, looking at good behaviours of the proposed estimators one may recommend them to survey statisticians and practitioners for their practical applications.

Acknowledgements

Authors are thankful to the reviewers for their valuable suggestions, which enhanced the quality of this paper. Authors are also thankful to the Indian School of Mines, Dhanbad for providing financial assistance and necessary infrastructure to carry out the present research work.

REFERENCES

BAHL, S., TUTEJA, R. K., (1991). Ratio and Product type exponential estimator. Information and Optimization sciences, XII(I), 159–163.

CHOUDHARY, R. K., BATHAL, H. V. L., SUD, U. C., (2004). On Non-response in Sampling on Two Occasions. Journal of the Indian Society of Agricultural Statistics, 58(3) 331–343.

COCHRAN, W. G., (1977). Sampling Techniques, 3rd edition, John Wiley & Sons, New York.

DAS, A. K., (1982). Estimation of population ratio on two occasions. Journal of the Indian Society of Agricultural Statistics, 34, 1–9.

ECKLER, A. R., (1955). Rotation Sampling. Annals Mathematical Statistics, 664–685.

FABIAN, C. O., HYUNSHIK, L., (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. Survey Methodology, 26 (2), 183–188.

FENG, S., ZOU, G., (1997). Sample rotation method with auxiliary variable. Communication in Statistics Theory and Methods, 26 (6) 1497–1509.

GARCÍA LUENGO A. V., (2013). Improvement on the non-response in the population ratio of mean for current occasion in sampling on two occasions. Pakistan Journal of Statistics and Operation Research, 9(1), 25–51.

GUPTA, P. C., (1979). Sampling on two successive occasions. Jour. Statist. Res., 13, 7–16.

HANSEN, M. H., HURWITZ, W. N., (1946). The problem of the non-response in sample surveys. Journal of the American Statistical Association, 41, 517–529.

HANSEN, M. H., HURWITZ, W. N., MADOW, W. G., (1953). Sample surveys methods and theory, 1, 2, John Wiley and Sons, Inc., New York and London.

JESSEN, R. J., (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Agricultural Experiment Station Road Bulletin No. 304, Ames, USA, 1–104.

PATTERSON, H. D., (1950). Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society Series B, 12, 241–255.

RAO, J. N. K., GRAHAM, J. E., (1964). Rotation design for sampling on repeated occasions. Journal of the American Statistical Association, 59, 492–509.

REDDY, V. N., (1978). A study on the use of prior knowledge on certain population parameters. Sankhya, 40C, 29–37.

SEN, A. R., (1971). Successive sampling with two auxiliary variables. Sankhya, Series B, 33, 371–378.

SEN, A. R., (1972). Successive sampling with p ($p \geq 1$) auxiliary variables. Annals Mathematical Statistics, 43, 2031–2034.

SEN, A. R., (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. Biometrics, 29, 381–385.

SINGH, D., SINGH, B. D., (1965). Double sampling for stratification on successive occasions, Journal of the American Statistical Association, 60, 784–792.

SINGH, G. N., SINGH, V. K., (2001). On the use of auxiliary information in successive sampling. Journal of the Indian Society of Agricultural Statistics, 54(1), 1–12.

SINGH, G. N., PRIYANKA, K., (2007). Effect of non-response on current occasion in search of good rotation patterns on successive occasions. Statistics in Transition-New Series, 8(2), 273–292.

SINGH, G. N., PRASAD, S., (2011). Some rotation patterns in two-phase sampling, Statistics in Transition-New Series, 12(1), 25–44.

SINGH, G. N., HOMA, F., (2013). Effective rotation patterns in successive sampling over two occasions. Journal of Statistics Theory and Practice, 7(1), 146–155.

SINGH, G. N., HOMA, F., (2014). An improved estimation procedure in two-phase successive sampling. International Journal of applied mathematics and statistics, 52(2), 76–85.

SINGH, H. P., VISHWAKARMA, G. K., (2007). Modified exponential ratio and product estimators for finite population mean in double sampling. Australian Journal of Statistics, 36(3), 217–225.

SINGH, H. P., KUMAR, S., (2009). Multivariate indirect methods of estimation in presence of non-response in successive sampling. Metron, LXVII(2), 153–175.

SINGH, H. P., KUMAR, S., (2010). Estimation of population product in presence of non-response in successive sampling. Statistical Papers, 51(4), 975–996.

SINGH, H. P., KUMAR, S., BHOUGAL, S., (2011). Estimation of population mean in successive sampling by sub-sampling non-respondents. Journal of Modern Applied Statistical Methods, 10(1), 51–60.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 183–210

TRANSMUTED KUMARASWAMY DISTRIBUTION

Muhammad Shuaib Khan[1], **Robert King**[2], **Irene Lena Hudson**[3]

ABSTRACT

The Kumaraswamy distribution is the most widely applied statistical distribution in hydrological problems and many natural phenomena. We propose a generalization of the Kumaraswamy distribution referred to as the transmuted Kumaraswamy (TKw) distribution. The new transmuted distribution is developed using the quadratic rank transmutation map studied by Shaw et al. (2009). A comprehensive account of the mathematical properties of the new distribution is provided. Explicit expressions are derived for the moments, moment generating function, entropy, mean deviation, Bonferroni and Lorenz curves, and formulated moments for order statistics. The TKw distribution parameters are estimated by using the method of maximum likelihood. Monte Carlo simulation is performed in order to investigate the performance of MLEs. The flood data and HIV/ AIDS data applications illustrate the usefulness of the proposed model.

Key words: Kumaraswamy distribution, moments, order statistics, parameter estimation, maximum likelihood estimation.

1. Introduction

The Kumaraswamy probability distribution was originally proposed by Poondi Kumaraswamy (1980) for double bounded random processes for hydrological applications. The Kumaraswamy double bounded distribution denoted by $Kw(\alpha, \theta)$ distribution is a family of continuous probability distributions defined on the interval [0,1] with cumulative distribution function given by

$$G_{Kw}(x; \alpha, \theta) = 1 - (1 - x^{\alpha})^{\theta}\,, \tag{1}$$

and probability density function (pdf) corresponding to (1) given by

$$g_{Kw}(x; \alpha, \theta) = \alpha\theta x^{\alpha-1}(1 - x^{\alpha})^{\theta-1}\,, \tag{2}$$

[1] School of Mathematical and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia. E-mail: Muhammad.S.Khan@newcastle.edu.au.
[2] School of Mathematical and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia. E-mail: robert.king@newcastle.edu.au.
[3] School of Mathematical and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia. E-mails: Irene.Hudson@newcastle.edu.au, Irenelena.hudson@gmail.com.

where $\alpha > 0$ and $\theta > 0$ are the shape parameters. The *Kw* probability density function has the same basic properties as the beta distribution. According to Jones (2009) and Cordeiro et al. (2010, 2012) it depends in the same way as the beta distribution on the values of its parameters: it is unimodal for $\alpha > 1$ and $\theta > 1$; uniantimodal for $\alpha < 1$ and $\theta < 1$; increasing for $\alpha > 1$ and $\theta \leq 1$; decreasing for $\alpha \leq 1$ and $\theta > 1$ and constant for $\alpha = \theta = 1$. Jones (2009) investigated not only properties of the Kumaraswamy distribution but also some similarities and differences between the beta and *Kw* distributions. According to Jones (2009) the Kumaraswamy distribution has several advantages over the beta distribution, such as a simple normalizing constant, simple explicit formulae for the distribution and quantile functions which do not involve any special functions, a simple formula for random variable generation, explicit formulae for L-moments and simpler formulae for moments of order statistics. On the other hand, the beta distribution has simpler formulae for moments and the moment generating function, a one-parameter sub-family of symmetric distributions, simpler moment estimation and more ways of generating the distribution through physical processes.

The *Kw* distribution is applicable to a number of hydrological problems and many natural phenomena whose process values are bounded on both sides. Cordeiro and de Castro (2009) studied a new class of the Kumaraswamy generalized distributions (denoted by the *Kw*–G distribution) based on the Kumaraswamy distribution (denoted by *Kw* distribution). They derived almost all formulas for the probability characteristics of the *Kw*–G distribution. In hydrology and related areas, the *Kw* distribution has received considerable interest, see Cordeiro et al. (2010, 2012), Fletcher and Ponnambalam (1996), Ganji et al. (2006), Ponnambalam et al. (2001), Sundar and Subbiah (1989) and Seifi et al. (2000). According to Nadarajah (2008), many papers in the hydrological literature have used this distribution because it is deemed to be a "better alternative" to the beta distribution; see, for example, Koutsoyiannis and Xanthopoulos (1989).

This article introduces a new three-parameter distribution which is a generalized two-parameter Kumaraswamy distribution, called the transmuted Kumaraswamy distribution and denoted by TKw. Recently the generalization of parametric models by transforming an appropriate model, into a more general model, by adding a shape parameter has been intensively studied for many different families of lifetime distributions. Aryal and Tsokos et al. (2009, 2011) considered the following transmuted extreme value distributions: the transmuted Gumbel distribution to model climate data, the transmuted Weibull distribution and their applications to analyse real data sets. Recently Khan et al. (2013 a, b, c) developed the transmuted modified Weibull distribution, the transmuted generalized inverse Weibull distribution and the transmuted generalized exponential distribution. More recently Khan et al. (2014) studied the characteristics of the transmuted Inverse Weibull distribution. Ashour et al. (2013), Elbatal et al. (2013) and Aryal (2013) studied the transmuted Lomax distribution, the transmuted quasi Lindley distribution and the transmuted log-logistic distribution with a discussion on some properties of this family. The most

recent families of the transmuted Rayleigh distribution, the transmuted generalized Rayleigh distribution and the transmuted Lindley distribution are derived in studies of Merovici (2013 a, b) & (2014). More recently Ahmad et al. (2015) also studied the transmuted Kumaraswamy distribution and discussed some mathematical results.

Using the quadratic rank transmutation map proposed by Shaw et al. (2009), we develop the three-parameter TKw distribution. According to this approach a random variable X is said to have a transmuted distribution if its cumulative distribution function (cdf) satisfies the relationship

$$F(x) = (1+\lambda)G(x) - \lambda[G(x)]^2, \qquad |\lambda| \leq 1 \tag{3}$$

and

$$f(x) = g(x)[(1+\lambda) - 2\lambda G(x)], \tag{4}$$

where $G(x)$ is the cdf of the base distribution, $g(x)$ and $f(x)$ are the corresponding probability density functions (pdf) associated with $G(x)$ and $F(x)$, respectively. It is important to note that at $\lambda = 0$ we have the distribution of the base random variable.

The paper is organized as follows. In Section 2, we present the analytical shapes of the probability density and hazard functions of the model under study. A range of mathematical properties are considered in Section 3, specifically we demonstrate the quantile functions, moment estimation and moment generating function. Maximum likelihood estimates (MLEs) of the unknown parameters are discussed in Section 4. Entropy and mean deviations are derived in Section 5 and 6. The probability density function (pdf) of order statistics and their moments are derived in Section 7. In Section 8 we evaluate the performance of MLEs using Simulation. Two applications of the TKw distribution to the flood data and HIV/ AIDS data are illustrated in Section 9. In Section 10, concluding remarks are addressed.

2. Transmuted Kumaraswamy distribution

A random variable X is said to have transmuted Kw probability distribution denoted by $TKw(x;\alpha,\theta,\lambda)$ with parameters $\alpha,\theta > 0$ and $-1 \leq \lambda \leq 1$, $x \in (0,1)$, if its pdf and cdf are given by

$$f_{TKw}(x;\alpha,\theta,\lambda) = \alpha\theta x^{\alpha-1}(1-x^{\alpha})^{\theta-1}\{1-\lambda+2\lambda(1-x^{\alpha})^{\theta}\}, \tag{5}$$

and

$$F_{TKw}(x;\alpha,\theta,\lambda) = \left[1-(1-x^{\alpha})^{\theta}\right]\left[1+\lambda(1-x^{\alpha})^{\theta}\right], \tag{6}$$

where α and θ are the shape parameters and λ the transmuting parameter, representing the different patterns of the subject distribution. The TKw distribution approaches the Kw distribution when the transmuted parameter $\lambda = 0$.

If X has $TKw(\mathrm{x};\alpha,\theta,\lambda)$distribution, then the reliability function (RF), hazard function and cumulative hazard function corresponding to (5) are given by

$$R_{TKw}(x;\alpha,\theta,\lambda) = 1-\left[1-(1-x^{\alpha})^{\theta}\right]\left[1+\lambda(1-x^{\alpha})^{\theta}\right], \tag{7}$$

$$h_{TKw}(x;\alpha,\theta,\lambda) = \frac{\alpha\theta x^{\alpha-1}(1-x^{\alpha})^{\theta-1}\left\{1-\lambda+2\lambda(1-x^{\alpha})^{\theta}\right\}}{1-\left[1-(1-x^{\alpha})^{\theta}\right]\left[1+\lambda(1-x^{\alpha})^{\theta}\right]}, \tag{8}$$

and

$$H_{TKw}(x;\alpha,\theta,\lambda) = \int_0^x \frac{\alpha\theta x^{\alpha-1}(1-x^{\alpha})^{\theta-1}\left\{1-\lambda+2\lambda(1-x^{\alpha})^{\theta}\right\}}{1-\left[1-(1-x^{\alpha})^{\theta}\right]\left[1+\lambda(1-x^{\alpha})^{\theta}\right]}dx,$$

$$H_{TKw}(x;\alpha,\theta,\lambda) = -\ln\left|1-\left[1-(1-x^{\alpha})^{\theta}\right]\left[1+\lambda(1-x^{\alpha})^{\theta}\right]\right|. \tag{9}$$

Figure 1 shows some possible shapes of probability density function of the TKw distribution for selected values of the parameters α,θ and λ, and the hazard function of the TKw distribution for the same value of the parameters, respectively.

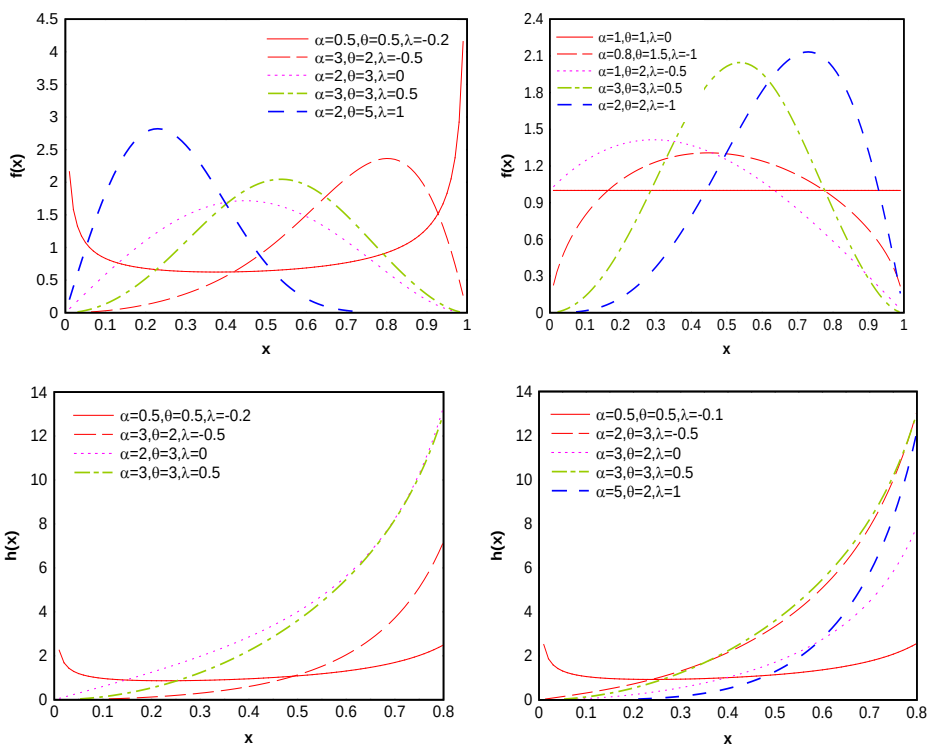


Figure 1. Plots of the TKw PDF & HF for some parameter values

Figure 1 also illustrates the TKw instantaneous failure rates. These failure rates are defined with different choices of parameters. For all choices of parameters the distribution has a monotonically increasing and decreasing behaviour of hazard rates. The transmuted beta and TKw distributions share their main special cases. T-Beta $(a, 1, \lambda)$ and $TKw(\alpha, 1, \lambda)$ distributions are both power function distributions. The TKw distribution approaches the transmuted uniform distribution $a = \theta = 1, \lambda \leq 1$, uniform distribution for $a = \theta = 1, \lambda = 0$.

3. Moments and quantiles

This section presents expressions for the moments, moment generating function and quantiles of the TKw distribution.

Theorem 1: If X has the $Tkw(x; \alpha, \theta, \lambda)$ distribution with $|\lambda| \leq 1$, then the k^{th} moment of X is given as follows

$$E\left(X^{\mathrm{k}}\right) = (1-\lambda)\theta\beta\left(\frac{\mathrm{k}}{\alpha}+1,\theta\right) + 2\lambda\theta\beta\left(\frac{\mathrm{k}}{\alpha}+1,2\theta\right).$$

Proof: Let X have a Tkw distribution, then the k^{th} moment of X is given as

$$E\left(X^{\mathrm{k}}\right) = \int_0^1 \alpha\theta x^{k+\alpha-1}(1-x^{\alpha})^{\theta-1}\{1-\lambda+2\lambda(1-x^{\alpha})^{\theta}\}dx,$$

$$E\left(X^{\mathrm{k}}\right) = (1-\lambda)\int_0^1 \alpha\theta x^{k+\alpha-1}(1-x^{\alpha})^{\theta-1}dx + 2\lambda\int_0^1 \alpha\theta x^{k+\alpha-1}(1-x^{\alpha})^{2\theta-1}\,dx.$$

Finally, we obtain the k^{th} moment of the Tkw distribution as

$$E\left(X^{\mathrm{k}}\right) = (1-\lambda)\theta\ \psi_{1,\mathrm{k}} + 2\lambda\theta\ \psi_{2,\mathrm{k}}, \tag{10}$$

where $\psi_{j,\mathrm{k}}$ is introduced for simplicity as

$$\psi_{j,\mathrm{k}} = \beta\left(\frac{\mathrm{k}}{\alpha}+1, j\theta\right), \quad j = 1, 2.$$

The expressions for the expected value and variance are

$$\mathrm{E}(X) = (1-\lambda)\theta\ \psi_{1,1} + 2\lambda\theta\ \psi_{2,1}, \tag{11}$$

and

$$Var(X) = (1-\lambda)\theta\ \psi_{1,2} + 2\lambda\theta\ \psi_{2,2} - \{(1-\lambda)\theta\ \psi_{1,1} + 2\lambda\theta\ \psi_{2,1}\}^2. \qquad (12)$$

The coefficient of variation, skewness and kurtosis measures can now be calculated using the following relationships

$$CV(X) = \frac{\sqrt{Var(X)}}{\mathrm{E}(X)},$$

$$\text{Skewness}(X) = \frac{\mathrm{E}\big(\mathrm{X} - \mathrm{E}(X)\big)^3}{[Var(X)]^{\frac{3}{2}}},$$

and

$$\text{Kurtosis}(X) = \frac{\mathrm{E}\big(\mathrm{X} - \mathrm{E}(X)\big)^4}{[Var(X)]^2}.$$

The mean, variance and coefficient of variation can be obtained using equations (11) and (12). The relationship between α and the mean is shown in Figure 2.

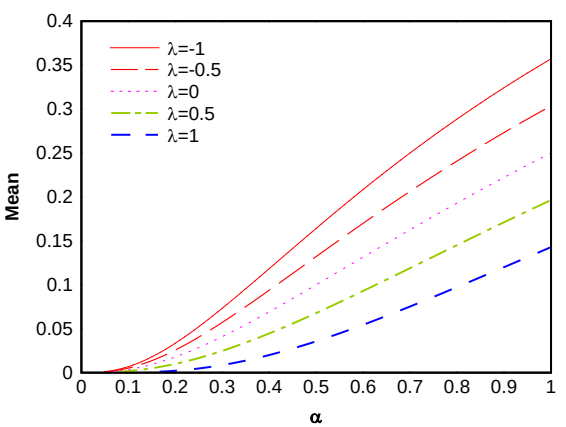
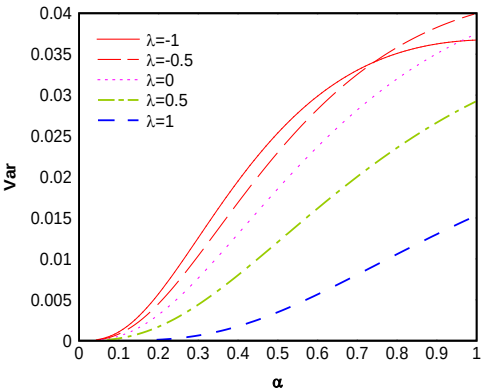
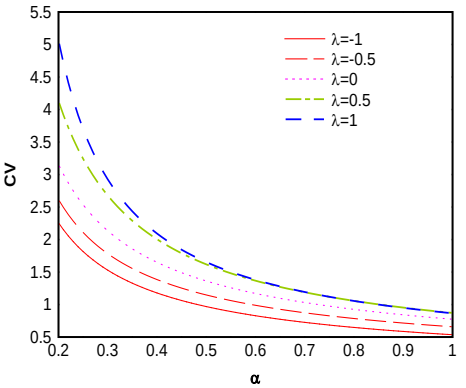


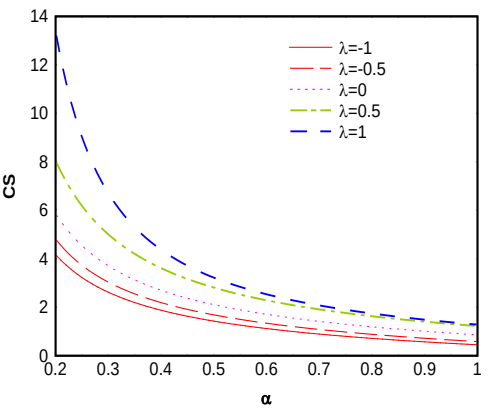
Figure 2. α vs mean



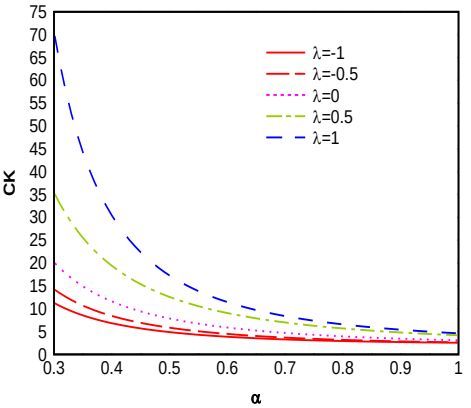
(a)α vs Var_{TKw}



(b) α vs CV_{TKw}



(c) α vs CS_{TKw}



(d) α vs CK_{TKw}

Figure 3. Plots of the TKw distribution α vs coefficients

The relationship between α and the variance is represented in Figure 3a. It is clear that, as α increases, the variance of the subject distribution also increases. The relationship between α and the coefficient of variation (CV_{TKw}) is shown in Figure 3b, which shows that as the parameter α increases, the coefficient of variation of the subject distribution is decreasing. Graphical representation of skewness (CS_{TKw}) and kurtosis (CK_{TKw}) when $\theta = 3$ and $\lambda = -1, -0.5, 0, 0.5, 1$ as a function of α are illustrated in Figures 3c and 3d, respectively. This shows the coefficients of skewness and kurtosis have a negative relationship with α.

Theorem 2: If X has the $Tkw(x;\alpha,\theta,\lambda)$ distribution with $|\lambda| \leq 1$, then the moment generating function of X say $\mathrm{M}_x(\mathrm{t})$ is given as follows

$$\mathrm{M}_x(\mathrm{t}) = (1-\lambda)\sum_{\mathrm{m}=0}^{\infty}\frac{\mathrm{t}^{\mathrm{m}}}{\mathrm{m}!}\theta\beta\left(\frac{\mathrm{m}}{\alpha}+1,\theta\right)+2\lambda\sum_{\mathrm{m}=0}^{\infty}\frac{\mathrm{t}^{\mathrm{m}}}{\mathrm{m}!}\theta\beta\left(\frac{\mathrm{m}}{\alpha}+1{,}2\theta\right).$$

Proof: Let X has a Tkw distribution, then the moment generating function of X is given as

$$\mathrm{M}_x(\mathrm{t}) = \int_0^1 \alpha\theta\exp(tx)\,x^{\alpha-1}(1-x^{\alpha})^{\theta-1}\{1-\lambda+2\lambda(1-x^{\alpha})^{\theta}\}dx.$$

Using the Taylor series of function e^{tx} reduces the above to

$$\mathrm{M}_x(\mathrm{t}) = (1-\lambda)\sum_{\mathrm{m}=0}^{\infty}\frac{\mathrm{t}^{\mathrm{m}}}{\mathrm{m}!}\int_0^1 \alpha\theta x^{m+\alpha-1}(1-x^{\alpha})^{\theta-1}dx + 2\lambda\sum_{\mathrm{m}=0}^{\infty}\frac{\mathrm{t}^{\mathrm{m}}}{\mathrm{m}!}\int_0^1 \alpha\theta x^{m+\alpha-1}(1-x^{\alpha})^{2\theta-1}\,dx.$$

By solving the above integral we obtain

$$\mathrm{M}_x(\mathrm{t}) = (1-\lambda)\sum_{\mathrm{m}=0}^{\infty}\frac{\mathrm{t}^{\mathrm{m}}}{\mathrm{m}!}\theta\beta\left(\frac{\mathrm{m}}{\alpha}+1,\theta\right)+2\lambda\sum_{\mathrm{m}=0}^{\infty}\frac{\mathrm{t}^{\mathrm{m}}}{\mathrm{m}!}\theta\beta\left(\frac{\mathrm{m}}{\alpha}+1{,}2\theta\right), \qquad (13)$$

which completes the proof.

Theorem 3: The qth quantile x_{q} of the Tkw random variable is given by

$$x_{\mathrm{q}} = \left[1-\left\{1-\frac{(1+\lambda)-\sqrt{(1+\lambda)^2-4\lambda\mathrm{q}}}{2\lambda}\right\}^{\frac{1}{\theta}}\right]^{\frac{1}{\alpha}}, \qquad 0 < q < 1. \qquad (14)$$

Proof: The qth quantile x_{q} of the Tkw distribution is defined as

$$q = \mathrm{P}(X \leq x_{\mathrm{q}}) = \mathrm{F}(x_{\mathrm{q}}), \qquad x_{\mathrm{q}} \geq 0.$$

Using the distribution function of the Tkw distribution we have

$$q = \mathrm{F}(x_{\mathrm{q}}) = (1+\lambda)\left[1-\left(1-x_q^{\alpha}\right)^{\theta}\right]-\lambda\left[1-\left(1-x_q^{\alpha}\right)^{\theta}\right]^2,$$

that is

$$\lambda\left[1-\left(1-x_q^{\alpha}\right)^{\theta}\right]^2-(1+\lambda)\left[1-\left(1-x_q^{\alpha}\right)^{\theta}\right]+q=0.$$

Consider this as a quadratic in $1-\left(1-x_q^{\alpha}\right)^{\theta}$ as

$$\Delta = 1+(2-4q)\lambda+\lambda^2.$$

It has roots $\frac{(1+\lambda)-\sqrt{\Delta}}{2\lambda}$. These exist if Δ is positive. Consider the following cases.

If $\lambda = -1$ then Δ reduces to

$$\Delta = 4q > 0, \quad if\ q > 0.$$

If $\lambda = 1$ then Δ takes the form

$$\Delta = 4(1-q) > 0, \quad if\ q > 0.$$

Otherwise for $-1 < \lambda < 1$, consider the roots of Δ, as a quadratic form in λ, are

$$\lambda = -1 + 2q \pm 2\sqrt{q^2 - q}\,,$$

Therefore, $q^2 - q < 0$ for $0 < q < 1$. So the only real roots could occure for $q = 0\ or\ 1$.

If $q = 0$ then roots = -1, contradiction between $(-1 < \lambda < 1)$, and if $q = 1$ then roots = 1, contradiction between $(-1 < \lambda < 1)$. Thus, there are no real roots of Δ as a quadratic in λ. Therefore, Δ has the same sign in the range $-1 \le \lambda \le 1$, hence $\Delta > 0$.

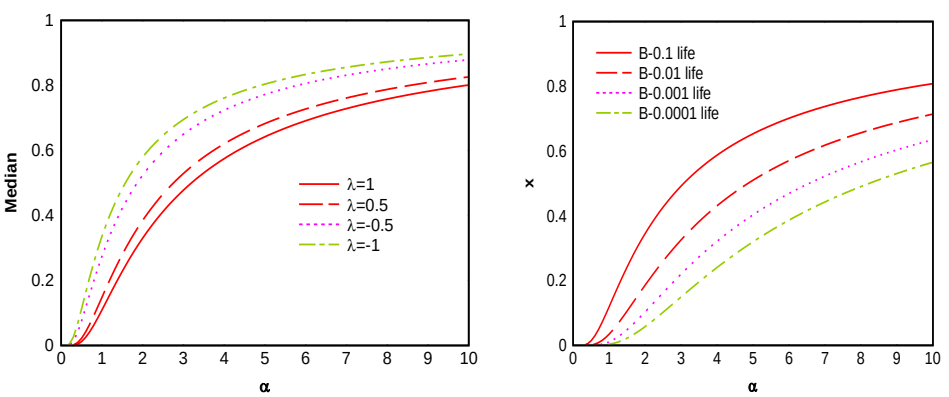
Since $\Delta \ge 0$, then

$$1 - \left(1 - x_q^{\alpha}\right)^{\theta} = \frac{(1+\lambda) - \sqrt{\Delta}}{2\lambda}.$$

Finally, we obtain the qth quantile x_q of the Tkw distribution as

$$x_q = \left[1 - \left\{1 - \frac{(1+\lambda) - \sqrt{\Delta}}{2\lambda}\right\}^{\frac{1}{\theta}}\right]^{\frac{1}{\alpha}},$$

which completes the proof.



(a) Median (b) B-lives

Figure 4. Plots of the Quantiles for the TKw distribution for some parameter values

Using the method of inversion we can generate random variables from the Tkw distribution. One can use equation (14) to generate random numbers when the parameters α, θ, λ are known.

Hence, the median of Tkw distribution is given by

$$x_{0.5} = \left[1 - \left\{1 - \frac{(1+\lambda) - \sqrt{1+\lambda^2}}{2\lambda}\right\}^{\frac{1}{\theta}}\right]^{\frac{1}{\alpha}}. \tag{15}$$

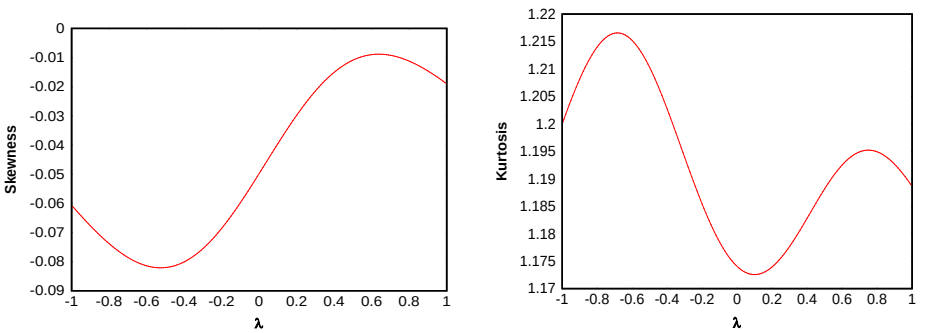
To demonstrate the effect of the shape parameter on the median and the percentile life (or B-life) as a function of α, they are calculated using quantiles and shown in Figure 4a and 4b, respectively. It can be concluded that as the shape parameter α increases the behaviour of median and percentile life (or B-life) also increases. To illustrate the effect of the transmuting parameter, on skewness and kurtosis, we also consider the measure based on quantiles. Skewness and kurtosis are calculated by using the relationship of Bowley ($\mathcal{B}$) and Moors ($\mathcal{M}$). The Bowley skewness is one of the earliest skewness measures defined by the average of the quantiles minus median, divided by the half of the interquantile range given by (see Kenney and Keeping (1962))

$$\mathcal{B} = \frac{Q(3/4) + Q(1/4) - 2Q(2/4)}{Q(3/4) - Q(1/4)}. \tag{16}$$

The Moors kurtosois is based on octiles and given by Moors (1998)

$$\mathcal{M} = \frac{Q(3/8) - Q(1/8) + Q(7/8) - Q(5/8)}{Q(6/8) - Q(2/8)}. \tag{17}$$

Figures 5a and 5b respectively illustrate the graphical representation of the Bowley ($\mathcal{B}$) skewness and Moors ($\mathcal{M}$) kurtosis as a function of the λ for $\alpha = \theta = 3$.



(a) Skewness (b) kurtosis

Figure 5. Plots of the Bowley skewness and Moors kurtosis for the TKw distribution

4. Parameter estimation

If the parameters of the TKw distribution are not known, then the maximum likelihood estimates, MLEs, of the parameters are given as follows.

Let $x_1, x_2, \ldots, x_n$ be the random samples of size n from the TKw distribution. Then the log-likelihood function of (5) is given by

$$\mathcal{L} = \mathrm{n}\ln\alpha + \mathrm{n}\ln\theta + (\alpha - 1)\sum_{\mathrm{i}=1}^{\mathrm{n}} \ln x_{\mathrm{i}} + (\theta - 1)\sum_{\mathrm{i}=1}^{\mathrm{n}} \ln(1 - {x_{\mathrm{i}}}^{\alpha}) + \sum_{\mathrm{i}=1}^{\mathrm{n}} \ln\{1 - \lambda + 2\lambda(1 - {x_{\mathrm{i}}}^{\alpha})^{\theta}\}. \quad (18)$$

By differentiating (18) with respect to α, θ and λ, then equating it to zero, we obtain the estimating equations

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \frac{\mathrm{n}}{\alpha} + \sum_{\mathrm{i}=1}^{\mathrm{n}} \ln x_{\mathrm{i}} - (\theta - 1)\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{{x_{\mathrm{i}}}^{\alpha}\ln x_{\mathrm{i}}}{1 - {x_{\mathrm{i}}}^{\alpha}} - \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{2\lambda\theta(1 - {x_{\mathrm{i}}}^{\alpha})^{\theta-1}{x_{\mathrm{i}}}^{\alpha}\ln x_{\mathrm{i}}}{\{1 - \lambda + 2\lambda(1 - {x_{\mathrm{i}}}^{\alpha})^{\theta}\}},$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\mathrm{n}}{\theta} + \sum_{\mathrm{i}=1}^{\mathrm{n}} \ln(1 - {x_{\mathrm{i}}}^{\alpha}) - \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{2\lambda(1 - {x_{\mathrm{i}}}^{\alpha})^{\theta}\ln(1 - {x_{\mathrm{i}}}^{\alpha})}{\{1 - \lambda + 2\lambda(1 - {x_{\mathrm{i}}}^{\alpha})^{\theta}\}},$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{-1 + 2(1 - {x_{\mathrm{i}}}^{\alpha})^{\theta}}{\{1 - \lambda + 2\lambda(1 - {x_{\mathrm{i}}}^{\alpha})^{\theta}\}}.$$

The maximum likelihood estimator $\widehat{\omega} = \left(\widehat{\alpha}, \widehat{\theta}, \widehat{\lambda}\right)^{\mathrm{T}}$ of $\omega = (\alpha, \theta, \lambda)^{\mathrm{T}}$ is obtained by solving this nonlinear system of equations. These solutions will yield the ML estimators $\widehat{\alpha}, \widehat{\theta}$ and $\widehat{\lambda}$. Here we used a nonlinear optimization algorithm such as the quasi-Newton algorithm to numerically maximize the log-likelihood function given in (18). The required numerical evaluations were implemented using the R language. Under the conditions that are fulfilled for parameters in the interior of the parameter space, but not on the boundary, the asymptotic distribution of the element of the 3×3 observed information matrix for the TKw distribution is

$$\sqrt{\mathrm{n}}(\widehat{\omega} - \omega) \sim N_3(0, V^{-1}).$$

where V is the expected information matrix. Thus, the expected information matrix is

$$V^{-1} = -\mathrm{E}\begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial \alpha^2} & \frac{\partial^2 \mathcal{L}}{\partial \alpha\, \partial \theta} & \frac{\partial^2 \mathcal{L}}{\partial \alpha\, \partial \lambda} \\ \frac{\partial^2 \mathcal{L}}{\partial \alpha\, \partial \theta} & \frac{\partial^2 \mathcal{L}}{\partial \theta^2} & \frac{\partial^2 \mathcal{L}}{\partial \theta\, \partial \lambda} \\ \frac{\partial^2 \mathcal{L}}{\partial \alpha\, \partial \lambda} & \frac{\partial^2 \mathcal{L}}{\partial \theta\, \partial \lambda} & \frac{\partial^2 \mathcal{L}}{\partial \lambda^2} \end{bmatrix}. \tag{19}$$

By solving the expected information matrix, these solutions will yield the asymptotic variance and covariances of these ML estimators for $\hat{\alpha}$, $\hat{\theta}$ and $\hat{\lambda}$. By using (19), approximate $100(1-\gamma)\%$ confidence intervals for α, θ and λ can be determined as

$$\hat{\alpha} \pm \mathrm{Z}_{\frac{\gamma}{2}} \sqrt{\hat{\mathrm{V}}_{11}}, \qquad \hat{\theta} \pm \mathrm{Z}_{\frac{\gamma}{2}} \sqrt{\hat{\mathrm{V}}_{22}}, \qquad \hat{\lambda} \pm \mathrm{Z}_{\frac{\gamma}{2}} \sqrt{\hat{\mathrm{V}}_{33}},$$

where $\mathrm{Z}_{\frac{\gamma}{2}}$ is the upper γth percentile of the standard normal distribution.

5. Entropy

The original definition of entropy was defined by Rényi (1961) for a variable X with the probability density function f(x) continuous on [0,1]. Then, the integrated probability is

$$P_{n,k} = \int_{k/n}^{(k+1)/n} f(x)dx, \qquad k = 0,1,\dots,n-1.$$

By defining this as a discrete mass function $\mathrm{P_n} = \mathrm{P_{n,k}}$, it is possible to show that (see Principe (2009))

$$\lim_{n\to\infty} (I_n(P_n) - \log \mathrm{n}) = \frac{1}{1-\rho} \log\left\{\int f(x)^{\rho} dx\right\}.$$

The entropy of a random variable X with density $f(x)$ is a measure of variation of the uncertainty. A large entropy value indicates greater uncertainty in the data. By definition the Rényi entropy is defined as

$$I_R(\rho) = \tfrac{1}{1-\rho} \log\{\textstyle\int f(x)^{\rho} dx\}, \tag{20}$$

where $\rho > 0$ and $\rho \neq 1$. The Rényi entropy for the TKw variable X with density $f(x)$ is given by the following theorem.

Theorem 4: If a random variable X has the Tkw distribution, then the Rényi entropy of X, $I_R(\rho)$ is given by

$$I_R(\rho) = \frac{\rho}{1-\rho}\log(\alpha) + \frac{\rho}{1-\rho}\log(\theta) + \frac{\rho}{1-\rho}\log(1+\lambda) + \frac{1}{1-\rho}\log$$

$$\left[\sum_{k,\ell=0}^{\infty}(\alpha\theta)^{\rho}\left(\frac{2\lambda}{1+\lambda}\right)^{k}(1+\lambda)^{\rho}\mathcal{W}_{k,\ell,\rho}\,\beta\left\{\frac{\rho}{\alpha}(\alpha-1)+1,\rho(\theta-1)+\theta\ell+1\right\}\right].$$

Proof: Rényi entropy is defined in equation (20). So, to complete the proof, we first evaluate the integral as

$$\int_0^1 f(x)^{\rho}dx = \int_0^1 \alpha^{\rho}\theta^{\rho}x^{\rho(\alpha-1)}(1-x^{\alpha})^{\rho(\theta-1)}\left\{1+\lambda-2\lambda\left[1-(1-x^{\alpha})^{\theta}\right]\right\}^{\rho}dx$$

$$= \sum_{k,\ell=0}^{\infty}(\alpha\theta)^{\rho}\left(\frac{2\lambda}{1+\lambda}\right)^{k}(1+\lambda)^{\rho}\mathcal{W}_{k,\ell,\rho}\int_0^1 x^{\rho(\alpha-1)}(1-x^{\alpha})^{\rho(\theta-1)+\theta\ell}dx,$$

where

$$\mathcal{W}_{k,\ell,\rho} = \frac{(-1)^{k+\ell}\Gamma(\rho+1)\Gamma(k+1)}{k!\ell!\,\Gamma(\rho+1-k)\Gamma(k+1-\ell)}.$$

The above integral reduces to

$$\int_0^1 f(x)^{\rho}dx = \sum_{k,\ell=0}^{\infty}(\alpha\theta)^{\rho}\left(\frac{2\lambda}{1+\lambda}\right)^{k}(1+\lambda)^{\rho}\mathcal{W}_{k,\ell,\rho}\,\beta\left\{\frac{\rho}{\alpha}(\alpha-1)+1,\rho(\theta-1)+\theta\ell+1\right\}. \quad (21)$$

By using equation (21), equation (20) can be simplified to

$$I_R(\rho) = \frac{\rho}{1-\rho}\log(\alpha) + \frac{\rho}{1-\rho}\log(\theta) + \frac{\rho}{1-\rho}\log(1+\lambda) + \frac{1}{1-\rho}\log$$

$$\left[\sum_{k,\ell=0}^{\infty}(\alpha\theta)^{\rho}\left(\frac{2\lambda}{1+\lambda}\right)^{k}(1+\lambda)^{\rho}\mathcal{W}_{k,\ell,\rho}\,\beta\left\{\frac{\rho}{\alpha}(\alpha-1)+1,\rho(\theta-1)+\theta\ell+1\right\}\right]$$

which completes the proof.

6. Mean deviation

The mean deviation, about the mean and the median, can be used as measures of the degree of scatter in a population. Let $\mu = E(X)$ and M be the mean and the median of the Tkw distribution given by (11) and (15) equations, respectively.

The mean deviation about the mean, and about the median, can be calculated as

$$\delta_1(X) = \mathrm{E}|X - \mu| = \int_0^1 |X - \mu| f(x)dx,$$

and

$$\delta_2(X) = \mathrm{E}|X - \mathrm{M}| = \int_0^1 |X - \mathrm{M}| f(x)dx,$$

respectively. Hence, we obtain the following equations (Cordeiro et al. (2013))

$$\delta_1 = 2\mu\mathrm{F}(\mu) - 2\psi(\mu) \quad \text{and} \quad \delta_2 = \mu - 2\psi(\mathrm{M}), \tag{22}$$

where $\psi(q)$ can be obtained from (5) by

$$\psi(q) = (1-\lambda)\int_0^{q^\alpha} \theta z^{\frac{1}{\alpha}}(1-z)^{\theta-1}dz + 2\lambda\int_0^{q^\alpha} \theta z^{\frac{1}{\alpha}}(1-z)^{2\theta-1}\,dz. \tag{23}$$

One can easily compute these integrals numerically in software such as MAPLE, MATLAB, Mathcad, R and others and hence obtain the mean deviation about the mean and about the median, as desired. The mean deviation can be also used to determine the Bonferroni and Lorenz curves which have application in econometrics, finance, insurance and others.

By using equation (23), the Bonferroni curve can be calculated from the following equation

$$\mathrm{B(P)} = \frac{1}{\mathrm{P}\mu}\left\{(1-\lambda)\int_0^{q^\alpha} \theta z^{\frac{1}{\alpha}}(1-z)^{\theta-1}dz + 2\lambda\int_0^{q^\alpha} \theta z^{\frac{1}{\alpha}}(1-z)^{2\theta-1}\,dz\right\},$$

and the Lorenz curve can be calculated from the following equation

$$\mathrm{L(P)} = \frac{1}{\mu}\left\{(1-\lambda)\int_0^{q^\alpha} \theta z^{\frac{1}{\alpha}}(1-z)^{\theta-1}dz + 2\lambda\int_0^{q^\alpha} \theta z^{\frac{1}{\alpha}}(1-z)^{2\theta-1}\,dz\right\}.$$

7. Order statistics

If $X_{1:n}, \ldots, X_{n:n}$ denote the order statistics of a random sample $X = (X_1, \ldots, X_n)$ from the TKw distribution with cumulative distribution function $\mathrm{F}(x)$, and probability density function $f(x)$, then the 1st order and nth order probability density functions are given by

$$f_{1:n}(x) = \mathrm{n}\{1 - [1 - (1-x^\alpha)^\theta][1 + \lambda(1-x^\alpha)^\theta]\}^{\mathrm{n}-1}$$

$$\times \alpha\theta x^{\alpha-1}(1-x^{\alpha})^{\theta-1}\{1-\lambda+2\lambda(1-x^{\alpha})^{\theta}\}, \qquad (24)$$

and

$$f_{n:n}(x) = \mathrm{n}\{[1-(1-x^{\alpha})^{\theta}][1+\lambda(1-x^{\alpha})^{\theta}]\}^{\mathrm{n}-1} \times \alpha\theta x^{\alpha-1}(1-x^{\alpha})^{\theta-1}\{1-\lambda+2\lambda(1-x^{\alpha})^{\theta}\}. \qquad (25)$$

Theorem 5: The probability density function and the k^{th} moment of *r*th order statistic $X_{r:n}$ of the random sample X from the TKw distribution are given by

$$f_{r:n}(x) = \mathrm{n}\binom{\mathrm{n}-1}{\mathrm{r}-1}\sum_{\mathrm{k}=0}^{\mathrm{n}-\mathrm{r}}\sum_{\ell=0}^{\infty}\sum_{m=0}^{\infty} w_{k,\ell,m,\lambda}\left\{(1-\lambda)\tau^{(1)}_{\alpha,\theta,m,\lambda}+2\lambda\tau^{(2)}_{\alpha,\theta,m,\lambda}\right\}.$$

$$\mu_{\mathrm{k}}^{(\mathrm{r:n})} = \mathrm{n}\binom{\mathrm{n}-1}{\mathrm{r}-1}\sum_{\mathrm{k}=0}^{\mathrm{n}-\mathrm{r}}\sum_{\ell=0}^{\infty}\sum_{m=0}^{\infty} w_{k,\ell,m,\lambda}\left[(1-\lambda)\theta\beta\left\{\frac{\mathrm{k}}{\alpha}+1,\theta(m+1)\right\} + 2\lambda\theta\beta\left\{\frac{\mathrm{k}}{\alpha}+1,\theta(m+2)\right\}\right].$$

Proof: From Balakrishnan and Nagaraja (1992), the pdf of *r*th order statistics is given by

$$f_{r:n}(x) = \frac{[\mathrm{F}(x)]^{\mathrm{r}-1}[1-\mathrm{F}(x)]^{\mathrm{n}-\mathrm{r}}\mathrm{f}(x)}{\mathrm{B}(\mathrm{r},\mathrm{n}-\mathrm{r}+1)}, \qquad (26)$$

where $\mathrm{B}(.,.)$ is the beta function. Using (26) the pdf of $x_{(r)}$ is given by

$$f_{r:n}(x) = \frac{1}{\mathrm{B}(\mathrm{r},\mathrm{n}-\mathrm{r}+1)}\sum_{\mathrm{k}=0}^{\mathrm{n}-\mathrm{r}}\binom{\mathrm{n}-\mathrm{r}}{\mathrm{k}}(-1)^{\mathrm{k}}\big(\mathrm{F}(x)\big)^{\mathrm{r}+\mathrm{k}-1}\mathrm{f}(x). \qquad (27)$$

Substituting (5) and (6) in (27) we obtain

$$f_{r:n}(x) = \mathrm{n}\binom{\mathrm{n}-1}{\mathrm{r}-1}\sum_{\mathrm{k}=0}^{\mathrm{n}-\mathrm{r}}\binom{\mathrm{n}-\mathrm{r}}{\mathrm{k}}(-1)^{\mathrm{k}}\{[1-(1-x^{\alpha})^{\theta}][1+\lambda(1-x^{\alpha})^{\theta}]\}^{\mathrm{r}+\mathrm{k}-1} \times \alpha\theta x^{\alpha-1}(1-x^{\alpha})^{\theta-1}\{1-\lambda+2\lambda(1-x^{\alpha})^{\theta}\}.$$

The above expression reduces to the pdf of the *r*th order statistic as

$$f_{r:n}(x) = \mathrm{n}\binom{\mathrm{n}-1}{\mathrm{r}-1}\sum_{\mathrm{k}=0}^{\mathrm{n}-\mathrm{r}}\sum_{\ell=0}^{\infty}\sum_{m=0}^{\infty} w_{k,\ell,m,\lambda}\left\{(1-\lambda)\tau^{(1)}_{\alpha,\theta,m,\lambda}+2\lambda\tau^{(2)}_{\alpha,\theta,m,\lambda}\right\}, \qquad (28)$$

where

$$w_{k,\ell,m,\lambda} = \binom{\mathrm{n}-\mathrm{r}}{\mathrm{k}}\binom{\mathrm{r}+\mathrm{k}-1}{\ell}\binom{\mathrm{r}+\mathrm{k}+\ell-1}{m}(-1)^{\mathrm{k}+\ell+m}\left(\frac{\lambda}{1+\lambda}\right)^{\ell}(1+\lambda)^{\mathrm{r}+\mathrm{k}-1}$$

$$\tau^{(\mathrm{g})}_{\alpha,\theta,m,\lambda} = \alpha\theta x^{\alpha-1}(1-x^{\alpha})^{\theta(m+\mathrm{g})-1}, \quad g = 1,2.$$

Using (28) the k^{th} moment of rth order statistic of $x_{(r)}$ is given by

$$\mu_k^{(r:n)} = n\binom{n-1}{r-1}\sum_{k=0}^{n-r}\sum_{\ell=0}^{\infty}\sum_{m=0}^{\infty} w_{k,\ell,m,\lambda}\left[(1-\lambda)\int_0^1 \alpha\theta x^{k+\alpha-1}(1-x^{\alpha})^{\theta(m+1)-1}\,dx + 2\lambda\int_0^1 \alpha\theta x^{k+\alpha-1}(1-x^{\alpha})^{\theta(m+2)-1}\,dx\right].$$

Therefore, by solving the above integral the k^{th} moment of the rth order statistic of the TKw distribution can be obtained as

$$\mu_k^{(r:n)} = n\binom{n-1}{r-1}\sum_{k=0}^{n-r}\sum_{\ell=0}^{\infty}\sum_{m=0}^{\infty} w_{k,\ell,m,\lambda}\left[(1-\lambda)\theta\beta\left\{\frac{k}{\alpha}+1,\theta(m+1)\right\} + 2\lambda\theta\beta\left\{\frac{k}{\alpha}+1,\theta(m+2)\right\}\right], \quad (29)$$

which completes the proof.

Theorem 6: The probability density function and the k^{th} moment of the median order statistic of a random sample X from the TKw distribution are given by

$$g(\tilde{x}) = \frac{(2m+1)!}{m!\,m!}\sum_{k=0}^{m}\sum_{\ell=0}^{\infty}\sum_{n=0}^{\infty} w_{k,\ell,m,n,\lambda}\left\{(1-\lambda)\mathcal{E}_{\alpha,\theta,n,\lambda} + 2\lambda\mathcal{J}_{\alpha,\theta,n,\lambda}\right\}$$

and

$$\tilde{\mu}_k^{(r:n)} = \frac{(2m+1)!}{m!\,m!}\sum_{k=0}^{m}\sum_{\ell=0}^{\infty}\sum_{n=0}^{\infty}\sum_{n=0}^{\infty} w_{k,\ell,m,n,\lambda}\left[(1-\lambda)\theta\beta\left\{\frac{k}{\alpha}+1,\theta(n+1)\right\} + 2\lambda\theta\beta\left\{\frac{k}{\alpha}+1,\theta(n+2)\right\}\right].$$

Proof: Let $X,\ldots,X_n$ be independently and identically distributed ordered random variables from the TKw distribution having median order X_{m+1} probability density function given by

$$g(\tilde{x}) = \frac{(2m+1)!}{m!\,m!}\sum_{k=0}^{m}\binom{m}{k}(-1)^k\{F(\tilde{x})\}^{m+k}f(\tilde{x}). \quad (30)$$

Substituting (5) and (6) in (30) we obtain

$$g(\tilde{x}) = \frac{(2m+1)!}{m!\,m!}\sum_{k=0}^{m}\binom{m}{k}(-1)^k\{[1-(1-\tilde{x}^{\alpha})^{\theta}][1+\lambda(1-\tilde{x}^{\alpha})^{\theta}]\}^{m+k} \times \alpha\theta\tilde{x}^{\alpha-1}(1-\tilde{x}^{\alpha})^{\theta-1}\{1-\lambda+2\lambda(1-\tilde{x}^{\alpha})^{\theta}\}.$$

The above expression reduces to the pdf of the median as

$$g(\tilde{x}) = \frac{(2\mathrm{m}+1)!}{\mathrm{m}!\,\mathrm{m}!}\sum_{\mathrm{k}=0}^{\mathrm{m}}\sum_{\ell=0}^{\infty}\sum_{n=0}^{\infty} w_{k,\ell,m,n,\lambda}\left\{(1-\lambda)\mathcal{J}^{(1)}_{\alpha,\theta,n,\lambda} + 2\lambda\mathcal{J}^{(2)}_{\alpha,\theta,n,\lambda}\right\}, \quad (31)$$

where

$$w_{k,\ell,m,n,\lambda} = \binom{\mathrm{m}}{\mathrm{k}}\binom{\mathrm{m}+\mathrm{k}}{\ell}\binom{\mathrm{m}+\mathrm{k}+\ell}{n}(-1)^{\mathrm{k}+\ell+n}\left(\frac{\lambda}{1+\lambda}\right)^{\ell}(1+\lambda)^{\mathrm{m}+\mathrm{k}}$$

$$\mathcal{J}^{(h)}_{\alpha,\theta,n,\lambda} = \alpha\theta\tilde{x}^{\alpha-1}(1-\tilde{x}^{\alpha})^{\theta(n+h)-1}, \qquad h = 1, 2.$$

Using (31) the k^{th} moment of the *r*th median $\tilde{x}_{(r)}$ given by

$$\tilde{\mu}^{(\mathrm{r:n})}_{\mathrm{k}} = \frac{(2\mathrm{m}+1)!}{\mathrm{m}!\,\mathrm{m}!}\sum_{\mathrm{k}=0}^{\mathrm{m}}\sum_{\ell=0}^{\infty}\sum_{n=0}^{\infty}\sum_{n=0}^{\infty} w_{k,\ell,m,n,\lambda}\left[(1-\lambda)\int_0^1 \alpha\theta\tilde{x}^{k+\alpha-1}(1-\tilde{x}^{\alpha})^{\theta(n+1)-1}\,d\tilde{x}\right.$$

$$\left. +2\lambda\int_0^1 \alpha\theta\tilde{x}^{k+\alpha-1}(1-\tilde{x}^{\alpha})^{\theta(n+2)-1}d\tilde{x}\right].$$

Therefore, the above integral reduces to the k^{th} moment of the median as follows

$$\tilde{\mu}^{(\mathrm{r:n})}_{\mathrm{k}} = \frac{(2\mathrm{m}+1)!}{\mathrm{m}!\,\mathrm{m}!}\sum_{\mathrm{k}=0}^{\mathrm{m}}\sum_{\ell=0}^{\infty}\sum_{n=0}^{\infty}\sum_{n=0}^{\infty} w_{k,\ell,m,n,\lambda}\left[(1-\lambda)\theta\beta\left\{\frac{\mathrm{k}}{\alpha}+1, \theta(n+1)\right\}\right.$$

$$\left. +2\lambda\theta\beta\left\{\frac{\mathrm{k}}{\alpha}+1, \theta(n+2)\right\}\right], \quad (32)$$

which completes the proof.

8. Simulation

This section evaluates the performance of the MLEs for the three parameters α, θ and λ of the TKw distribution by using Monte Carlo simulation. The simulation of the TKw distribution can be performed by using (14). The samples of the TKw distribution were generated for different sizes $\mathrm{n} = 25, 50, 75, 100, 200, 300, 400, 500$ for fixed choice of parameters $\alpha = 3, \theta = 3$ and $\lambda = 0.5$. The estimates of the unknown parameters has been obtained by using BFGS method to minimize the total log-likelihood function. The estimated values of the parameters α, θ, λ with their corresponding standard error, bias and mean square error (MSE) are displayed in Table 1. The plot in Figure 6 evaluates the overall performance of the TKw distribution for simulated data sets that show the exact densities and histogram for some selected values of parameters.

Table 1. Mean, standard Error, Bias and MSE of the TKw distribution

n	Parameter	Mean	S.E	Bias	MSE
25	α	3.1119	0.6526	0.1119	0.4384
	θ	3.2504	1.3394	0.2504	1.8566
	λ	0.2605	0.6013	-0.2395	0.4189
50	α	2.9603	1.0269	-0.0397	1.0561
	θ	3.6296	0.8641	0.6296	1.1430
	λ	-0.2737	0.8336	-0.7737	1.2935
75	α	3.2675	0.3535	0.2675	0.1965
	θ	3.1488	0.9566	0.1488	0.9372
	λ	0.6362	0.3023	0.1362	0.1099
100	α	2.6158	0.4246	-0.3842	0.3278
	θ	3.1376	0.5942	0.1376	0.3720
	λ	0.0164	0.5084	-0.4836	0.4923
200	α	2.9212	0.2045	-0.0788	0.0480
	θ	3.0396	0.8866	0.0396	0.7876
	λ	0.4882	0.4256	-0.0118	0.1812
300	α	3.0556	0.1983	0.0556	0.0424
	θ	3.6425	0.6270	0.6425	0.8059
	λ	0.3285	0.3121	-0.1715	0.1268
400	α	2.8835	0.1679	-0.1165	0.0417
	θ	3.0600	0.4560	0.0600	0.2115
	λ	0.3212	0.2736	-0.1788	0.1068
500	α	2.9709	0.1579	-0.0291	0.0257
	θ	3.2659	0.4845	0.2659	0.3054
	λ	0.3331	0.2789	-0.1669	0.1056

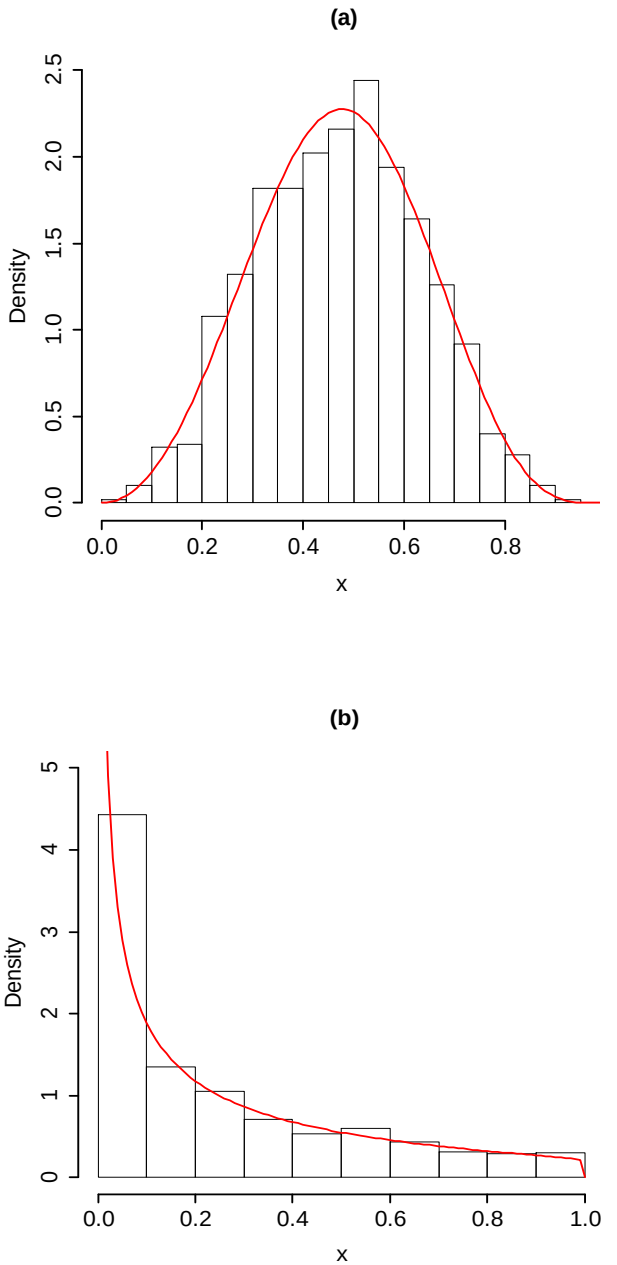


Figure 6. Plots of the TKw densities for simulated data sets: (a) $\alpha = 3, \theta = 3, \lambda = 1$ and (b) $\alpha = 0.5, \theta = 1, \lambda = 0.5$.

9. Applications

In this section we provide two data analyses in order to assess the goodness-of-fit of the TKw distribution. The first data set is from Dumonceaux and Antle, [32]; with respect to the flood data with 20 observations 0.265, 0.269, 0.297, 0.315, 0.3235, 0.338, 0.379, 0.379, 0.392, 0.402, 0.412, 0.416, 0.418, 0.423, 0.449, 0.484, 0.494, 0.613, 0.654, 0.74. The summary statistics for the TKw distribution are given in Table 2. The MLEs and the values of the maximized log-likelihoods for the TKw and Kw distributions are given in Table 3.

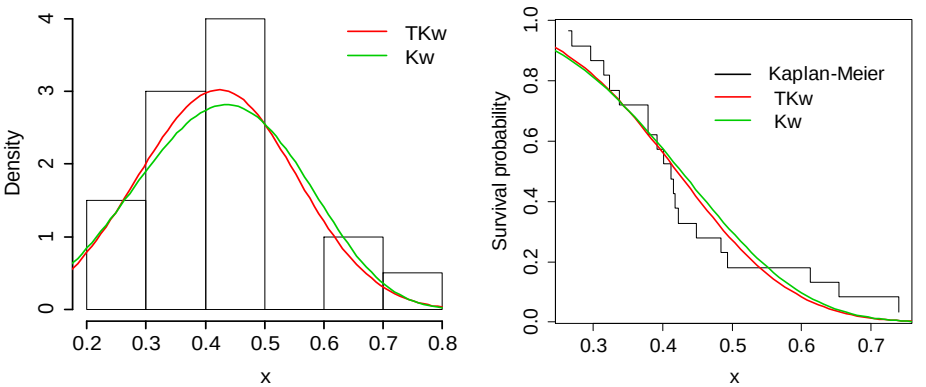
Table 2. Summary Statistics for the TKw and Kw distributions for flood data

Distribution	Median	Coefficient of Quartile Deviation	Bowley's skewness	Moors($\mathcal{M}$) kurtosis
TKw	0.4207	4.7328	-0.0048	1.2212
Kw	0.4268	4.4817	-0.0176	1.2019

In order to compare the distributions, we consider the Kolmogorov-Smirnov (K-S) test, Cramér-von Mises and Anderson-Darling goodness-of-fit statistics for the flood data. Table 3 gives the MLEs of the unknown parameters and the corresponding standard errors of the TKw and Kw distributions. These results indicate that the TKw distribution provides an adequate fit for the flood data.

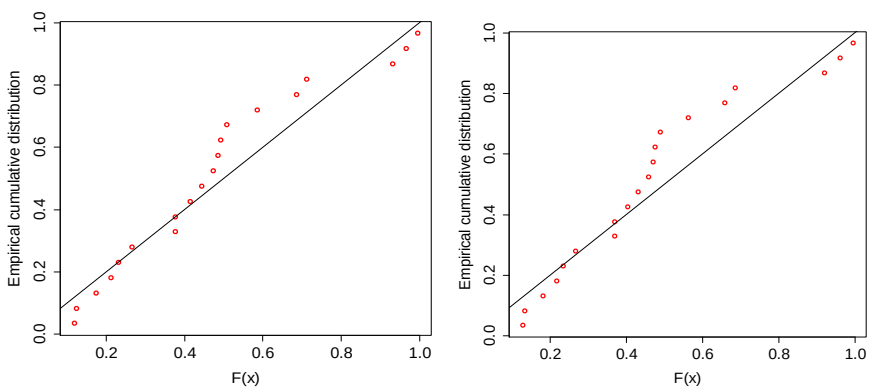
Table 3. MLEs of the unknown Parameters for the flood data with the corresponding standard errors in parenthesis and the goodness-of-fit measures, The K-S test, Cramér-von Mises and Anderson-Darling goodness-of-fit.

Distribution	Parameter Estimates			K-S test	$\mathcal{W}$	$\mathcal{A}$
	$\hat{\alpha}$	$\hat{\theta}$	$\hat{\lambda}$			
TKw	3.7252 (0.6489)	10.9575 (6.0334)	0.6143 (0.3752)	0.1930	0.1409	0.8408
Kw	3.3631 (0.6033)	11.7882 (5.3589)	-	0.2109	0.1658	0.9722



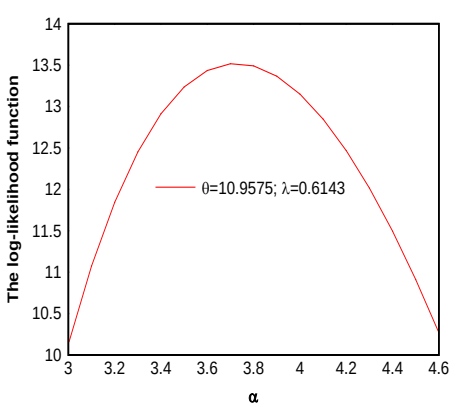
(a) Estimated pdf (b) Estimated Survival function

Figure 7. Estimated densities and survival functions of *TKw* and *Kw* models fitted to flood data

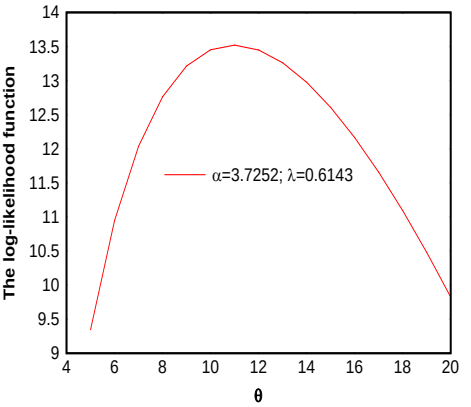


(a) TKw (b) Kw

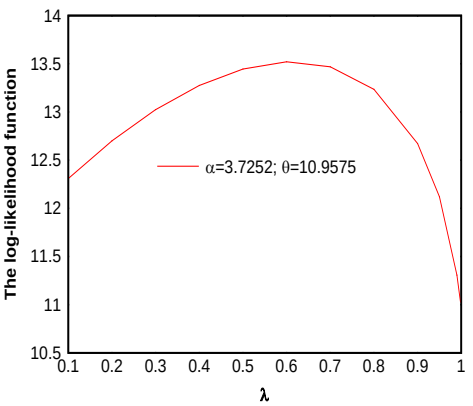
Figure. 8 P-P Plots of the (a) TKw and (b) Kw models fitted to flood data.



(a) $\mathcal{L}$ for α



(b) $\mathcal{L}$ for θ



(c) $\mathcal{L}$ for λ

Figure 9. The profile of the log-likelihood function for α, θ and λ for flood data

Based on the plots of the estimated TKw and Kw densities the relative histogram of the flood data suggests that the fit of the proposed model performs better than the baseline distribution shown in Figure 7(a). Furthermore, the empirical survival function and the fitted survival functions are plotted in Figure 7(b). By comparing the fitted models of these two distribution we have supporting evidence that the TKw distribution provides a good fit for flood data. We have supporting evidence that the Kolmogorov-Smirnov (K-S) distance between the empirical and fitted TKw distribution is smaller than the Kw distribution. Table 3 also indicates that the Cramér-von Mises test statistics and Anderson-Darling goodness statistics have the smallest values for the TKw distribution for the flood data with regard to the Kw distribution. Based on these three goodness-of-fit measures we conclude that the TKw distribution provides a better fit than the Kw lifetime distribution. Figure 8 displays the P-P Plots of the TKw distribution and

Kw distribution. The P-P Plot distance between the ab-line and the fitted model of the TKw distribution and Kw distribution are very similar and close to the baseline model. Figure 9 illustrates the profile of the log-likelihood function for the TKw distribution with parameters α, θ and λ fitted to the flood data and exhibits the unique maximum for these parameters. Based on these results we conclude that the TKw tends to provide a relatively better fit than the Kw distribution for the flood data.

The second data set has been collected from Joint United Nations programme on HIV/ AIDS (UNAIDS), for Infants born to HIV+ women receiving a virological test for HIV within 2 months of birth. The data set consists of 906 observations for the years 2009-13 and is freely available online at http://data.un.org/Data.aspx?d=UNAIDS&f=inID%3a41. The summary statistics for the TKw and Kw distributions for HIV data are given in Table 4.

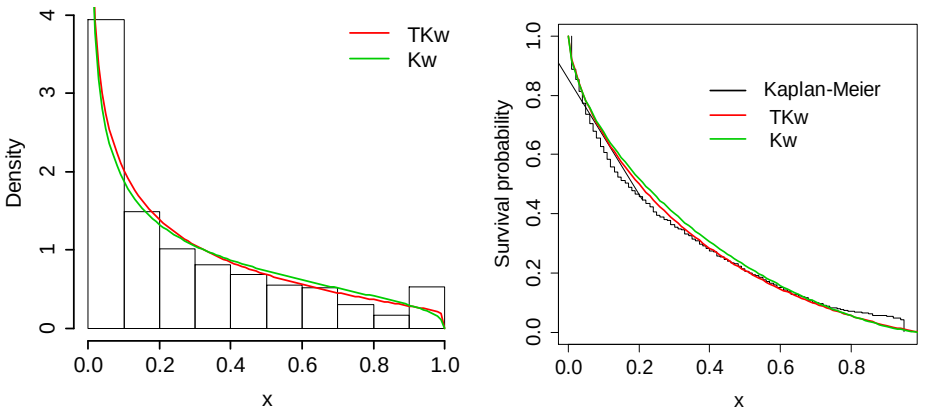
Table 4. Summary Statistics for the TKw and Kw distributions for HIV data

Distribution	Median	Coefficient of Quartile Deviation	Bowley's skewness	Moors($\mathcal{M}$) kurtosis
TKw	0.2002	1.3272	0.2754	1.1491
Kw	0.2146	1.3137	0.2529	1.0884

The MLEs and the values of the maximized log-likelihoods for the TKw and Kw distributions for HIV data are given in Table 5, with the MLEs of the unknown parameters and the corresponding standard errors of the TKw and Kw distributions. The standard error estimates obtained using the observed information matrix appear to be smaller than the parameter estimates. In order to compare the distributions, we consider the AIC (Akaike Information Criterion) for the HIV/ AIDS data. These results indicate that the TKw distribution provides better fit for the HIV/ AIDS data. Furthermore, we applied the LR statistics in order to verify which model fits better for the HIV data. The hypotheses to be tested are $H_0: \omega = (\alpha, \theta, \lambda)^{\mathrm{T}}$ versus H_A: H_0 *is not true*, and the LR statistics reduces to $\Lambda = 2\{l(\widehat{\omega}) - l(\widetilde{\omega})\}$=52.8088, where $\widehat{\omega}$ is the MLE of ω under H_0. The null hypothesis is rejected as the p-value=1.7048E-12 < $\alpha = 0.05$.

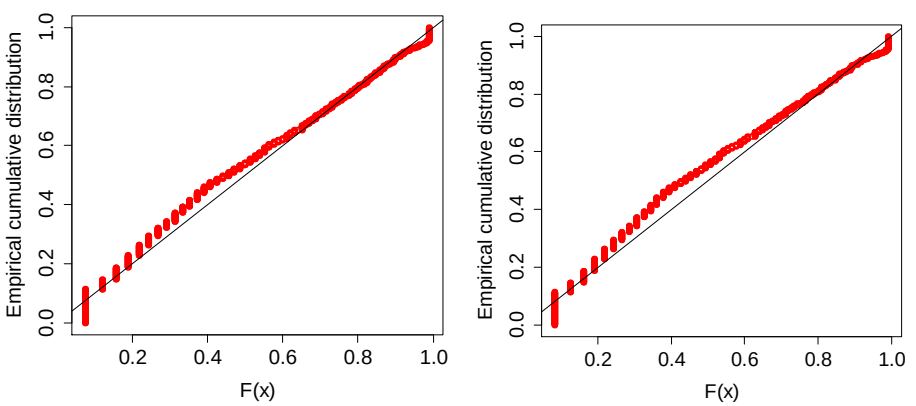
Table 5. MLEs of the unknown Parameters for the HIV/ AIDS data with the corresponding standard errors in parenthesis and the goodness-of-fit measure AIC

Distribution	Parameter Estimates			AIC
	$\hat{\alpha}$	$\hat{\theta}$	$\hat{\lambda}$	
TKw	0.6724	1.1222	0.5491	-633.605
	(0.0251)	(0.0837)	(0.0882)	
Kw	0.6096	1.3961	-	-607.201
	(0.0237)	(0.0629)		



(a) Estimated pdf (b) Estimated Survival function

Figure 10. Estimated densities and survival functions of *TKw* and *Kw* models fitted to HIV data



(a) TKw (b) Kw

Figure 11. P-P Plots of the (a) TKw and (b) Kw models fitted to HIV data

Finally, in order to assess whether the proposed model is appropriate for HIV data we display the visualization of the estimated TKw and Kw densities and the relative histogram for the HIV/ AIDS data suggests that the fit of the proposed model performs better than the baseline distribution shown in Figure 10(a). Furthermore, the plots of the empirical survival function and the fitted survival functions are shown in Figure 10(b). Both figures suggest that the TKw

distribution provides a good fit for HIV/ AIDS data. Moreover, the graphical comparison of the PP-Plots corresponding to these fits confirms our claim as demonstrated in Figure 11.

10. Concluding remarks

In this paper we have presented a new generalization of the Kw distribution, called the TKw distribution. This generalization is obtained by transforming the two parameter Kw model through the quadratic rank transmuted map technique. The properties of the proposed distribution are discussed. We obtain the analytical shapes of the density and hazard functions of the TKw distribution. We also consider mean deviations, Bonferroni and Lorenz curves and Rényi entropy. Maximum likelihood estimation is discussed within the framework of asymptotic log-likelihood inferences including confidence intervals. The three parameter TKw distribution produced monotonically increasing and decreasing hazard rates. In terms of the statistical significance of the model adequacy, the TKw distribution leads to a better fit than the Kw distribution.

Acknowledgements

The authors would like to thank the referees for valuable comments and suggestions which greatly improved the paper.

REFERENCES

ASHOUR, S. K., ELTEHIWY, M. A., (2013). Transmuted Lomax Distribution, American Journal of Applied Mathematics and Statistics, 1(6), pp. 121−127.

AHMAD, A., AHMAD, S. P., AHMED, A., (2015). Characterization and estimation of transmuted Kumaraswamy distribution, Mathematical Theory and Modeling, Vol. 5, No. 9, 168−174.

ARYAL, G. R., TSOKOS, C. P., (2011). Transmuted Weibull distribution: A Generalization of the Weibull Probability Distribution. European Journal of Pure and Applied Mathematics, Vol. 4, No. 2, 89−102.

ARYAL, G. R., TSOKOS. C. P., (2009). On the transmuted extreme value distribution with applications. Nonlinear Analysis: Theory, Methods and applications, Vol. 71, 1401−1407.

ARYAL, G. R., (2013). Transmuted Log-Logistic Distribution, J. Stat. Appl. Pro. 2, No. 1, 11−20.

BALAKRISHNAN, A. N., NAGARAJA, H. N., (1992). A first course in order statistics. New York: Wiley-Interscience.

CORDEIRO, G. M., ORTEGA, E. M. M., NADARAJAH, S., (2010). The Kumaraswamy Weibull distribution with application to failure data. J Frankl Inst 347, 1399–1429.

CORDEIRO, G. M., NADARAJAH, S., ORTEGA, E. M. M., (2012). The Kumaraswamy Gumbel distribution, Stat Methods Appl 21: 139–168.

CORDEIRO G. M., GOMES, A. E., QUEIROZ DA-SILVA, C., ORTEGA, E. M. M., (2013). The beta exponentiated Weibull distribution, Journal of Statistical Computation and Simulation. 83, 1, 114–138.

CORDEIRO, G. M., DE CASTRO M., (2009). A new family of generalized distributions, Journal of Statistical Computation & Simulation, Vol. 00, No. 00, August, 1–17.

DUMONCEAUX, R., ANTLE, C. E., (1973). Discrimination between the log-normal and the Weibull distributions. Technometrics 15 (4), 923–926.

ELBATHAL, I., ELGARHY, M., (2013). Transmuted Quasi Lindley Distributions: A generalization of the Quasi Lindley Distribution. International Journal of Pure and Applied Sciences and Technology , 18(2), pp. 59–69.

FLETCHER, S. C., PONNAMBALAM, K., (1996). Estimation of reservoir yield and storage distribution using moments analysis. J Hydrol 182, 259–275.

GANJI, A., PONNAMBALAM, K., KHALILI, D., (2006). Grain yield reliability analysis with crop water demand uncertainty. Stoch Environ Res Risk Assess 20, 259–277.

JONES, M. C., (2009). Kumaraswamy's distribution: a beta-type distribution with some tractability advantages. Stat Methodol 6, 70–91.

KENNEY, J. F., KEEPING, E. S., (1962). Mathematics of Statistics. Princeton, NJ.

KUMARASWAMY, P., (1980). Generalized probability density-function for double-bounded random-processes. J Hydrol 46, 79–88.

KHAN, M. S., KING, R., (2013a). Transmuted Modified Weibull Distribution: A Generalization of the Modified Weibull Probability Distribution, European Journal of Pure And Applied Mathematics, Vol. 6, No. 1, 66–88.

KHAN, M. S., KING, R., (2013b). Transmuted generalized Inverse Weibull distribution, Journal of Applied Statistical Sciences, Vol. 20, No. 3, 15–32.

KHAN, M. S., KING, R., HUDSON, I., (2013c). Transmuted generalized Exponential distribution, 57th Annual Meeting of the Australian Mathematical Society, Australia.

KHAN, M., SHUAIB, KING, R., HUDSON, I., (2014). Characterizations of the transmuted Inverse Weibull distribution, ANZIAM J., Vol. 55 (EMAC2013 at the Queensland University of Technology from 1st – 4th December 2013), C197–C217.

KOUTSOYIANNIS, D, XANTHOPOULOS, T., (1989). On the parametric approach to unit hydrograph identification. Water Resour Manag 3, 107–128.

MOORS, J. A., (1998). A quantile alternative for kurtosis. Journal of the Royal Statistical Society, D, 37, 25–32.

MEROVCI, F., (2013a). Transmuted Rayleigh distribution. Austrian Journal of Statistics, Vol. 42, No. 1, 21–31.

MEROVCI, F., (2013b). Trasmuted Lindley distribution. Int. J. Open Problems Compt. Math, 6, 63–72.

MEROVCI, F., (2014). Transmuted generalized Rayleigh distribution, J. Stat. Appl. Pro. 3, No. 1, 9–20.

NADARAJAH, S., (2008). On the distribution of Kumaraswamy. J Hydrol 348, 568–569.

PRINCIPE, J. C., (2009). http://www.cnel.ufl.edu/courses/EEL6814/renyis_entropy.pdf.

PONNAMBALAM, K, SEIFI, A, VLACH, J., (2001). Probabilistic design of systems with general distributions of parameters. Int J Circuit Theory Appl 29, 527–536.

RÉNYI, A., (1961). On measures of entropy and information. University of California Press, Berkeley, California, 547–561.

SUNDAR, V., SUBBIAH, K., (1989). Application of double bounded probability density-function for analysis of ocean waves. Ocean Eng 16, 193–200.

SEIFI, A., PONNAMBALAM, K., VLACH, J., (2000). Maximization of manufacturing yield of systems with arbitrary distributions of component values. Ann Oper Res 99, 373–383.

SHAW, W. T., BUCKLEY, I. R. C., (2009). The alchemy of probability distributions: beyond Gram–Charlier expansions, and a skew-kurtotic normal distribution from a rank transmutation map. Technical report, http://arxiv.org/abs/0901.0434.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 211–219

BAYESIAN INFERENCE FOR STATE SPACE MODEL WITH PANEL DATA

Ranjita Pandey[1], Anoop Chaturvedi[2]

ABSTRACT

The present work explores panel data set-up in a Bayesian state space model. The conditional posterior densities of parameters are utilized to determine the marginal posterior densities using the Gibbs sampler. An efficient one step ahead predictive density mechanism is developed to further the state of art in prediction-based decision making.

Key words: Bayesian analysis, Gibbs sampler, conditional posterior densities, predictive distribution.

1. Introduction

The importance of panel data stems from its ability for addressing questions of economic and social behaviour which cannot be easily answered by using the usual cross section or time series data. The panel data consists of the observations on the same cross section of units under study at different and usually successive time periods. The longitudinal nature of panel data allows for the use of simple techniques to solve otherwise complicated problems and permits cross section and/or time heterogeneity. Tiwari, Yang and Zalkikar (1996) have studied the level of water pollution by recording biochemical oxygen demand (BOD) and dissolved oxygen (DO), at a selected point along the stream at different time points. The present paper adapts and extends their concept to a multiple point scenario. The situation could be visualized in a comprehensive manner, if the BOD and DO measurements are taken at more than one selected point along the length of the stream, at successive time periods. The resulting longitudinal data accounts for the individual effects at the various locations as well as considers impact of other relatively slowly changing left-out variables. Panel data-based studies have been undertaken by Maddala (1971), Mundlak (1978), Hausman (1978), Hausman and Taylor (1981) and Chamberlain (1982) in various fields. Baltagi (2008) advocates the use of panel data for controlling individual

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

heterogeneity and to incorporate dynamic adjustment in the model while elaborating that the panel data structure gives more informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency compared to the classical cross sectional model-based study.

2. Model with panel data and assumptions

The variables involved in the model are denoted by

y_{it} : observed value of the dependent variable for unit i at time point *t.*

x_{it} : $p \times 1$ vector of observations on p explanatory variable corresponding to the i-th unit at time t, $(i = 1,\ldots,n;\ t = 1,\ldots,T)$

The state space model for panel data is given by

$$y_{it} = x_{it}'\theta_t + \varepsilon_{it} \tag{2.1}$$

The dynamics of the system as it evolves through time is represented by

$$\theta_t = G_t\, \theta_{t-1} + v_t \tag{2.2}$$

G_t is a known $p \times p$ transition matrix, θ_t is a $p \times 1$ unknown parameter vector, θ_0 denotes the initial state of the system, v_t is the systems error, ε_{it} is the observation error. We write ε_{it} as

$$\varepsilon_{it} = \alpha_i + \eta_{it} \tag{2.3}$$

We also assume that α_i, η_{it} and v_t are all independently distributed. Hence, equation (2.1) may be rewritten as

$$y_{it} = x_{it}'\theta_t + \alpha_i + \eta_{it} \tag{2.4}$$

We further assume that

$$\alpha_i \sim N\left(0, \frac{\sigma_\alpha^2}{\lambda}\right) \quad \text{for all } i \tag{2.5}$$

$$\eta_{it} \sim N\left(0, \frac{\sigma_\eta^2}{\lambda}\right) \quad \text{for all } i \text{ and } t \tag{2.6}$$

$$v_t | \lambda \sim N(0, \lambda^{-1}\Sigma) \quad \text{for all } t \tag{2.7}$$

$$\lambda \sim G\left(\frac{a_0}{2}, \frac{b_0}{2}\right) \tag{2.8}$$

$$\theta_0 | \lambda \sim N(m_0, \lambda^{-1}\Sigma_0) \tag{2.9}$$

Thus, $\theta_t | \theta_{t-1}, \lambda \sim N(G_t\theta_{t-1}, \lambda^{-1}\Sigma)$ (2.10)

where Σ is a known $p\times p$ positive definite matrix. We can write the model (2.1) as

$$y_t = X_t\,\theta_t + \varepsilon_t \tag{2.11}$$

and

$$\varepsilon_t = \alpha + \eta_t \tag{2.12}$$

where $y_t = (y_{1t},\ldots,y_{nt})' : n\times 1,\ X_t = \begin{bmatrix} x_{1t}' \\ \mathrm{M} \\ x_{nt}' \end{bmatrix} : n\times p$

$$\alpha = (\alpha_1,\ldots,\alpha_n)' : n\times 1,\ \eta_t = (\eta_{1t},\ldots,\eta_{nt})' : n\times 1,\ \varepsilon_t = (\varepsilon_{1t},\ldots,\varepsilon_{nt})' : n\times 1$$

The distributions of α and η_t are given by

$$\alpha \sim N_n\left(0, \frac{\sigma_\alpha^2}{\lambda} I_n\right) \tag{2.13}$$

$$\eta_t \sim N_n\left(0, \frac{\sigma_\eta^2}{\lambda} I_n\right) \tag{2.14}$$

3. Conditional Posterior densities

In the posterior analysis of the model we treat α as an unknown parameter and derive its conditional posterior density also along with the conditional posterior densities of other parameters, and utilize these conditional posterior densities in employing Gibbs sampler. Under model specifications and the underlying assumptions, we have

$$f^*(y_t|\{\theta_t\}_{t=0}^T,\alpha,\lambda) = \frac{\lambda^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}\sigma_\eta^n}\exp\left[-\frac{\lambda}{2\sigma_\eta^2}(y_t - X_t\theta_t - \alpha)'(y_t - X_t\theta_t - \alpha)\right] \tag{3.1}$$

so that $f^*(\underset{\sim}{y}|\{\theta\}_{t=0}^T,\alpha,\lambda) = \frac{\lambda^{\frac{nT}{2}}}{(2\pi)^{\frac{nT}{2}}\sigma_\eta^{nT}}\exp\left[-\frac{\lambda}{2\sigma_\eta^2}\sum_{t=1}^T(y_t - X_t\theta_t - \alpha)'(y_t - X_t\theta_t - \alpha)\right]$

(3.2)

The joint density of $(\underset{\sim}{y},\alpha,\lambda,\theta_0,\{\theta_t\}_{t=1}^T)$, obtained by combining expressions (2.8), (2.9), (2.10), (2.13) and (3.2), is given as follows

$$f(\underset{\sim}{y},\alpha,\lambda,\theta_0,\{\theta_t\}_{t=1}^{T}) = \frac{\lambda^{\frac{nT}{2}}}{(2\pi)^{\frac{nT}{2}}\sigma_\eta^{nT}} \exp\left[-\frac{\lambda}{2\sigma_\eta^2}\sum_{t=1}^{T}(y_t - X_t\theta_t - \alpha)'(y_t - X_t\theta_t - \alpha)\right] \times$$

$$\frac{\lambda^{\frac{pT}{2}}}{(2\pi)^{\frac{pT}{2}}\sum^{\frac{T}{2}}} \exp\left[-\frac{\lambda}{2}\sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})'\Sigma^{-1}(\theta_t - G_t\theta_{t-1})\right] \times$$

$$\frac{\lambda^{\frac{p}{2}}}{(2\pi)^{\frac{p}{2}}\sum^{\frac{1}{2}}} \exp\left[-\frac{\lambda}{2}(\theta_0 - m_0)'\Sigma_0^{-1}(\theta_0 - m_0)\right] \times$$

$$\left(\frac{b_0}{2}\right)^{\frac{a_0}{2}} \times \frac{1}{\Gamma\left(\frac{a_0}{2}\right)} \times \lambda^{\frac{a_0}{2}-1} \exp\left[-\frac{b_0}{2}\lambda \times\right] \frac{\lambda^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}\sigma_\alpha^n} \exp\left[-\frac{\lambda}{2}\left(\frac{\alpha'\alpha}{\sigma_\alpha^2}\right)\right] \tag{3.3}$$

We define

$$B_t^{-1} = \begin{cases} \Sigma_0^{-1} + G_1'\Sigma^{-1}G_1 & , \text{ for } t = 0 \\ \Sigma^{-1} + \frac{1}{\sigma_\eta^2}X_t'X_t + G_{t+1}'\Sigma^{-1}G_{t+1} & , \text{ for } t = 1,2,\ldots,T-1 \\ \Sigma^{-1} + \frac{1}{\sigma_\eta^2}X_T'X_T & , \text{ for } t = T \end{cases} \tag{3.4}$$

$$b_t = \begin{cases} G_1'\Sigma^{-1}\theta_1 + \Sigma_0^{-1}m_0 & , \text{ for } t = 0 \\ \Sigma^{-1}G_t\theta_{t-1} + G_{t+1}'\Sigma^{-1}\theta_{t+1} + \frac{1}{\sigma_\eta^2}(X_t'y_t - X_t'\alpha) & , \text{ for } t = 1,2,\ldots,T-1 \\ \Sigma^{-1}G_T\theta_{T-1} + \frac{1}{\sigma_\eta^2}(X_T'y_T - X_T'\alpha) & , \text{ for } t = T \end{cases} \tag{3.5}$$

$$\theta^{(s)} = \{\theta_t\}_{t=0}^{T} \quad \text{for } t \neq s \tag{3.6}$$

Theorem 1: The conditional posterior density of θ_S given ($\theta^{(s)}$, α, λ) is normal with mean vector $B_S\, b_S$ and covariance matrix $\lambda^{-1} B_S$.

Proof: . Utilizing (3.3), the conditional posterior density of θ_0 is obtained as

$$f^*(\theta_0|\underset{\sim}{y},\alpha,\lambda) \propto \exp\left[-\frac{\lambda}{2}\left\{\theta_0'\left(\Sigma_0^{-1}+G_1'\Sigma^{-1}G_1\right)\theta_0 - 2\theta_0'\left(\Sigma_0^{-1}m_0+G_1'\Sigma^{-1}\theta_1\right)\right\}\right]$$

$$\propto \exp\left[-\frac{\lambda}{2}\left(\theta_0 - b_0B_0\right)' B_0^{-1}\left(\theta_0 - b_0B_0\right)\right] \tag{3.7}$$

We now write expression (3.7) as

$$f^*(\theta_0|\underset{\sim}{y},\alpha,\lambda) = C_0 \exp\left[-\frac{\lambda}{2}\left(\theta_0 - B_0b_0)'B_0^{-1}(\theta_0 - B_0b_0\right)\right]$$

where C_0 is the normalizing constant. Its value is obtained as

$$C_0^{-1} = \int_{R^p} \exp\left[-\frac{\lambda}{2}(\theta_0 - b_0B_0)'B_0^{-1}(\theta_0 - b_0B_0)\right]d\theta_0 = \frac{(2\pi)^{\frac{p}{2}}|B_0|^{\frac{1}{2}}}{\lambda^{\frac{p}{2}}}$$

For $s = 1,\ldots, T-1$, We have

$$f^*(\theta_s|\underset{\sim}{y},\alpha,\lambda,\theta^{(s)}) \propto \exp\left[-\frac{\lambda}{2}\left\{\theta_s'\left(\Sigma^{-1}+\frac{1}{\sigma_\eta^2}X_s'X_s + G_{s+1}'\Sigma^{-1}G_{s+1}\right)\theta_s\right.\right.$$

$$\left.\left. -2\theta_s'\left(\frac{1}{\sigma_\eta^2}(X_s'y_s - X_s'\alpha) + G_{s+1}'\Sigma^{-1}\theta_{s+1} + \Sigma^{-1}G_s\theta_{s-1}\right)\right\}\right]$$

$$\propto \exp\left[-\frac{\lambda}{2}(\theta_s - B_sb_s)'B_s^{-1}(\theta_s - B_sb_s)\right] \tag{3.8}$$

We now write expression (3.8) as

$$f^*(\theta_s|\underset{\sim}{y},\alpha,\lambda,\theta^{(s)}) = C_s \exp\left[-\frac{\lambda}{2}(\theta_s - B_sb_s)'B_s^{-1}(\theta_s - B_sb_s)\right]$$

where C_S is evaluated as

$$C_s^{-1} = \int_{R^p} \exp\left[-\frac{\lambda}{2}(\theta_s - B_sb_s)'B_s^{-1}(\theta_s - B_sb_s)\right]d\theta_s = \frac{(2\pi)^{\frac{p}{2}}|B_s|^{\frac{1}{2}}}{\lambda^{\frac{p}{2}}}$$

Finally, for $s= T$ we have $f^*\left(\theta_T|\underset{\sim}{y},\alpha,\lambda,\{\theta_t\}_{t=0}^{T-1}\right)$

$$\propto \exp\left[-\frac{\lambda}{2}\left\{\theta_T'\left(\Sigma^{-1}+\frac{1}{\sigma_\eta^2}X_T'X_T\right)\theta_T - 2\theta_T'\left(\frac{1}{\sigma_\eta^2}\left(X_T'y_T - X_T'\alpha\right)+\Sigma^{-1}G_T\theta_{T-1}\right)\right\}\right]$$

$$\propto \exp\left[-\frac{\lambda}{2}(\theta_T - B_T b_T)' B_T^{-1}(\theta_T - B_T b_T) - b_T' B_T b_T\right]$$

$$\propto \exp\left[-\frac{\lambda}{2}(\theta_T - B_T b_T)' B_T^{-1}(\theta_T - B_T b_T)\right] \tag{3.9}$$

We now write equation (3.9) as

$$f^*\left(\theta_T | \underset{\sim}{y}, \alpha, \lambda, \{\theta_t\}_{t=0}^{T-1}\right) = C_T \exp\left[-\frac{\lambda}{2}(\theta_T - B_T b_T)' B_T^{-1}(\theta_T - B_T b_T)\right]$$

where C_T is the normalizing constant, which is obtained as

$$C_T^{-1} = \int_{R^p} \exp\left[-\frac{\lambda}{2}(\theta_T - B_T b_T)' B_T^{-1}(\theta_T - B_T b_T)\right] d\theta_T = \frac{(2\pi)^{\frac{p}{2}} |B_T|^{\frac{1}{2}}}{\lambda^{\frac{p}{2}}}$$

Thus, the theorem is proved.

Let us write

$$Q = \frac{T}{\sigma_\eta^2} + \frac{1}{\sigma_\alpha^2},\ q = \frac{1}{\sigma_\eta^2}[y_t - X_t \theta_t] \tag{3.10}$$

and $a^* = nT + pT + p + a_0 + n$

$$b^*(\alpha) = \frac{1}{\sigma_\eta^2}\sum_{t=1}^{T}(y_t - X_t\theta_t - \alpha)'(y_t - X_t\theta_t - \alpha)' + \sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})'\Sigma^{-1}(\theta_t - G_t\theta_{t-1})'$$

$$+(\theta_0 - m_0)'\sum_0^{-1}(\theta_0 - m_0) + b_0 + \frac{\alpha'\alpha}{\sigma_\alpha^2} \tag{3.11}$$

Theorem 2: The conditional posterior of α is given by normal distribution with mean $Q^{-1}q$ and variance covariance matrix $\lambda^{-1}Q^{-1}$.

Proof : From equation (3.3) we obtain

$$f^*\left(\alpha | \underset{\sim}{y}, \lambda, \theta_0, \{\theta_T\}_{t=1}^{T}\right) \propto \exp\left[-\frac{\lambda}{2}\left\{\alpha'\left(\frac{T}{\sigma_\eta^2} + \frac{1}{\sigma_\alpha^2}\right)\alpha - \frac{2\alpha'}{\sigma_\eta^2}(y_t - X_t\theta_t)\right\}\right]$$

$$\propto \exp\left[-\frac{\lambda}{2}\left\{(\alpha - Q^{-1}q)' Q (\alpha - Q^{-1}q)\right\}\right] \tag{3.12}$$

We rewrite expression (3.12) as follows

$$f^*\left(\alpha|\underset{\sim}{y},\lambda,\theta_0,\{\theta_T\}_{t=1}^T\right)= K_\alpha \exp\left[-\frac{\lambda}{2}\left\{(\alpha-Q^{-1}q)'Q(\alpha-Q^{-1}q)\right\}\right]$$

where the normalizing constant K_α is obtained as

$$K_\alpha^{-1}=\int_{R^n}\exp\left[-\frac{\lambda}{2}\left\{(\alpha-Q^{-1}q)'Q(\alpha-Q^{-1}q)\right\}\right]d\alpha \;=\frac{(2\pi)^{\frac{n}{2}}}{\lambda^{\left[\frac{T}{\sigma_\eta^2}+\frac{1}{\sigma_\alpha^2}\right]}}$$

which establishes the theorem.

Theorem 3: The conditional posterior of λ is given by

$$f^*(\lambda|\underset{\sim}{y},\alpha,\{\theta_t\}_{t=0}^T)=K_\lambda\lambda^{\frac{a^*}{2}-1}\exp\left[-\frac{\lambda}{2}b^*\right] \tag{3.13}$$

Proof: We obtain the following from expression (5.3.3)

$$f^*(\lambda|\underset{\sim}{y},\alpha,\{\theta_t\}_{t=0}^T)\propto\lambda^{\frac{1}{2}(nT+pT+p+a_0+n)-1}\exp\left[-\frac{\lambda}{2}\left\{\frac{1}{\sigma_\eta^2}\sum_{t=1}^T(y_t-x_t\theta_t-\alpha)'(y_t-x_t\theta_t-\alpha)\right.\right.$$

$$\left.\left.+\sum_{t=1}^T\left\{(\theta_t-G_t\theta_{t-1})'\Sigma^{-1}(\theta_t-G_t\theta_{t-1})\right\}+(\theta_0-m_0)'\Sigma_0^{-1}(\theta_0-m_0)+b_0+\frac{\alpha'\alpha}{\sigma_\alpha^2}\right\}\right]$$

$$\propto\lambda^{\frac{a^*}{2}-1}e^{-\frac{b^*}{2}\lambda}$$

or
$$f^*(\lambda|\underset{\sim}{y},\alpha,\{\theta_t\}_{t=0}^T)=\frac{\left(\frac{b^*}{2}\right)^{\frac{a^*}{2}}}{\Gamma\left(\frac{a^*}{2}\right)}\lambda^{\frac{a^*}{2}-1}e^{\frac{b^*}{2}\lambda}$$

Thus, the theorem is proved.

4. Implementation of Gibbs Sampler

Let the generated Gibbs sample be denoted by $\left(\left\{\theta_{tj}^{(k)}\right\}_{t=0}^T,\lambda_j^{(k)},\alpha_j^{(k)}\right)$; *j = 1, 2,...,N ,* where N is the total number of replications. Then, by employing the Gibbs sampler, posterior density of θ_S given $\underset{\sim}{y}$ can be estimated as

$$\hat{f}(\theta_s|\underset{\sim}{y})=\frac{1}{N}\sum_{j=1}^N f^*(\theta_s|\theta_{s-1,j}^{(k)},\theta_{s+1,j}^{(k)},\lambda_j^{(k)},\underset{\sim}{y},\alpha_j^{(k)}) \tag{4.1}$$

for $s = 0, 1, \ldots\ldots, T$. Notice that the estimated posterior density of θ_S in (4.1) depends on the two values adjacent to θ_s. Hence, an estimate of $\theta_{s/T}$ is the mean of the estimated density (4.1) which is obtained as

$$\overline{\theta}_{s/T} = B_s\left(\frac{1}{N}\sum_{j=1}^{N} b_{sj}^{(k)}\right) \tag{4.2}$$

where k is the number of iterations during implementation of the Gibbs sampler. $b_{sj}^{(k)}$ is the value of b_S based on $\left(\left\{\theta_{tj}^{(k)}\right\}_{j=1}^{N},\ t = s-1,\ s+1\right)$. Then the fitted value of y_t for our model is

$$\tilde{y}_t = X_t \overline{\theta}_{t|T} \tag{4.3}$$

Similarly, by employing the Gibbs sampler posterior density of α given $\underset{\sim}{y}$ can be estimated as

$$\hat{f}(\alpha|\underset{\sim}{y}) = \frac{1}{N}\sum_{j=1}^{N} f^*\left(\alpha|\left\{\theta_{tj}^{(k)}\right\}_{j=1}^{N},\ \lambda_j^{(k)},\ \underset{\sim}{y}\right) \tag{4.4}$$

Further, an estimator of α is $\overline{\alpha} = \frac{1}{N}\sum_{j=1}^{N} \overline{\alpha}_j^{(k)}$

where $\overline{\alpha}_j^{(k)} = \left(Q_j^{(k)}\right)^{-1} q_j^{(k)},\ Q_j^{(k)} = \left(\frac{T}{\sigma_\eta^2} + \frac{1}{\sigma_\alpha^2}\right),\ q_j^{(k)} = \left(y_t - X_{tj}^{(k)}\theta_{tj}^{(k)}\right)$

5. Conclusion

The state space model is utilized to obtain Bayesian estimators for the parameters which can improve panel data-based prediction wherein the observations are available on the behaviour of a 'panel' of decision units at multiple successive time epochs. The use of panel data has become increasingly popular in econometrics in recent years. This analysis provides an elaborate theoretical framework and is therefore expected to contribute effectively to improved and more precise panel data-based prediction for applied researchers and practitioners.

Acknowledgements

Authors are thankful to the reviewers for their valuable comments and useful suggestions which have improved the quality of the original manuscript. The first author acknowledges with thanks financial assistance from DU-DST PURSE Grant and R& D Grant from the University of Delhi.

REFERENCES

BALTAGI, B. H., (2008). Econometric analysis of panel data. Wiley.

CHAMBERLAIN, G., (1982). Multivariate regression models for panel data, Journal of Econometrics 18, No. 1.

HAUSMAN, J., (1978). Specification tests in econometrics. Econometrica 46, No. 6.

HAUSMAN, J., TAYLOR, W., (1981). Panel data and unobservable individual effects. Econometrica 49, No. 6.

MADDALA, G. S., (1971). The use of variance components models in pooling cross-section and time series data. Econometrica 39, No. 2.

MUNDLAK, Y., (1978). On the pooling of time series and cross section data. Econometrica 46, No. 1.

TIWARI, R. C., YANG, Y., ZALKIKAR, J. N., (1996). Time series analysis of BOD data using the Gibbs sampler. Enviormetrics 7: 567−78.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 221–236

ON MEASURING INCOME POLARIZATION: AN APPROACH BASED ON REGRESSION TREES

Mauro Mussini[1]

ABSTRACT

This article proposes the application of regression trees for analysing income polarization. Using an approach to polarization based on the analysis of variance, we show that regression trees can uncover groups of homogeneous income receivers in a data-driven way. The regression tree can deal with nonlinear relationships between income and the characteristics of income receivers, and it can detect which characteristics and their interactions actually play a role in explaining income polarization. For these features, the regression tree is a flexible statistical tool to explore whether income receivers concentrate around local poles. An application to Italian individual income data shows an interesting partition of income receivers.

Key words: polarization, regression trees, recursive partitioning, ANOVA, JEL D31, D63, C14.

1. Introduction

The measurement of income polarization has developed by following two distinct approaches. One approach focuses on the concept of bipolarization that considers the extent to which incomes spread from the middle to the tails of the distribution, implying the disappearance of the middle class (Wang and Tsui, 2000; Wolfson, 1994). The other approach relies on the concept of identification-alienation: individuals identify themselves with those having similar income levels, whereas they feel alienated from individuals with different income levels (Deutsch *et al*. 2013; Duclos *et al*., 2004; Esteban and Ray, 1994; Poggi and Silber, 2010); therefore, polarization is investigated from the perspective of grouping of individuals around local poles and within-group identification. Following the second approach, we show that the regression tree is a useful statistical tool for measuring polarization in income distribution.

[1] Department of Economics, University of Verona, Via dell'Artigliere 8, Verona (Italy). E-mail: mauro.mussini@univr.it.

Recently, Palacios-González and García-Fernández (2012) have pointed out that the coefficient of determination (R^2) of an ANOVA linear model can be interpreted as a measure of polarization. Since R^2 increases as within-group variance decreases (i.e. groups are internally more homogeneous), Palacios-González and García-Fernández state that R^2 can be seen as a (normalised) measure of polarization. Moreover, linking the ANOVA coefficient of determination with polarization enables one to analyse polarization by the characteristics of income receivers when groups are defined by such characteristics (Palacios-González and García-Fernández, 2012).

The variance decomposition approach proposed by Palacios-González and García-Fernández is analogous in the spirit to the Zhang and Kanbur (2001) approach to polarization measurement, since the latter is based on the income inequality decomposition by groups. Both the Palacios-González and García-Fernández approach and the Zhang and Kanbur one assume that groups are pre-established, and then measure polarization for that population partition; therefore, both approaches tell us whether polarization is high or low for the population partition defined *a priori*. Duclos *et al*. (2004) suggested letting the population partition arise in a data-driven way rather than taking the population partition as exogenous. In our approach to polarization analysis, we initially face the issue of identifying the most homogeneous groups in a data-driven way and then we measure the degree of income polarization for the population partition showing maximal within-group identification.

We show that groups can be naturally formed from the data exploration by using regression trees to recursively partition the population. We assume that income is the response variable and income receiver's characteristics are the explanatory variables; then, the population is recursively partitioned to maximally reduce the within-group variance, which is maximizing the gain in homogeneity within groups. Once groups clustering income receivers with similar income levels have been detected, R^2 is used to measure the extent to which incomes are polarized.

In our empirical analysis, regression trees are applied to Italian individual income data in order to detect the characteristics relevant for polarization. Our findings show that the interactions among employment status, education and age form well-identified groups of income receivers.

The article is organised as follows. Section 2 briefly reviews the Palacios-González and García-Fernández approach to polarization measurement. Section 3 introduces regression trees and shows how this technique is suitable for analysing income polarization. In Section 4, the regression tree approach is applied to Italian income data from the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy in 2010 (Banca d'Italia, 2012).

2. Measuring polarization VIA ANOVA

The link between polarization and ANOVA is outlined by Palacios-González and García-Fernández (2012) in the generalized linear model framework. Palacios-González and García-Fernández follow the Zhang and Kanbur (2001)

approach to polarization, which assumes that for k predetermined groups of income receivers the larger the ratio of between-group income inequality to within-group income inequality, the larger the polarization. Similarly to the Zhang and Kanbur approach, Palacios-González and García-Fernández assume the mean income of a group as the representative income for the income receivers within that group; moreover, they observe that the larger the disparities among the mean income of a group and the mean incomes of the other groups, the more the income receivers belonging to that group feel alienated from income receivers included in the other groups. However, the Palacios-González and García-Fernández approach differs from the Zhang and Kanbur one since the former is based on variance decomposition by group. Indeed, Palacios-González and García-Fernández propose to measure polarization using the ratio between the variance between groups and the variance within groups

$$P = \frac{\sigma_b^2}{\sigma_w^2}, \tag{1}$$

where σ_w^2 denotes the within-group variance and σ_b^2 is the between-group variance. Then, Palacios-González and García-Fernández suggest to normalise expression in (1) by replacing σ_w^2 with the overall variance $\sigma^2 = \sigma_w^2 + \sigma_b^2$,

$$P^* = \frac{\sigma_b^2}{\sigma^2} = 1 - \frac{\sigma_w^2}{\sigma^2}. \tag{2}$$

The polarization measure in (2) is equivalent to the (unadjusted) R^2 used in ANOVA when one investigates the effect of grouping on income. Palacios-González and García-Fernández formulate a fixed-effects ANOVA model in the framework of generalized linear models, where n income receivers are partitioned into k groups on the basis of the k different values (levels) taken by one of the characteristics of income receivers (e.g. gender, age, employment status, etc.). Let Y_i denote the income receiver i's income and D_{ih} be the dummy variable that equals 1 if the income receiver i belongs to group h and 0 otherwise. In matrix notation, the model is expressed as

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} d_{11} & \cdots & d_{1h} & \cdots & d_{1k} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ d_{i1} & \cdots & d_{ih} & \cdots & d_{ik} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ d_{n1} & \cdots & d_{nh} & \cdots & d_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_h \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} \tag{3}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{X}$ is the $n \times k$ matrix with the known constants d_{ih}, $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown parameters, $\mathbf{u}$ is the $n \times 1$ vector of unobservable errors. Given the model specification in (3), it can be immediately verified that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & & n_h & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & n_k \end{bmatrix} \tag{4}$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^{n_1} y_{i1} \\ \vdots \\ \sum_{i=1}^{n_h} y_{ih} \\ \vdots \\ \sum_{i=1}^{n_k} y_{ik} \end{bmatrix}, \tag{5}$$

where n_h is the size of group *h*. Therefore, the elements of the least squares estimator $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$ are the group mean incomes $\bar{y}_h$,

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_h \\ \vdots \\ \bar{y}_k \end{pmatrix}. \tag{6}$$

As shown in Palacios-González and García-Fernández (2012, p.1546), even though the model in (3) does not include the intercept, the decomposition of the total sum of squares (*TSS*) into the explained sum of squares (*ESS*) and the residual sum of squares (*RSS*) is valid. Then, the coefficient of determination of the model in (3)

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\left(\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}}{\left(\mathbf{y} - \mathbf{1}\bar{y}\right)'\left(\mathbf{y} - \mathbf{1}\bar{y}\right)} \end{aligned} \tag{7}$$

is equivalent to P^* in (2).[2] Using (7) the link between the income polarization and the levels of one of the characteristics of income receivers can be investigated: values of R^2 close to 1 suggest that grouping income receivers by the levels of one of their characteristics creates groups which are internally homogenous; on the contrary, low values of R^2 indicate that an income receiver does not identify himself much with the other members of his group (i.e. those sharing the same level of the characteristic under consideration).

3. Using regression trees for detecting homogenous groups

The regression tree is a nonparametric method for finding patterns or predicting new observations in data mining (Hsiao and Shih, 2006). Regression trees are able to capture nonlinear relationship between the response variable and explanatory variables, and to summarize results with an intuitive graphic. In addition, unlike other statistical methods (e.g. linear regression, ANOVA) regression trees do not require specific distribution assumptions. For these reasons, the regression tree method is a flexible statistical tool which has been applied in various research fields, such as ecology (De'ath and Fabricius, 2000), finance (Campanella, 2014) and epidemiology (Gass *et al*., 2014). Here we present the regression trees as an explorative statistical tool for uncovering the relationships between income and the characteristics of income receivers. Let $(Y, \mathbf{X}) : \Omega \to \left(S_Y \times S_{X_1} \times \cdots \times S_{X_p} \right) \equiv S$ be a vector random variable defined on the probability space (Ω, F, P), where Y is a numerical response variable and $\mathbf{X} = \left\{ X_1, \ldots X_j, \ldots, X_p \right\}$ is a set of p explanatory variables. Assume that Y is income and $\mathbf{X}$ is the vector collecting p income receiver's characteristics. The regression tree is built by recursively partitioning the space S into disjoint subsets, such that each subset comprises income receivers who are as homogenous as possible with respect to Y. The income receivers comprised in a subset constitute a group which is characterized by the group mean income and the combination of the levels of the characteristics that defines the group. From this standpoint, maximizing within-group homogeneity is equivalent to minimizing variance within groups. Therefore, a rule based on ANOVA is used to repeatedly split income receivers into more homogeneous groups.

Define the variance of the values of Y within subset t as follows:

$$\sigma_t^2 (Y) = n_t^{-1} \sum_{\mathbf{X}_i \in t} \left(y_{it} - \bar{y}_t \right)^2 , \tag{8}$$

[2] A model with more explanatory variables can produce a higher R^2, but this result may be caused by overfitting. To avoid this problem, the adjusted R^2 can also be used as a measure of polarization, as noted by an anonymous referee.

where $\bar{y}_t$ is the mean income within subset *t* and n_t is the number of income receivers in subset *t*. Let $c \in S_{X_j} \mid t$ stand for a value of X_j within the domain of X_j restricted to subset *t*. The variance reduction due to splitting *t* into two parts, t_L and t_R, at the threshold *c* is

$$\Delta_t(Y,c) = \sigma_t^2(Y) - \left[\frac{n_{t_L}}{n_t} \sigma_t^2(Y \mid X_j \le c) + \frac{n_{t_R}}{n_t} \sigma_t^2(Y \mid X_j > c) \right], \qquad (9)$$

where $n_{t_L} = \sum_{i=1}^{n_t} I_{\{X_{ij} \le c\}}$ and $n_{t_R} = \sum_{i=1}^{n_t} I_{\{X_{ij} > c\}}$ are the numbers of income receivers in subsets t_L and t_R, respectively. For subset *t*, the splitting variable and the variable split *c* are selected from all possible splits of the explanatory variables in order to maximize the variance reduction in (9). We note that maximizing (9) is equivalent to maximizing $\Phi_t(Y,c) = n_t \Delta_t(Y,c)$; that is, one searches for the split that minimizes the residual sum of squares

$$RSS_t(Y,c) = \sum_{\mathbf{X}_i \in t_L} \left(y_{it_L} - \bar{y}_{t_L} \right)^2 + \sum_{\mathbf{X}_i \in t_R} \left(y_{it_R} - \bar{y}_{t_R} \right)^2 . \qquad (10)$$

It follows that a subset is formed in *S* by splitting a parent subset into two parts through a binary split of the support of an explanatory variable X_j; therefore, a subset is characterized by the explanatory variables and variable splits which define it.

At the beginning of the recursive partitioning procedure ungrouped income receivers are considered, and then the whole space *S* is split into two parts by selecting the most effective variable (and variable split) in reducing the overall variance of *Y* by minimizing within-group variance. The binary splitting is repeated for each subset until the tree has grown large enough so that no further splitting yields a variance reduction, which overcomes a pre-established minimal threshold. As pointed out in Breiman *et al*. (1993), it is convenient to set a small value for the threshold,[3] growing an overlarge tree and then searching for the best tree. Tree pruning is used to find the best tree. Pruning can be performed by minimizing the following cost-complexity function for a tree *T*

$$R_\alpha(T) = R(T) + \alpha |T|, \qquad (11)$$

[3] Setting a large threshold serves the scope of excluding a split if it does not produce an appreciable reduction in variance; however, if that split is made, one of the descendent subsets may be split in a way to yield an appreciable decrease in variance. This can occur when a split based on interactions among variables yields an appreciable decrease in variance, but none of the associated variable main effects produces an appreciable variance reduction (De'ath and Fabricius, 2000, pp. 3183).

where $|T|$ is the tree size, that is the number of terminal subsets, α is a complexity parameter ranging within the interval $[0,\infty)$, and $R(T)$ is the resubstitution estimate of error which coincides with the residual sum of squares of Y for a regression tree with size $|T|$ (De'ath and Fabricius, 2000).[4] As shown in Breiman *et al*. (1993), for any α there is a unique smallest tree which minimizes (11), therefore, finding the best tree reduces to choosing the best tree size. The strategy for selecting the optimal tree size is discussed in the empirical analysis shown in the next section (Section 4).

Once the regression tree has been pruned, $|T|$ homogenous groups are identified. The measure of polarization P^* (i.e. R^2) is calculated for this population partition. Unlike the Palacios-González and García-Fernández approach, where groups are pre-established, the identification of the $|T|$ groups arises from the structure of the data by clustering observations with similar income values. Therefore, using regression trees, polarization patterns can be naturally uncovered from data.

Another difference between the regression tree and the Palacios-González and García-Fernández ANOVA model is that the former can deal with high-order interaction effects among explanatory variables, whereas the latter can only capture the main effects of the variable used to define groups. It is worth mentioning that the Palacios-González and García-Fernández model could be extended to include interaction effects among explanatory variables; however, the interactions need to be specified *a priori*. Using regression tree, only the interactions which actually contribute to growing the tree are included in the fitting process; therefore, we can say that interactions are specified in a data-driven way, as noted in Strobl *et al*. (2009).

3.1. Comparison with other methods for measuring income polarization

Since other approaches for analysing income polarization have been proposed in the literature, it is worth underlining the differences between these approaches and the approach based on regression trees. Esteban and Ray (1994) define a class of indices to measure income polarization. The Esteban and Ray polarization index is based on the pairwise comparisons between groups, where each group is identified by its income level and size:

$$ER = \sum_{i=1}^{k}\sum_{j=1}^{k} n_i^{1+\alpha} n_j \left| y_i - y_j \right| \quad \text{with } \alpha \in [1;1.6], \tag{12}$$

where k is the number of groups, y_i is the income level of group i and n_i is the size of group i. The *ER* index depends on the choice of parameter α. To apply

[4] The introduction of α which handles the trade-off between $R(T)$ and the tree size is necessary since the residual sum of squares will always be minimized by the largest tree (Sutton, 2005, p. 311); however, the larger the tree, the lower its interpretability. The use of a cost-complexity measure avoids choosing trees with very small $R(T)$, but too large to be interpreted clearly.

the *ER* index, the choice of the criterion to form k groups is required. In doing so, the groups may be formed by partitioning the income distribution into k non-overlapping income ranges or by establishing an external criterion (e.g. age, occupation, geographical area, education level) which a priori splits the population into k groups. Unlike the Esteban and Ray approach to polarization, the approach based on regression trees finds groups in a data-driven way by searching for the partition maximizing within-group homogeneity. Using the tree-based approach, R^2 is used to measure polarization whereas the index of polarization in (12) is used by applying the Esteban and Ray approach.

In the income distribution literature, polarization has also developed by following an alternative approach focusing on the concept of bipolarization; that is, the extent to which incomes spread from the middle to the tails of the distribution, implying the disappearance of the middle class (Wolfson, 1994). Wolfson (1994) suggests an index to measure bipolarization in income distribution. Let $Me(\mathbf{y})$ stand for the median income. Let $\mathbf{y}^+$ be the vector with the incomes above the median income and $\mathbf{y}^-$ be the vector with the incomes below the median income. $\mu(\mathbf{y}^+)$ and $\mu(\mathbf{y}^-)$ being the mean incomes above and below the median respectively, the Wolfson index is

$$W = \frac{2\mu(\mathbf{y})}{Me(\mathbf{y})}\left[\frac{\mu(\mathbf{y}^+)-\mu(\mathbf{y}^-)}{\mu(\mathbf{y})} - G(\mathbf{y})\right], \qquad (13)$$

where $\mu(\mathbf{y})$ is the overall mean and $G(\mathbf{y})$ is the Gini index of inequality. When measuring bipolarization, the median is considered as a threshold for partitioning the distribution into a lower portion and an upper portion; then, the concentration of incomes around two poles on opposite sides of the median is observed.

4. Application to income data

We apply regression trees to individual incomes collected by the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy in 2010 (Banca d'Italia, 2012). The SHIW is carried out every two years, and each survey sample comprises households interviewed for the first time and households interviewed in previous surveys (panel households). The SHIW data is one of the most frequently used information source to investigate income inequality in Italy (Mussini, 2013; Zenga 2007), since the survey collects information on income and socioeconomic status for every household member. The sample size of the 2010 survey is 7,951 households, including 19,836 individuals. We perform the analysis on individual incomes, considering 13,733 income receivers. Table 1 shows some descriptive statistics for the subsample under consideration.

Table 1. Descriptive statistics for the income distribution

number of observations	minimum	first quartile	median	mean	third quartile	maximum
13,733	-7,345.2	10,131.7	16,073.0	19,155.3	23,711.4	573,383.9

Source: Calculations on SHIW 2010 data.

The set of characteristics of income receivers used as explanatory variables is shown in Table 2. Applying regression trees enables one to detect which characteristics play a role in explaining the income received by an individual. The combinations of the characteristics defining the $|T|$ terminal subsets identify $|T|$ exhaustive and mutually exclusive groups of income receivers.

Table 2. Explanatory variables description and coding

name	description	type	categories coding (for categorical variables) or range (for numeric variables)
age	age	numerical	(0, 102] years;
area	geographical area of residence	nominal	N="North", C="Centre", S="South and Islands";
employment	employment status	nominal	BC="blue-collar worker", OW="office worker or school teacher", M="cadre or manager", P="sole proprietor/member of the arts or professions", SE="other self-employed", R="retired", NE="other not-employed";
status	marital status	nominal	M="married", S="single", D="separated or divorced", W="widowed"
education	educational qualification	ordinal*	N="none", P="primary school certificate", LS="lower secondary school certificate", VS="vocational secondary school diploma", US="upper secondary school diploma", B="3-year university degree", G="5-year university degree", PG="postgraduate qualification";
activity	sector of activity	nominal	A="agriculture, fishing", I="industry", G="general government", O="other", NA="do not know";
gender	gender	dichotomous	F="female", M="male";
size	size of the town of residence	ordinal	ST="0-20,000 inhabitants", MT="20,000-40,000", LT="40,000-500,000", C="more than 500,000 inhabitants";
Italian	citizenship	dichotomous	I="Italian", F="not Italian";
health	state of health	ordinal	VP="very poor", P="poor", F="fair", VG="good", E="excellent";
home	individual's home status	nominal	O="owned", R="rented or sublet", UR="under redemption agreement", U="occupied in usufruct";

*Source: SHIW 2010. *Ordinal variable categories are listed in ascending order.*

The tree grows large by setting a small value of the complexity parameter (*cp*) to avoid that interaction effects among explanatory variables are not discovered because none of the associated main effects produces a split with an appreciable decrease in variance.[5] Table 3 shows the resubstitution relative error (RE=1-R^2), the 10-fold cross-validation relative error (RE^{CV}), and the standard error of the 10-fold cross-validation relative error (*SE*) for different tree sizes. From Table 3, we observe that the pre-pruning tree has twenty six terminal subsets. Cross-validation is used to obtain more accurate estimates of (prediction) relative error for trees of a given size (see Breiman *et al*., 1993, pp. 234-237).[6] Cross-validation estimates of relative error can be used to select the optimal tree size by choosing the size with minimum cross-validation relative error. However, to select the optimal tree size we follow the 1-*SE* rule proposed by Breiman *et al*. (1993). The 1-*SE* rule suggests choosing the smallest tree T such that

$$RE^{CV}\left(T\right) \leq RE^{CV}\left(T_{\min}\right) + SE \ , \tag{14}$$

where $T_{\min}$ is the tree with minimum cross-validation relative error and *SE* is the associated standard error estimate. The rationale for the use of the 1-*SE* rule is that it usually selects a much smaller (and more interpretable) tree than that suggested by the minimum cross-validation relative error, entailing a minimal increase in the cross-validation relative error (less than *SE*).

[5] We use the R package rpart (Therneau *et al*., 2012) for recursive partitioning and we set the *cp* equal to 0.0025. The *cp* value in rpart has a meaningful interpretation since it is equal to the increase in R^2 that a split has to produce in order to be made. It immediately follows that the relationship between *cp* and α in equation (11) is $\alpha = TSS \cdot cp$, where *TSS* denotes the total sum of squares of *Y*. Therefore, when setting *cp*, one also defines α.

[6] 10-fold cross-validation is performed as follows: (*i*) observations are divided into ten subsets of approximately equal size; (*ii*) each subset in turn is left out, a tree of size $|T|$ is built using the remaining subsets, and this tree is used to predict the response variable values for the omitted subset; (*iii*) the prediction errors are calculated for each omitted subset by adding up the squared differences between the observed and predicted values; (*iv*) the sums of prediction errors calculated for the ten subsets are added up, and the total sum of prediction errors $R^{CV}(T)$ is divided by *TSS* to obtain the 10-fold cross-validation relative error $RE^{CV}(T)$ for a tree with size $|T|$; (*v*) steps (*i*)-(*iv*) are repeated for every tree size.

Table 3. Resubstitution relative error ($RE(T)$) and 10-fold cross-validation relative error ($RE^{CV}(T)$) by tree size

$\|T\|$	*cp*	Number of splits	$RE(T)$	$RE^{CV}(T)$	*SE*
1	0.106872	0	1.00000	1.00017	0.09527
2	0.049124	1	0.89313	0.89395	0.09285
3	0.035976	2	0.84400	0.84488	0.09240
4	0.029526	3	0.80803	0.80903	0.09083
5	0.022208	4	0.77850	0.78202	0.08846
6	0.013902	5	0.75629	0.76205	0.08779
7	0.013235	6	0.74239	0.75649	0.08743
8	0.011640	7	0.72916	0.73876	0.08732
9	0.010856	8	0.71752	0.72990	0.08707
10	0.007842	9	0.70666	0.71721	0.08689
11	0.007525	10	0.69882	0.71631	0.08696
12	0.007216	11	0.69129	0.71462	0.08692
13	0.004419	12	0.68408	0.70086	0.08669
14	0.004253	13	0.67966	0.70096	0.08675
15	0.003642	14	0.67541	0.69789	0.08681
16	0.003585	17	0.66438	0.70033	0.08680
17	0.003570	18	0.66079	0.69947	0.08679
18	0.003459	19	0.65722	0.69953	0.08679
19	0.003393	20	0.65376	0.69938	0.08679
20	0.003384	21	0.65037	0.69825	0.08678
21	0.002977	22	0.64699	0.69782	0.08679
22	0.002853	23	0.64401	0.69650	0.08676
23	0.002774	24	0.64116	0.69552	0.08663
24	0.002766	25	0.63838	0.69132	0.08666
25	0.002549	26	0.63562	0.68918	0.08665
26	0.002500	27	0.63307	0.68661	0.08665

Source: Calculations on SHIW 2010 data.

From Table 3, we see that the tree with six terminal subsets is the smallest tree which satisfies (14). Once the optimal tree size has been selected, the tree is pruned.[7] Figure 1 shows the pruned tree where six groups are detected by the terminal subsets 4, 6, 7, 10, 22, 23. Each terminal subset in Figure 1 shows its size and mean income.

[7] Practically speaking, pruning is performed through the R package rpart by replacing the *cp* value used to grow the overgrown tree with the *cp* value that generates a tree with six terminal subsets in Table 3 (i.e., *cp*=0.013902).

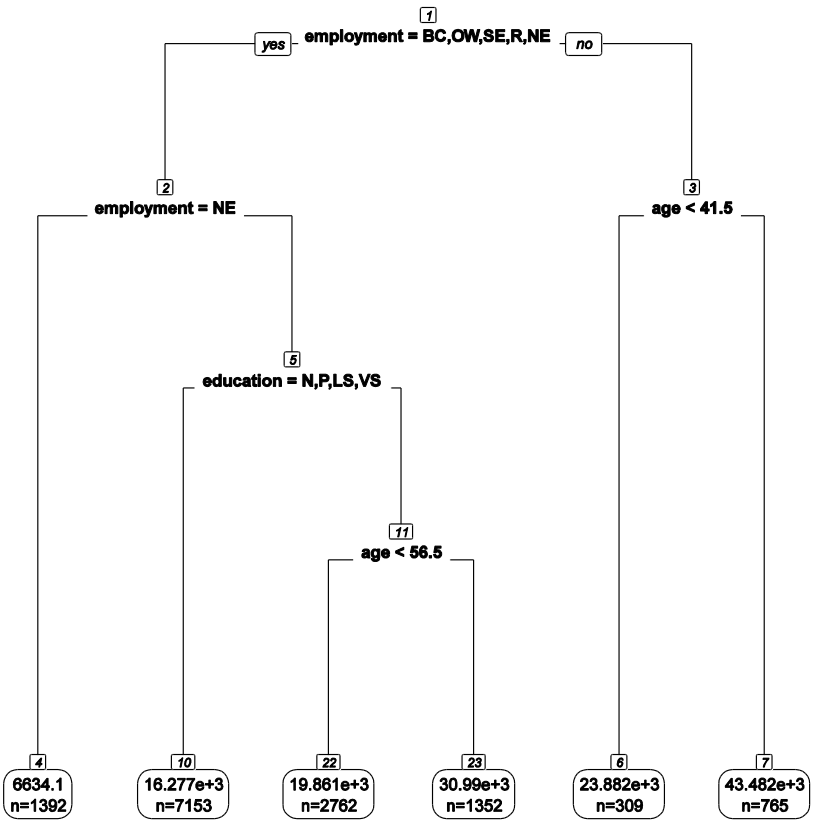


Figure 1. Regression tree analysis of income polarization

Only three variables (education, employment, age) from the set of explanatory variables in Table 2 are discriminating in recursive partitioning income receivers into subsets. The employment main effect distinguishes between individuals whose employment status is equal to M or P and the remaining individuals; that is, the employment status determines the initial partition between high-skilled workers or business owners (M or P) and the other workers (BC, OW, SE) or not working individuals (R or NE). This means that the main effect of the employment status is more important than those of the other variables in determining differences in income. Subset 4 comprises unemployed individuals and has the lowest mean income (6,634.1 EUR). The use of regression tree enables one to identify subsets 6 and 7, since the regression tree also accounts for interaction between employment and age: among the income receivers whose employment status is M or P, individuals younger than 41.5 years old (subset 6) receive much

lower incomes than those older than 41.5 years old (subset 7). Education has an effect on income for individuals whose employment status is BC, OW, SE or R: individuals with educational qualifications lower than or equal to VS (subset 10) receive lower incomes than those with educational qualifications higher than VS (hereafter, high-educated workers); then, among high-educated workers, incomes are higher for individuals older than 56.5 years old (subset 23). Subset 10 is the largest subset, including more than half of the income receivers in the sample. It is worth mentioning that age does not play a role in determining income in subset 10 (low-educated workers), whereas age is discriminating in subset 11 (high-educated workers). This finding suggests that high-educated workers have chances of increasing their income during their career; this age effect is not present for low-educated workers. More specifically, the mean income of high-educated BC, OW and SE workers older than 56.5 years old (30,990 EUR) is almost twice the mean income of low-educated workers in the same occupations (16,277 EUR).

The above discussed partition is detected by discovering the different patterns of income existing in the income distribution: income receivers comprised in the same group share the same income pattern which differs from those of the other groups. Therefore, each income receiver identifies himself with those sharing the same income pattern and feels alienated from income receivers with different income patterns. The R^2 calculated for the partition detected by the regression tree is equal to 0.24371 and measures the polarization in the income distribution.

5. Concluding remarks

The contribution of the article is two-fold. First, we show that the regression tree is a useful statistical tool to investigate whether incomes concentrate around local poles. The regression tree identifies groups which are internally homogeneous in a data-driven way: income receivers are recursively partitioned into groups by selecting the explanatory variables that actually contribute to defining groups of income receivers with similar income levels. Other distinguishing features of regression trees are the ability to capture nonlinear relationships between explanatory variables and income, and the intuitive graphic interpretation of results. Therefore, regression tree can be seen as a flexible and practical technique to explore income polarization.

Second, we extend the ANOVA-based approach to polarization measurement proposed by Palacios-González and García-Fernández (2012), since we point out that using regression trees instead of one-way ANOVA we are able to detect not only the main effects of explanatory variables but also their interaction effects. This enables analysts to discover polarization patterns that cannot be assumed *a priori*. For instance, our empirical analysis of Italian income data shows that the interactions among employment status, educational qualification and age form well-identified groups of income receivers, whereas the other characteristics do not play a clear role in explaining income polarization.

Further research will be devoted to extending the approach based on recursive partitioning to the analysis of polarization when the response variable is ordinal (e.g. level of satisfaction, health status) instead of numeric (e.g. income). In the first instance, this requires the definition of a proper polarization-sensitive impurity function that can be used for recursive partitioning, as the residual sum of squares is suited to the tree-based model for income polarization.

Acknowledgements

The author would like to thank two anonymous reviewers for helpful comments which helped to improve the manuscript.

REFERENCES

BANCA D'ITALIA, (2012). Survey on Household Income and Wealth 2010, Rome, Italy, 2012.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J., (1993). Classification and regression trees, Chapman & Hall/CRC press, Boca Raton.

CAMPANELLA, F., (2014). Assess the Rating of SMEs by using Classification and Regression Trees (CART) with Qualitative Variables, Review of Economics & Finance, 4, 16–32.

DE'ATH, G., FABRICIUS, K. E., (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis, Ecology, 81, 3178–92.

DEUTSCH, J., FUSCO, A., SILBER, J., (2013). The BIP trilogy (Bipolarization, Inequality and Polarization): one saga but three different stories, Economics, Discussion Paper No. 2013–22.

DUCLOS, J. Y., ESTEBAN, J. M., RAY, D., (2004). Polarization: Concepts, measurement, estimation, Econometrica, 72, 1737–72.

ESTEBAN, J. M., RAY, D., (1994). On the measurement of polarization, Econometrica, 62, 819–51.

GASS, K., KLEIN, M., CHANG, H. H., FLANDERS, W. D., STRICKLAND, J., (2014). Classification and regression trees for epidemiological research: an air pollution example, Environmental Health, 13:17.

HSIAO, W. C., SHIH, Y. S., (2006). Splitting variable selection for multivariate regression trees, Statistics and Probability Letters, 77, 265–71.

MUSSINI, M., (2013). A matrix approach to the Gini index decomposition by subgroup and by income source, Applied Economics, 45, 2457–2468.

PALACIOS-GONZÁLEZ, F., GARCÍA-FERNÁNDEZ, R. M., (2012). Interpretation of the coefficient of determination of an ANOVA model as a measure of polarization, Journal of Applied Statistics, 39, 1543–55.

POGGI, A., SILBER, J., (2010). On polarization and mobility: a look at polarization in the wage-career profile in Italy, Review of Income and Wealth, 56, 123–140.

STROBL, C., MALLEY, J., TUTZ, G., (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests, Psychological Methods, 14, 323–48.

SUTTON, C. D., (2005). Classification and regression trees, bagging, and boosting, in Handbook of statistics 24: data mining and data visualization (Eds.) C. R. Rao, E. J. Wegman and J. L. Solka, Elsevier, Amsterdam, pp. 303–29.

THERNEAU, T., ATKINSON, B., RIPLEY, B., (2012). Rpart: recursive partitioning and regression trees, R package version 3.1–55.

WANG, Y. Q., TSUI, K. Y., (2000). Polarization orderings and New Classes of Polarization Indices, Journal of Public Economic Theory, 2, 349–63.

WOLFSON, M. C., (1994). When inequalities diverge? American Economic Review, 84, 353–58.

ZENGA, M., (2007), Inequality curve and inequality index based on the ratios between lower and upper arithmetic means, Statistica & Applicazioni, 5, 3–27.

ZHANG, X., KANBUR, R., (2001). What difference do polarization measures make? An application to China, Journal of Development Studies, 37, 85–98.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 237–247

QUALITY OF LIFE AND POVERTY IN UKRAINE – PRELIMINARY ASSESSMENT BASED ON THE SUBJECTIVE WELL-BEING INDICATORS[1]

Oleksandr Osaulenko[2]

ABSTRACT

The paper provides an overview of the information sources, methodology and main findings of the research of quality of life and poverty using indicators of subjective well-being applied by state statistics agencies in Ukraine. The paper describes the system of indicators for self-evaluation of the attained level of well-being, the level of satisfaction from meeting the basic living needs, and the limitations in consumption abilities of selected population groups due to hard conditions. In addition, methodological approaches in national statistics practice are discussed for the case of analysis of economic deprivation and for infrastructure development as indicator of geographic accessibility of services and non-geographic barriers causing the deprivation of access. Also, this paper reviews the factors that underlie the deprivations and define the percentage of population that is particularly affected by multiple deprivation in Ukraine. It covers the data on dynamics and analyses the distribution of deprivation by different population group, for several years. Finally, it describes further steps on the way to enhance the information capacity of subjective wellbeing studies, particularly as regards implementation of the contemporary approaches in international perspective, including Europe.

Key words: poverty, level of well-being, household, subjective well-being, deprivation.

1. Introduction

Given current socio-economic conditions, one of the most pressing tasks is to improve the efficiency and targeting of social support and improvement of social administration at all levels, from state level to local communities. The practical solution to this problem requires improvements in relevant information and

[1] The paper basis on the presentation given at the 60th World Statistics Congress in Rio de Janeiro (July 26-31, 2015).

[2] National Academy of Statistics, Accounting and Audit. Kyiv, Ukraine.
E-mail: O.Osaulenko@nasoa.edu.ua.

analytical support: application of integrated approach and different sources of data for in-depth research of material conditions of population, efficiency analysis of measures on social protection of vulnerable groups, the risks and factors that affect the well-being and social stability to develop appropriate pre-emptive measures (Osaulenko et al., 2004; Ministry of Economic Development and Trade of Ukraine and UNDP, 2013). Considering that any democratic society strives for self-awareness, the employment of an integrated approach in social research improves the quality of such public information, and, therefore, the development of democratic principles in Ukraine.

A comprehensive approach in social research requires different types of information. They will determine the scope and direction of social changes from different perspectives and define direct and hidden cause-effect relations and factors that have the most significant influence on development of effective social policies. This will allow to achieve the most positive results – the advance in living standards of population, social cohesion, creation of favourable and equal opportunities for personal development, positive improvements in public assessment of social protection policies, in particular with regard to targeted aid, provided minimization of state funds for functioning of the system (Cherenko, 2006; Libanova, 2008).

Market transformations in economy, increased differentiation of the living conditions of some population groups and related aggravation of the poverty issue, as well as increased interest of the authorities and society in objective information have given a powerful impetus to the development of specialized state statistical observations and surveys of population. Currently, Ukraine state statistics regularly conducts three population sample surveys on: household living conditions, economic activity and agricultural activity in rural areas.

The multi-vector nature of living conditions survey of general population, in particular the most vulnerable population groups, requires a multidimensional approach to define and characterize not only the key indicators, but also concepts, processes and phenomena. In particular, there are numerous internationally recognized approaches to measuring poverty, low-income and social exclusion. Each of them has its advantages and disadvantages, they do not provide definitive assessments of events and heavily depend on research objectives and national specificity (Kangas and Ritakallio, 1998; Ramplakash, 1994).

The analysis of living standards should be based not only on the objective information, i.e. administrative data and data of continuous state statistical observations. An important role in addressing the integrated multidimensional nature of such a research phenomenon as well-being is played by thematic modular sample surveys of population. Subjective evaluations of living standards, made directly by respondents of the sample survey, reflect the degree of satisfaction of population with living standards, particularly with their possibilities of satisfying not only the minimum physiological needs but also the needs for personal development and enhancement of living comfort. Subjective evaluations indirectly display the actual level of satisfaction of population with the existing socio-economic provisions and the results of the public authorities' activity. Despite the limitations that are typical for measurements based on

"subjective attitude", this method has become widespread in many countries, including Ukraine, over recent years.

The analysis of living standards of population, poverty and other closely related issues of subjective well-being (a person's self-evaluation of the level to which essential means, physical, social, cultural and spiritual benefits are accessible by her/him (completeness)), is gaining exceptional relevance today (Cherenko, 2015; Libanova et al., 2013; State Statistics Service of Ukraine et al., 2013). Current uneasy social and economic realities, the antiterrorist operation (ATO) in Donetsk region and related immense migration pose a number of challenges to state statistics bodies of Ukraine, and give more focus to the determinants of subjective well-being. The challenging tasks also emerge during the implementation of the EU regulations and standards in national statistical practice (Commission regulation (EC), 2003; Vogel, 1997).

The information base for the analysis of subjective well-being in Ukraine is obtained from the sample household living conditions survey conducted by the state statistics bodies and modular polls based on the survey (State Statistics Service of Ukraine, 2013, 2014). This survey is a unique source for comprehensive studies of Ukraine's population well-being. It enables analysis of various spheres of household life, which depends on their level of income (expenses), composition, presence of children, place of residence and other criteria. An annual effective sample amounts to about 10,000 households. The analytical potential is significantly expanded by the combination of sociological questions with "subjective" ratings on attitudes, expectations and aspirations of certain groups, identification of their needs, and self-evaluation of their well-being within the survey research. These thematic surveys provide a unique opportunity of combining the information on the actual financial situation of each surveyed household with its subjective evaluations by household members.

2. The system of indicators to characterize the self-evaluation of the achieved well-being and the degree of satisfaction of basic living needs

Self-evaluation of well-being by households is made by subjective determination of the adequacy of their income to meet basic needs, information on limitations of consumption abilities due to lack of funds, and by social self-identification. Household self-evaluation involves the selection of alternative responses to questions referred to the following system of structural indicators:

- subjective determination of the adequacy of annual household income (had enough and made savings; enough, but did not make any savings; constantly denied the essentials, except for food; could not afford even adequate food);
- consumption abilities of certain groups of households (the presence or absence of cases of inability to meet individual needs due to lack of funds,

namely the possibility of daily consumption of hot food; hunger cases among adults and children due to lack of funds; inclusion of fruits or juices in a child's diet; the ability to give children food or money for food at school; the ability to give children treats at least once a week; the ability to pay for children in kindergarten);

- self-identification of households as representatives of certain population groups (rich, middle class, not poor but not middle class yet, poor).

A comprehensive analysis of well-being is not possible without such important aspects as public accessibility of health services and grounds for unmet needs in health care, the ability to purchase medicines and medical devices. Even the effective and comprehensive system of administrative data collection is not able to reflect the whole picture in the area, as public health institutions provide information only about people who employ the health care system. Information on individuals, who do not use the services, and, in particular, about the reasons for that, may be obtained only from other sources, namely from sample population surveys. For this purpose, in Ukraine, the following indicators are developed:

- the level of accessibility of medical aid for household members, ability to purchase medicines and medical supplies if such needs emerge (share of households whose needs were satisfied);
- distribution of households which did not satisfy these needs, because of lack of access to services (too high cost, failed to find the desired one, too long queue to see a doctor, there were no appropriate specialist, or no required department in the hospital, no free place).

Well-being largely depends on living conditions, the availability of modern amenities in housing as well as the availability of subsidiary farms and other property in a household. The survey program studies these issues in sufficient detail. Subjective evaluation is presented by distribution of households by the degree of satisfaction with their living conditions (very dissatisfied, dissatisfied, not very satisfied, satisfied, very satisfied).

As the survey results reveal, the psychological impact of the crisis and economic insecurity have a greater influence on self-evaluation than the actual financial state. Thus, the level of subjective poverty, defined as the share of households which consider themselves as such, increased from 59% to 65% during 2009-2013. However, the poverty rate for other household's self-evaluation of income (always denied themselves the essentials, except for food, or they could not afford even adequate food) decreased from 44% to 39% during this period. Figure 1 below reflects more clearly the interrelation between objective

(poor by absolute[3] and relative criteria[4] (Ministry of Social Policy of Ukraine et al., 2012)) and subjective poverty.

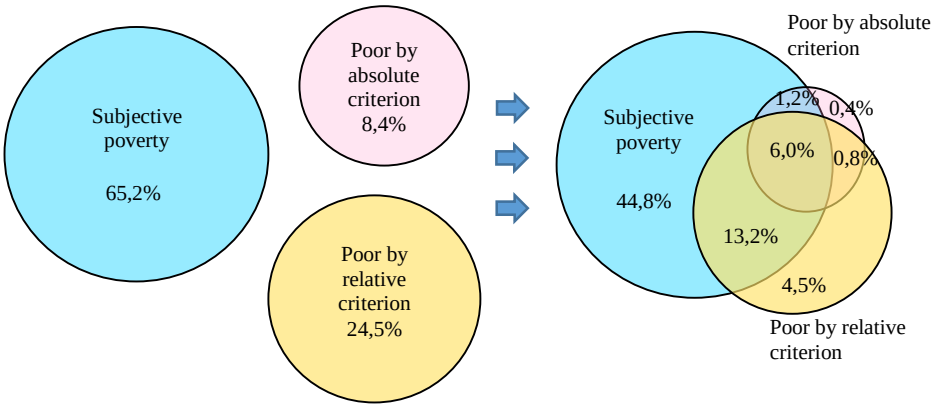


Figure 1. Interrelation of objective and subjective poverty evaluations in 2013

3. Methodological approaches to evaluation of economic deprivation and deprivation of access.

Another area of well-being analysis that is based on subjective evaluation of population is the research of deprivation poverty. This form of poverty is characterized by limited abilities of the population to access certain essentials, which not only cover the minimum physiological needs but also the needs associated with the personal development and assurance of the adequate level of living comfort. To provide the information base for the deprivation study, a modular survey is conducted on the basis of household living conditions survey. It extends the classic (monetary) understanding of poverty through non-monetary subjective indicators. The program of the survey is designed having regard to the modern international experience and the needs of national users. The module for household deprivation and factors that are often subjectively perceived by public as signs of poverty, allows for the following directions of the survey:

- Economic deprivation due to inadequate level or quality of:

[3] Absolute poverty line corresponds to the amount of the legal subsistence level per month per person, annually approved by the Verkhovna Rada of Ukraine in the Law on State Budget of Ukraine for the relevant year.

[4] Relative poverty line is defined by the fixed (75%) share of average per capita total expenditures of the median total expenditures of a particular person who takes the medium position in the list of population ranked by average per capita expenditures calculated for one conventional person.

- meals (lack of funds to ensure a certain quality of meals);
- non-food goods (lack of funds to acquire the needed inexpensive goods, and lack of certain types of these goods);
- housing conditions (lack of normal housing conditions, lack of funds to improve housing conditions);
- health care and education services (lack of funds to obtain the needed inexpensive goods and services);
- income or lack of possibilities of satisfying other important needs.

➢ Development of infrastructure as an attribute of geographical accessibility of services and non-geographical barriers that identify the deprivation of access.

The survey program implies not only the determination of the public perception of signs of poverty and isolation, but also the collection of information on their actual distribution. The national list covered 18 items of deprivation. All items went through frequency control (items indicated by prevailing number of households were selected) and consensus control (items about which majority of respondents felt that their presence is necessary for the normal standard of living). In addition, each item was checked for the interrelation with the level of population well-being. Pearson correlation ratio[5] for almost all types of deprivations indicated a close linkage between distribution of each deprivation and income of households.

Dynamics of changes in distribution of certain deprivations is shown in Figure 2. Each of 18 items of the national list of deprivations contains data for 2009-2013. New items introduced into the program of observation in 2013 are presented for one period.

The incidence degree of certain types of deprivation significantly depended on the place of residence of households. Urban households, as compared to the rural ones, more suffered from financial failure to enlarge the available floor space. Rural residents suffered more than urban residents from all other manifestations of poverty and deprivation, especially from deprivations related to ensuring normal living conditions, availability of amenities in the housing and deprivation associated with low infrastructure development.

[5] Pearson correlation rate was calculated by the distribution of equivalent per capita income and incidence of deprivations among decile population groups.

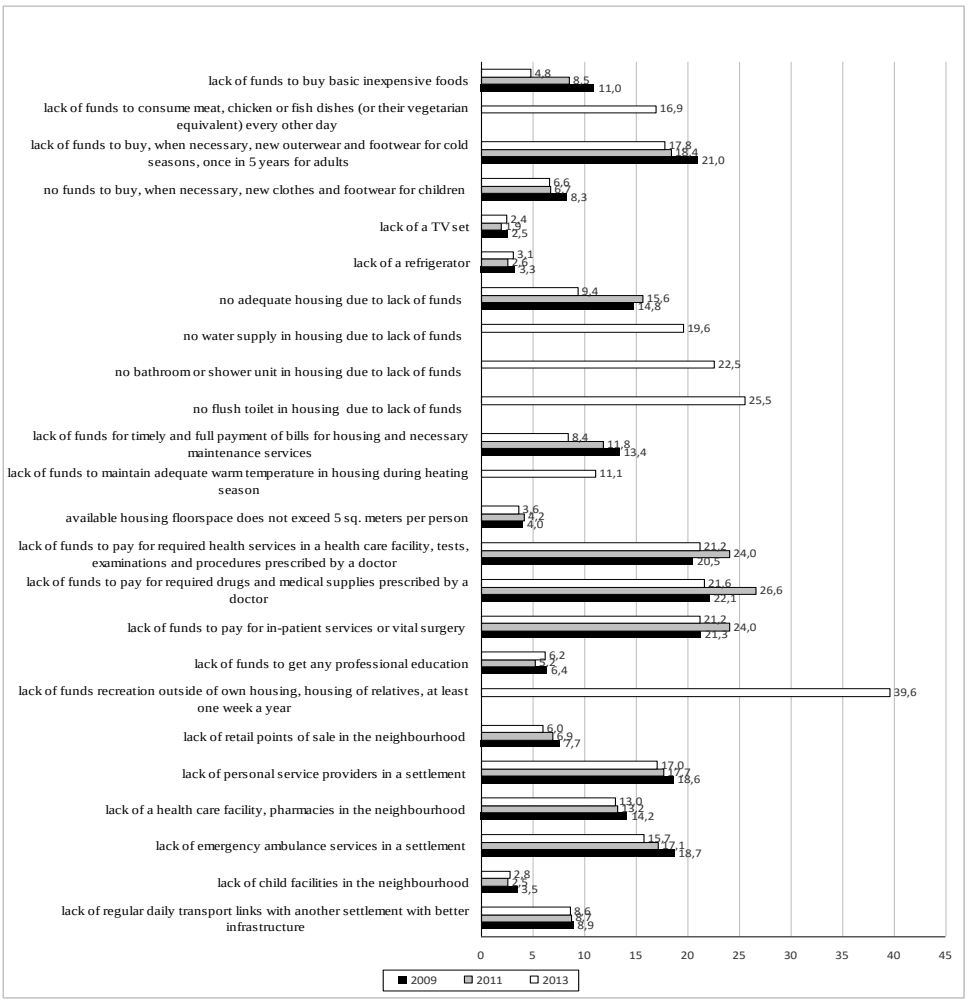


Figure 2. Incidence of certain deprivations among households of Ukraine in 2009-2013

4. Multidimensional assessment of the size of the household group with the lowest standard of living.

The development of information base for improving multidimensional poverty assessment is primarily associated with the use of a combined approach and employment of monetary and subjective criteria as well as the criterion for poverty deprivation. Since the different criteria can differently display poor population, the combined approach reflects the group of households with the highest risk of poverty by all its types.

Figure 3 shows the scope of population poverty defined by different criteria. In 2013, out of population in relative poverty, 32% had 4 or more deprivations. Out of population in absolute poverty, such deprivations were characteristic of 40% of population. On the other side, out of 22% of population which had 4 and more deprivations signs 15% were absolutely poor, and 36% of population were relatively poor. 3% of population were simultaneously at risk of absolute, relative and deprivation poverty. Among the population who were in three types of poverty, the majority (56%) were residents of rural areas. As to the composition, these households mostly had children (79% versus 21% of households without children), half of which had one child, 38% - two and 12% - three and more children. Nearly a quarter of them did not have employed persons in the household, in 44% of households one person was employed. Out of the households with children which were poor simultaneously according to three criteria, almost a quarter had children without one or both parents. 39% of population had at least one of these types of poverty.

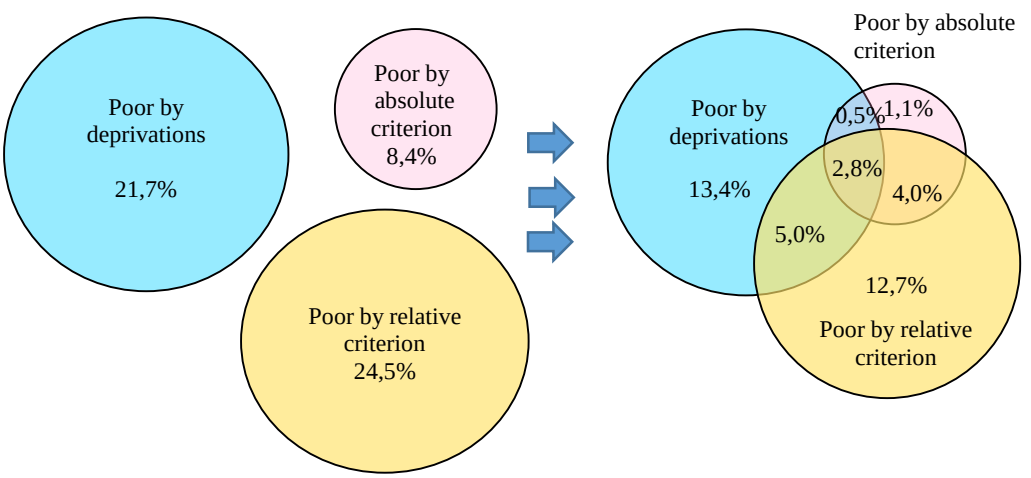


Figure 3. Scope of poverty in 2013, defined by different criteria

The main profiles of poverty by socio-demographic and socio-economic groups of population remain steady. Regardless of the criteria, the level of poverty among people of working age and among people of retirement age is below the national average value, and vulnerable groups include children and "old" pensioners. The most vulnerable traditionally include large families, households with children up to the age of 3 and double demo-economic burden (with children and unemployed).

The profiles of poverty vary depending on the monetary and non-monetary approaches: the high risks of monetary poverty apply to children and non-monetary poverty is much higher among older age groups. When using monetary approaches, the highest poverty risks are typical for large families and for the families with children and unemployed persons. Increased risks are also assigned to households with two or more children, and with children up to the age of 3. When using non-monetary criteria, the group with the highest risk of poverty is represented by households that consist solely of older age groups (75 and older).

4. Conclusions

At present stage of socio-economic development in Ukraine, as in many other countries, social partners pay great attention to the analysis of social inequality, changes in well-being standards that require new approaches to surveys of living standards of population:

- use of subjective evaluations and estimates of material deprivation for the analysis of well-being and poverty, in addition to traditional monetary approaches;
- identification of the most problematic and vulnerable population groups on the basis of multidimensional analysis and combined estimates, obtained by different factors and criteria (for example, poor by income and deprivation criteria, etc.);
- significant differences in living standards in urban and rural areas require more focus on the factorial analysis of key indicators of subjective well-being and material deprivations of the population living in rural areas, especially in the context of the task of optimizing the development of rural areas, which is relevant to Ukraine;
- introduction of the study on the regional differences and territorial determinants in the material deprivations of the population. This direction is of particular importance with regard to administrative and territorial reform in Ukraine. However, the research of territorial determinants will require a significant increase in the size of household sampling and the attractiveness of significant additional financial and human resources for representative and qualitative results for the regional level, which is currently quite challenging for our country.

Acknowledgements

Author would like to express his thankfulness to the referees and the editor for their valuable comments and suggestions which led to improvements on earlier version of the paper.

REFERENCES

CHERENKO, L. M., ed., (2006). Рівень життя населення України [Living standard of population in Ukraine]. Kyiv: Consultant Publishing House, LLC.

CHERENKO, L. M., (2015). Нові форми бідності в Україні: основні прояви та оцінка масштабів явища [New forms of poverty in Ukraine: Key directions and assessment of scale phenomena]. Demography and social economy, 1 (23), 11-21.

COMMISSION REGULATION (EC) 1983/2003 of 7 November 2003, implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the list of target primary variables.

KANGAS, O., RITAKALLIO, V. M., (1998). Different methods, different results? Approaches to multidimensional poverty. In: H. J. Andreß, ed. Empirical Poverty Research in a Comparative Perspective. Aldershot: Ashgate.

LIBANOVA, E. M., (2008). Бідність населення України: методологія, методика та практика аналізу [Poverty of population in Ukraine: Methodology, methods and practices of analysis]. Kyiv: Kyiv National Economic University.

LIBANOVA, E. M., GLADUN, O. M., LISOGOR L. S. and others, (2013). Вимірювання якості життя в Україні: аналітична доповідь [Measuring quality of life in Ukraine: Analytical report]. Kyiv: Ptoukha Institute for Demography and Social Studies of National Academy of Sciences of Ukraine, UNDP, Ministry of Economic Development and Trade of Ukraine.

MINISTRY OF SOCIAL POLICY OF UKRAINE, THE MINISTRY OF ECONOMIC DEVELOPMENT AND TRADE OF UKRAINE, THE MINISTRY OF FINANCE OF UKRAINE, THE STATE STATISTICS SERVICE OF UKRAINE, THE NATIONAL ACADEMY OF SCIENCES OF UKRAINE, (2012). Методика комплексної оцінки бідності [Methodology of comprehensive assessment of poverty]. Available at rada.gov.ua website: http://zakon5.rada.gov.ua/laws/show/z1785-12.

OSAULENKO, O. G., NOVIKOVA, O. F., VLASENKO, N. S., KALACHOVA, I. V. and others, (2004). Інформаційне забезпечення державного та регіонального соціального управління [Information support to state and regional social administration]. Kyiv, Donetsk. Institute of Industrial Economics, State Statistics Committee of Ukraine.

RAMPRAKASH, D., (1994). Poverty in the Countries of the European Union: A Synthesis of Eurostat's Research on Poverty. Journal of European Social Policy, 4 (2), 117–128.

REDUCE POVERTY, (2013). In: Millennium development goals. Ukraine – 2013. National Report. Kyiv: Ministry of Economic Development and Trade of Ukraine, UNDP. 43–54.

STATE STATISTICS SERVICE OF UKRAINE, UKRAINIAN CENTER FOR SOCIAL REFORMS, SWISS COOPERATION OFFICE, USAID and others, (2013). Суб'єктивна оцінка добробуту [Subjective assessment of well-being]. In: Multicluster Household Survey, 2012**.** Final report. Kyiv: KIS. 251–258.

STATE STATISTICS SERVICE OF UKRAINE, (2013). Самооцінка домогосподарствами України рівня своїх доходів у 2013 році [Ukraine household self-assessment of their revenues in 2013]. Available at ukrstat.org website: http://ukrstat.org/uk/operativ/operativ2006/gdn/sdrsd/arh_sdrsd.html.

STATE STATISTICS SERVICE OF UKRAINE, (2014). Самооцінка домогосподарствами України рівня своїх доходів (за даними вибіркового опитування домогосподарств у січні 2014 року) [Ukraine household self-assessment of their revenues (by the data of the household sample survey of January 2014)]. Statistical bulletin. Kyiv: State Statistics Service of Ukraine.

VOGEL, J., (1997). Living Conditions and Inequality in the European Union 1997. Eurostat Working Papers, Population and Social Conditions E/1997-3.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 249–264

ON THE RELATIONSHIPS BETWEEN SMART GROWTH AND COHESION INDICATORS IN THE EU COUNTRIES[1]

Beata Bal-Domańska[2], **Elżbieta Sobczak**[3]

ABSTRACT

Within the framework of the Europe 2020 strategy smart growth is listed as one of the leading policy objectives aimed at improving the situation in education, digital society and research and innovation. The objective of this article is to evaluate the relationships between smart growth and economic and social cohesion factors. Aggregate measures were used to describe smart growth pillars. Here, social cohesion is described by the level of employment rate as one of the conditions essential to the well-being and prosperity of individuals. Economic cohesion is defined by the level of GDP per capita in PPS. Observation of these three phenomena forms the basis for the construction of panel data models and undertaking the assessment of the relationships between smart growth and economic and social cohesion factors. The study was performed on the group of 27 European Union countries in the period of 2002-2011.

Key words: economic and social cohesion, smart growth, European Union countries, panel data analysis

1. Introduction

European economies face many challenges in the contemporary world. Actions outlined in the Europe 2020 strategy present the response of the EU member countries (a strategy for smart, sustainable and inclusive growth). It emphasises the importance of a balanced development of all countries and

[1] The study was conducted within the framework of research grant NCN no. 2011/01/B/HS4/04743 entitled: *European regional space classification in the perspective of smart growth concept – dynamic approach*.

[2] Wrocław University of Economics, Faculty of Economics, Management and Tourism, Department of Regional Economics, Nowowiejska 3, 58-500 Jelenia Gora, Poland. E-mail: beata.bal-domanska@ue.wroc.pl.

[3] Wrocław University of Economics, Faculty of Economics, Management and Tourism, Department of Regional Economics, Nowowiejska 3, 58-500 Jelenia Gora, Poland. E-mail: elzbieta.sobczak@ue.wroc.pl.

regions, particularly by unblocking and initiating growth processes through actions aimed to strengthen three priorities:

- smart growth – i.e. development of the knowledge-driven economy,
- sustainable growth – i.e. transformation towards low-carbon economy, which efficiently uses resources and benefits from competition,
- inclusive growth – i.e. fostering a high-employment economy bringing about social and territorial cohesion.

Countries that provide favourable conditions for smart growth are expected to gain a developmental advantage that manifests itself in the form of a higher level of social progress (for example noticeable in the larger number of workplaces available to individuals); and economic advancement (expressed in a higher output of goods and services).

The new endogenous growth theory (Romer, 1986), (Romer, 1990) directs the focus to the knowledge related factors. It implies the possibility of accumulation of the growth incentives, which creates a favourable environment for a constant development, but at the same time it may add to sustaining or even increasing differences between countries. In this approach, the long-term socio-economic development is based on the gains in human capital resources, physical and technological innovation, which in turn will increase the productivity of traditional growth factors through education, R&D, diffusion of innovation, along with positive spillovers related to the transfer of technology and assets. As (Fiedor, 2010) states, "this growth is based on the increase of the intellectual capital resources in the region by strengthening business support institutions oriented towards creating entrepreneurship and innovation, as well as, forming the web of linkages between the economy and the sphere of education, science and research."

Economic and social cohesion – according to the European Union policy – is about reducing disparities between countries and the lagging behind of the advantaged regions. It should also promote more balanced, more sustainable 'territorial development'.

This article attempts to assess the relationships between smart growth and social and economic cohesion in the EU countries. The focus of the research is not straightforwardly on the process of levelling off of the disparities but rather on establishing whether changes observed in smart growth level can or cannot influence the socio-economic situation and enable the levelling off processes as far as territorial disparities are concerned.

The definition of smart growth is based on the three conceptual pillars:

- innovativeness, as the driving force of economies towards knowledge and innovation,
- creativity, in the form of human capital resources,
- smart specialization, as the existing cutting-edge structures of highly advanced and specialised branches of economy.

The concept of smart growth pillars as well as social and economic cohesion were based on the assumptions made over the course of research study on: European regional space classification in the perspective of smart growth concept – dynamic approach (Markowska, Strahl, 2013).[4]

It is rather difficult to clearly indicate the directions of relationships that link smart growth and social and economic cohesion. It is more appropriate to state that they coexist and are interconnected. Smart growth is seen as the causative factor for achieving social and economic cohesion. Social and economic cohesion supports the expansion of spheres related to knowledge, human capital and innovation, which in turn are needed to create conditions for smart growth. Shifting growth to knowledge and high-tech sectors is not possible without achieving a certain level of socio-economic development, with reference to the aspects related to human capital formation, among others.

The review of selected regional development theories on the role of innovation was presented by Dominiak et al. (2012), Kawa (2007) and Strahl (2010), among others, while human capital aspects were discussed by, e.g.: Herbst (2007) and Cichy (2008).

This analysis of relationships between economic and social cohesion and smart growth is presented as the cross-section of the EU countries in the period of 2002-2011.

2. The research procedure and techniques

The analysis was conducted for all 27 EU Member States (excluding Croatia which joined the EU structures in 2013) in the period of 2002-2011. The Eurostat database[5] was the source of data for all the variables. This ensured comparability of data concerning the analysed countries.

The study was performed in three stages which covered:

I. Defining measures for smart growth, economic and social cohesion
II. Constructing aggregate measures for smart growth, economic and social cohesion
III. Estimating econometric models of economic and social cohesion with smart growth pillars

[4] Grant NCN no. 2011/01/B/HS4/04743.
[5] Internet service http://ec.europa.eu/eurostat.

Stage I. Defining measures for smart growth, economic and social cohesion

Multidirectional and multidimensional relations within socio-economic processes make their measurement a complex task. It is further hindered by limited access to the statistical data necessary to evaluate processes occurring in that area (especially at the administrative level, which is lower than the country level).

Economic cohesion is described by means of Gross Domestic Product *per capita* in PPS (*GDP*). This indicator is widely regarded as a relatively good measure of economic activity. For comparison purposes, these values were calculated as values per 1 inhabitant.

Social cohesion can be defined in the socio-cultural context as the willingness of members of a society to cooperate with each other in order to survive and prosper (Stanley, 2003). The OECD Development Centre describes a cohesive society as one which "works towards the well-being of all its members, fights exclusion and marginalisation, creates a sense of belonging, promotes trust, and offers its members the opportunity of upward social mobility" (OECD, 2011). On the basis of the works of the European System of Social Indicators (EUSI), social cohesion was measured in the context of a system of indicators, which distinguishes between two principle goals of social cohesion across a wide spectrum of life domains (Berger-Schmitt, 2000). The first goal is about reducing disparities, inequalities, and social exclusion within a society, while the second deals with the strengthening of the social capital in a society. Regarding the first goal, regional disparities are taken into account, for example with respect to access to transport, leisure and cultural facilities, educational and health care institutions, employment opportunities or the condition of the natural environment. The social dimension covers many diverse aspects reflected in local residents' quality of life. Therefore, a question arises which social cohesion aspects present the strongest connections with smart growth. In the presented study the employment factor (expressed as the employment rate among population aged 20-64 in % (*EM*)) is defined as the key aspect of social cohesion. The impact of employment issues on social cohesion may be considered in terms of its significance to an individual. In the light of this approach, employment is the basic condition that provides financial means necessary to obtain goods and services. Being at work lays foundations for individual aspirations and advancement, and determines one's social position, thus influencing the overall level of satisfaction derived from life and its quality.

A set of diagnostic indicators for smart growth was suggested. Among them the indicators for each pillar were selected, based on the availability and comparability of data over time for 27 countries (Table 1).

Table 1. The set of diagnostic indicators for smart growth pillars

SMART GROWTH		
Pillar I	Pillar II	Pillar III
SMART SPECIALIZATION	**CREATIVITY**	**INNOVATION**
KIS – employment in knowledge-intensive services as the share of total employment (%)	*TETR* – the share of tertiary education employment in total employment in a region (%)	R&De – research and development expenditure in enterprise sector (% of GDP)
HTMS – employment in high and medium high-technology manufacturing as the share of total employment (%)	*HRST* – human resources in science and technology as the percentage of active population (%)	R&Dgov – research and development expenditure in government sector (% of GDP)
	LLL – participation in education and training of population aged 25-64 (as the share of total population (%))	*EPO* - patent applications to the European Patent Office per million labour force

Source: Authors' compilation based on: European regional space classification in the perspective of smart growth concept – dynamic approach (grant NCN no. 2011/01/B/HS4/04743)

Smart specialization emphasises the real scope and role of the high and medium technology sector in the employment structure of individual countries. Currently, knowledge- and innovation-based economies, i.e. the ones where a large proportion of GDP and workplaces comes from these sectors, are considered to be capable of gaining a competitive advantage on an international scale, thus guaranteeing the availability of workplaces to individuals. For knowledge-intensive services (KIS) knowledge is the main production factor as well as the good that they offer. In line with the Eurostat methodology, services are mainly aggregated into knowledge-intensive services (KIS) and less knowledge-intensive services (LKIS) based on the share of tertiary educated persons at NACE 2-digit level. KIS covers such activity as:

- knowledge-intensive high-tech services (post and telecommunications; computer and related activities; research and development);
- knowledge-intensive market services (excluding financial intermediation and high-tech services) (water transport; air transport; real estate activities; renting of machinery and equipment without operator, and of personal and household goods; other business activities);
- knowledge-intensive financial services (financial intermediation, except insurance and pension funding; insurance and pension funding, except compulsory social security; activities auxiliary to financial intermediation);
- other knowledge-intensive services (education; health and social work; recreational, cultural and sporting activities).

The high and medium high-technology manufacturing (HMMS) refers to such groups of economic activity as:

- high technology (basic pharmaceutical product and pharmaceutical preparation; computer, electronic and optical products; air and spacecraft and related machinery);
- medium and high technology (chemicals and chemical products; weapons and ammunition; electrical equipment, machinery equipment, motor vehicles, trailer and other; medical and dental instruments and supplies).

Creativity is the aspect that focuses on the quality of human capital across countries, as well as readiness to improve qualifications. Human capital is approximated by three variables: human resources in science and technology (HRST) - citing *the Canberra Manual*, this refers to those individuals who fulfil one of the following conditions: (1) successfully completed education at the tertiary (third) level in an S&T field of studies, (2) did not formally qualify as above, but are employed in a S&T profession, where the above qualifications are normally required. This variable helps to better understand the demand for and supply of highly skilled, specialized staff in S&T. Highly skilled human resources are defined as essential to the diffusion of knowledge, and form the crucial link between technological progress and economic growth, social development and environmental well-being. The second variable underlines the general level of formal knowledge in the society expressed by percentage of people who successfully completed tertiary education, and the third variable describes the level of inclination toward life-long learning.

Innovation is the pillar that represents the amount of R&D funds invested in the region, taking into consideration the character of the investor (business and public sector), along with the results of innovation activities in the form of patent applications (*EPO*). The total European patent applications refer to requests made for protection of an invention forwarded either directly to the European Patent Office (EPO) or filed under the Patent Cooperation Treaty and designating the EPO (Euro-PCT), regardless of whether they are granted or not.

To obtain the comparability of data among countries and their economies all features were defined as indicators (in relation to other phenomena, e.g. population, employed).

Stage II. Constructing measures for smart growth, economic and social cohesion

This stage of analysis covers (Hellwig 1968; Walesiak 2006; Bal-Domańska, Wilk 2011):

A. Defining the character of a variable in terms of its connection to the described phenomena as: (S) *stimulant* – when the increase in a variable indicates an improved situation; (D) de*stimulant* – when the increase in the value is interpreted as deterioration in the situation. (N) *nominant* – when a

specified value is the only one to be regarded as having positive impact; the values below and above the nominal one have negative impact on the assessment of the situation. All variables applied to describe economic and social cohesion, as well as smart growth, were treated as stimulants.

Their higher values strengthen development processes.

B. Normalising diagnostic indicators by scaling between 0 and 1 in line with the following formula:

$$z_{itj} = \frac{x_{itj} - \min\limits_{i} x_{itj}}{\max\limits_{i} x_{itj} - \min\limits_{i} x_{itj}} \tag{1}$$

where:

z_{itj} – value of j-diagnostic feature (indicator, variable) (j = 1, 2,…, K) in i-th object (country) (i = 1, 2,…, N) in t-th period (t = 1, 2,…, T) after the normalization by scaling between 0 and 1,

x_{itj} – implementation of j-diagnostic feature in i-th object in t-th period,

$\min x_{itj}$ ($\max x_{itj}$) – the lowest (highest) value of j-diagnostic feature x_{itj}.

The standardisation was simultaneously performed for values of the variable referring to all countries and years, which allowed comparison of the country's position in consecutive years.

C. Calculating aggregate growth measure (AGM) for l-th pillar of smart growth (l = SS, C, I; SS – smart specialization; C- creativity; I – Innovation) by:

- defining the global benchmark of smart growth z_{0t} for T periods together for each variable,

$$z_{0t} = [z_{0t1}\, z_{0t2} \ldots z_{0tK}], \tag{2}$$

$$\text{such that:} \quad z_{0tj} = \max z_{itj}. \tag{3}$$

- calculating aggregate growth measure for each of the K_l sub-measures of smart growth l-th pillar:

$$AGM^{\,l}_{SMART_{it}} = \frac{1}{K_l} \sum_{j=1}^{K_l} z_{itj}, \tag{4}$$

Each of the values is normalised between 0 and 1, so that 1 is the most favourable value.

Stage III. – Models of social and economic cohesion

Linear econometric models describe relations which combine smart growth with economic and social cohesion by means of applying panel data in the EU countries, which is presented in the form of the following model constructions:

$$AM_{ECON,it} = (AGSS_{it}, AGC_{it}, AGI_{it}, \alpha_i, \alpha_t, \varepsilon_{it}), \tag{5}$$

$$AM_{SOC,it} = (AGSS_{it}, AGC_{it}, AGI_{it}, \alpha_i, \alpha_t, \varepsilon_{it}). \tag{6}$$

where:

$AM_{ECON,it}$ - aggregate measure for economic cohesion for *i*-th country in *t*-th year, which is *GDP* (Gross Domestic Product *per capita* in PPS),

$AM_{SOC,it}$ - aggregate growth measure for social cohesion for *i*-th country in *t*-th year, which represents *EM* (the employment rate among population aged 20-64 in %),

$AGSS_{it}$ ($AGM^{SS}_{SMART,it}$) - aggregate growth measure for *smart specialization* pillar of smart growth for *i*-th country in *t*-th year,

AGC_{it} ($AGM^{C}_{SMART,it}$) - aggregate growth measure for *creativity* pillar of smart growth for *i*-th country in *t*-th year,

AGI_{it} ($AGM^{I}_{SMART,it}$) - aggregate growth measure for *innovation* pillar of smart growth for *i*-th country in *t*-th year,

α_i - constant in time individual effects for *i*-th country,

α_t - different intercepts in each year common for all objects (countries),

ε - error term.

In the model both individual effects for each country α_i, and time for each year α_t, were included. Incorporating individual effects into the model structure made it possible to take into account characteristics which are specific for each country and constant in time (such as geographic location and accompanying resources). Time effects introduce an additional incidental parameter bias (Wooldridge, 2002).

In order to estimate the parameters, adequate estimation techniques, typical for panel data, were applied. LSDV (*Least Squares with Dummy Variable*) model was used in the study (Greene, 2003), (Wooldridge, 2002). To assess the validity of introducing the individual effects α_i to the model, F test was performed.

$$F = \frac{(\sum e^2_{OLS} - \sum e^2_{LSDV})/(N-1)}{(\sum e^2_{OLS})/(NT - N - K)} \tag{7}$$

where:

$\sum e_{OLS}^2 (\sum e_{LSDV}^2)$ - the sum of squared residuals in the LSDV (Least Square Dummy Variable) and OLS (Ordinary Least Square) regression.

It is the test of null hypothesis, i.e. all the units share the same intercept against the alternative that they are different from.

Wald's test (chi-square) was applied to assess the validity of introducing α_t time effects to the model.

In the process of estimating econometric models, certain problems, may occur, e.g. autocorrelation, heteroskedasticity. In order to minimize their possible negative effects, robust standard errors (Arellano, 2003) were used in assessing the significance of structural parameters evaluation.

All calculations were performed in GRETL.

3. Econometric analysis results

The analysis begins with the distribution of aggregate values of growth measures for particular pillars of smart growth (Figure 1), as well as of economic and social cohesion (Figure 2) for 27 EU countries, in the period of 2002-2011.

The levels of smart specialisation (*AMSS*) and innovation (*AMI*) in the studied countries do not change significantly in the analysed years. A significant increase in the aggregate measure of growth is observed for creativity (*AMC*).

Innovation occurs to be the most diverse variable pillar of smart growth (in terms of variation coefficient) in the cross-section of the EU countries, while smart specialisation is the least one. In the analysed time periods (years) the levelling off of creativity, and to a lesser extent innovation, can be observed.

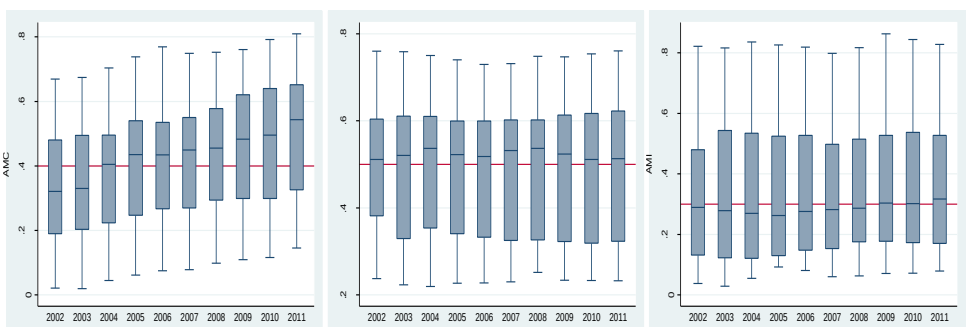


Figure 1. Values of aggregate growth measure of smart growth pillars for the EU countries in the period of 2002-2011

Source: Authors' work in STATA program.

Looking at the distribution of the values of economic cohesion (Figure 2) one can observe that *GDP* grows over the entire analysed period, except the years directly after the crisis (2008-2009). Attention should be paid to the level of *GDP* per capita for Luxemburg, which differs from other countries in each of the studied years (to be seen as outlier observations). In 2011 *GDP* per capita in PPS of Luxemburg was 68,100, in Netherlands – the second country in the range – 32,900, in Austria – 32,400 and in Ireland – 32,300, which is half of Luxemburg's *GDP* amount. The lowest *GDP* level was recorded in Romania and Bulgaria – about 11,700, a slightly higher one in Latvia – 14,700.

Within the analysed period, the processes of achieving economic cohesion are observed, which manifests itself in narrowing differences in the level of economy development among countries (measured as GDP per capita in PPS). These positive processes came to a halt in the years 2008-2011. However, disparities among countries in GDP per capita at the end of the analysed period are shown to be narrower than in the first year of the research.

The value of the employment rate (Figure 2) increased significantly (referring to the median and maximum value) during the period of 2004-2008. It can also be noticed that the minimum value of the indicator grows year on year, which seems to be a positive aspect, which indicates the increase of the employment rate even in the countries with the least favourable situation. In 2011, the highest employment rate was in Sweden (79%), Netherlands (77%), with values exceeding 75% also reported in Germany, Austria and Denmark. The lowest employment rate in 2011 (about 60%) was recorded in Greece, Hungary, Italy and Malta.

Until 2008, the processes leading to social cohesion among the EU countries were observed; it was manifested in decreasing disparities in employment levels among countries. However, in the years of the crisis and immediately after them the differences in employment levels were growing again.

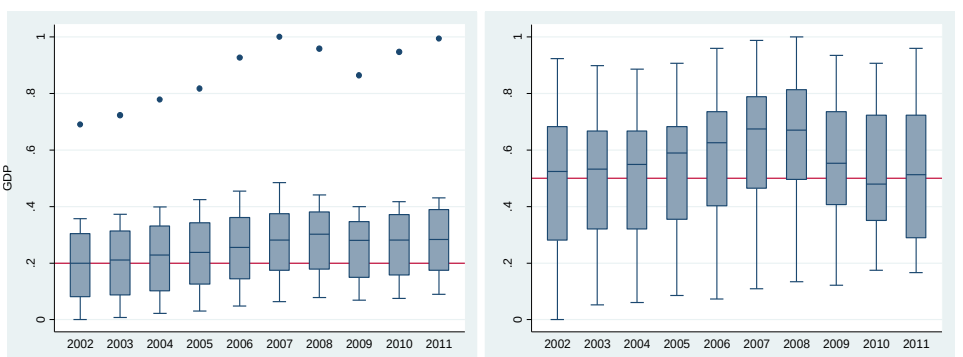


Figure 2. Values of economic and social cohesion indicators for the EU countries in the period of 2002-2011

Source: Authors' work in STATA program.

Out of the three smart growth pillars: creativity, innovation and smart specialization, only creativity could be identified as statistically significant (at the level of 0.1) in terms of its influence on economic cohesion (Table 2). This pillar represents the measure of the quality of the country's human capital, with special attention paid to the science and technology sector, the level of tertiary education and life-long learning. The increase in creativity level by 1 point was reflected in the growth of economic cohesion by 0.171 (ceteris paribus). The other pillars did not show any statistically significant relations. All time effects were statistically significant.

The values of F statistics amounting to 517 confirm that including α_i individual effects in the model is fully justified, since they improve estimation results as statistically significant. That means that major differences in economic cohesion between countries were observed. The value of determination coefficient informs that almost 98.8% of economic cohesion variability was explained by the model with dummy variable.

Table 2. The results of model estimations of economic cohesion and smart growth for 27 UE countries in the period of 2002-2011

Specification	$AM_{ECON,it} = (AGSS_{it}, AGC_{it}, AGI_{it}, \alpha_i, \alpha_t, \varepsilon_{it})$
AGC	**0.171* [0.037]**
AGSS	-
AGI	-
α_{2002}	0.156***
α_{2003}-α_{2002}	0.004***
α_{2004}-α_{2002}	0.015***
α_{2005}-α_{2002}	0.027***
α_{2006}-α_{2002}	0.048***
α_{2007}-α_{2002}	0.071***
α_{2008}-α_{2002}	0.069***
α_{2009}-α_{2002}	0.037***
α_{2010}-α_{2002}	0.051***
α_{2011}-α_{2002}	0.059***
R^2	0.988
Test F (*p-value*)	516.9 (0.000)
The Akaike information criterion	-1330.8

*** significant at the level of 0.001, ** significant at the level of 0.05, * significant at the level of 0.1. Arellano robust standard error HAC is quoted in parentheses [].

Source: Authors' estimations in GRETL programme.

The attempt to describe (by applying econometric models) the relationships between smart growth and social cohesion expressed in terms of employment rates proved to be a considerable challenge.

The main reson for this is the diverse nature of growth processes in each of the countries, particularly in the years after the crisis of 2008. Consequently, the attempt to apply the pillars concept in order to describe social cohesion failed. Figure 3 presents the changes of the employment rate $AGM_{EMPL,it}$.

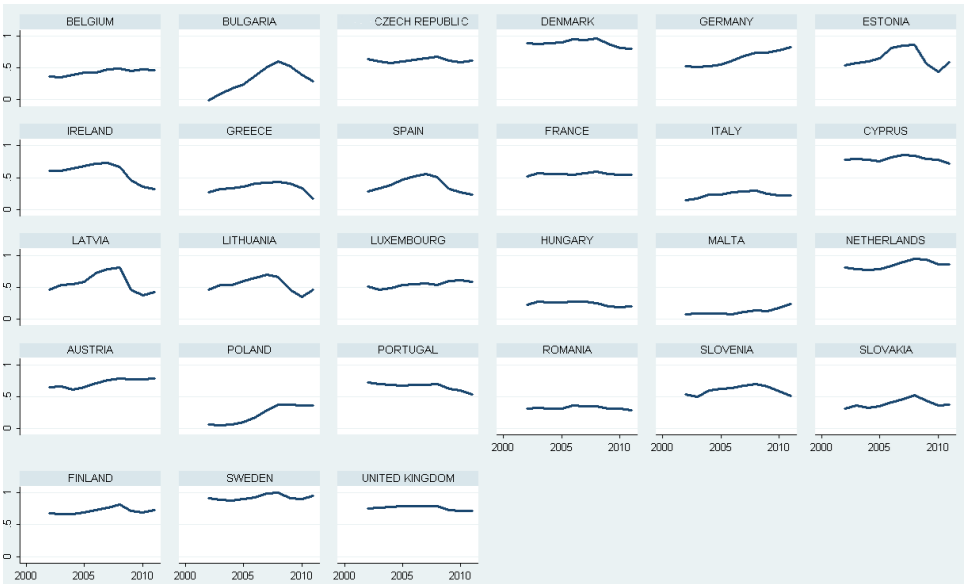


Figure 3. Values of the employment rate (*EM*) as a social cohesion measure of the EU countries in the period of 2002-2011

Source: Authors' work in STATA program.

As can be seen, the run (distribution) of indicators differed among the studied countries in the period of 2002-2011. Taking into account the values of the employment rate, three main types of run can be identified:

- **increase** - this tendency was true for the employment rate in 5 countries: Austria, Poland, Germany, Malta and Belgium.
- **hill** - until 2008 an increase in the indicator was observed (sometimes very explicit, e.g. in Spain, Estonia, Bulgaria, Latvia, Lithuania, Slovakia, Ireland and Greece). Later a significant decline was observed.
- the third type referrs to the absence of changes (**stable**) - in that case changes are irrelevant and oscillate around a particular level. 10 such countries were identified.

It is an approximate division.

The situation was different during the analysis of smart growth pillars. In terms of creativity an increase was observed for the majority of countries. Only in few of them the changes smaller than 10% of *AGC* were recorded.

The level of innovativeness was constant, or increased in most countries. A decrease of over 10% of *AGI* was observed in the United Kingdom, Hungary, Cyprus and Bulgaria.

Looking at the smart specialization factor the situation improved in 7 countries (Czech Republic, Greece, Cyprus, Luxemburg, Portugal, Slovenia and Slovakia), whereas in another group of 7 countries (Denmark, Estonia, Ireland, Malta, Romania, Sweeden and United Kingdom) a decline in the value of *AGSS* was observed in the last assessment period compared to the initial one. In the remaining countries the value of *AGSS* remained at a relatively constant level.

The models for clusters of countries analysed in terms of the employment rate and smart growth pillars allowed for the identification of the following statistically significant relations (Table 3).

Table 3. The results of model estimations for the employment rate and smart growth pillars regarding clusters of the EU countries in the period of 2002-2011

Specification	*Increase*	*Hill*	*Stable*
AGC	-	-1.212** [0.594]	0.386*** [0.060]
AGSS	0.791***[0.285]	-	-0.358** [0.147]
AGI	-	-	-
α_{2002}	-0.1099	0.8835***	0.4280***
α_{2003}-α_{2002}	0.0039***	0.0155***	0.0075
α_{2004}-α_{2002}	0.0093	0.0361***	0.0137
α_{2005}-α_{2002}	0.0069***	0.0570***	0.0120
α_{2006}-α_{2002}	0.0157***	0.0656***	0.0142
α_{2007}-α_{2002}	0.0192***	0.0789***	0.0112***
α_{2008}-α_{2002}	0.0334***	0.0928***	0.0178*
α_{2009}-α_{2002}	0.0376***	0.1050**	0.0229
α_{2010}-α_{2002}	0.0331***	0.1079	0.0191**
α_{2011}-α_{2002}	0.0407***	0.1208	0.0199**
R^2	0.977	0.899	0.989
Test F (*p-value*)	277.10 (0.000)	44.5 (0.000)	275.4 (0.000)
The Akaike information criterion	-155.8	-275.9	-416.3

Designation as in Table 2.

Source: Authors' estimations in GRETL program.

For the "increase" class, a statistically significant relation (at the level of 0.001) related to smart specialization pillar was identified. A significance increase in employment in technology and knowledge-intensive sectors by unit was related to the increase in total employment rate by 0.791 (ceteris paribus).

In the case of the "hill" class, the relation between countries and creativity was negative, which suggests that despite the increase in the creativity level (observed for the majority of countries) the employment rate declined. It was influenced by other factors not included in the model. The employment rate did not depend on the level of innovativeness and smart specialization in a given country. The absence of statistically significant time effects for the years 2010-2011 indicates the trend breakdown regarding the employment rate in the period of crisis.

The role of employment in technology and knowledge-intensive sectors had a negative effect on the total employment rate in the "stable" class. Expanding the role of employment in the medium and high-tech manufacturing sector and, at the same time, the knowledge-intensive sector by unit reduces the employment rate by 0.358 (*ceteris paribus*). The negative sign of the parameter estimate indicates that changes in the employment rate resulted in changes in the employment structure in sectors other than knowledge. At the same time changes in the level of creativity were consistent with changes in the employment rate of 0.386 (*ceteris paribus*).

4. Conclusions

As a result of the research conducted by applying econometric tools the following conclusions for the EU regions in the period of 2002-2011 were drawn:

- A statistically significant relationship between the level of economic cohesion and the creativity level of the EU countries was confirmed. Enhancing human capital potentially favours a higher level of economic cohesion.
- It was not possible to identify (at a country level) statistically significant relationships for the two remaining pillars of smart growth: smart specialization and innovation.
- It was also not possible to identify any statistically significant connections between smart growth and social cohesion (employment). This might be due to the diverse and complex nature of links connecting these phenomena among the EU countries in the studied years.
- Within the clusters of countries, specified in terms of the employment rate, statistically significant relationships were identified for the chosen smart growth pillars. An increase in the employment rate (in the "increase clusters") was related to the increasing role of employment in smart specialization sectors. Simultaneously, the countries from this cluster demonstrated the highest resilience against the consequences of the crisis manifested in the form of a decline in the employment rate.

REFERENCES

A strategy for smart, sustainable and inclusive growth, (2010). European Commission. Communication from the Commission EUROPE 2020, Brussels, 3.3.2010.

ARELLANO, M., (2003). Panel Data Econometrics, Oxford: Oxford University Press 2003.

BERGER-SCHMITT, R., (2000). Social cohesion as an aspect of the quality of societies: concept and measurement. Centre for Survey Research and Methodology Mannheim, EuReporting Working Paper, No. 14/2000.

CICHY, K., (2008). Kapitał ludzki i postęp techniczny jako determinanty wzrostu gospodarczego [Human capital and technological progress as the determinants of economic growth], Instytut Wiedzy i Innowacji, Warszawa.

DOMINIAK, J., CHURSKI, P., (2012). Rola innowacji w kształtowaniu regionów wzrostu i stagnacji w Polsce [The role of innovation in shaping the regions of growth and stagnation in Poland], Studia Regionalne i Lokalne, No.4(50)/2012, pp. 54–77.

FIEDOR, B., (2010). Pomoc zewnętrzna i endogenizacja wzrostu a polityka spójności – ze szczególnym uwzględnieniem Unii Europejskiej, Kilka refleksji [External aid and endogenisation of growth, and cohesion policy - with focus on the European Union, Some reflections] [in:] M. Klamut. E. Szostak. Spójność w rozwoju regionalnym w Polsce obecnie i w przyszłości [Cohesion in regional development in Poland at present and in the future], Wrocław University of Economics Publishing House. Wrocław, pp. 11–23.

GORYNIA, M., JANKOWSKA, B., (2008). Klastry a międzynarodowa konkurencyjność i internacjonalizacja przedsiębiorstw [Clusters and international competitiveness and internationalization of enterprises], Centrum Doradztwa I Informacji. Difin, Warszawa.

GREENE, W. H., (2003). Econometric analysis. Pearson Education International. New Jersey.

HELLWIG, Z., (1968). Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr [The application of taxonomic method for typological division of countries regarding their development level as well as the resources and structure of qualified personnel]. "Przegląd Statystyczny" [“Statistical Review”] 1968 Bulletin, No. 4, pp. 307–327.

HERBST, M., edit., (2007). Kapitał ludzki i kapitał społeczny a rozwój regionalny [Human capital and social capital vs. regional development], SCHOLAR, Warsaw.

MARKOWSKA, M., STRAHL, D., (2013). Multicriteria European regional space classification regarding economic and social cohesion and smart growth level [Klasyfikacja wielokryterialna europejskiej przestrzeni regionalnej uwzględniająca spójność ekonomiczną i społeczną oraz rozwój inteligentny], The 7th Professor A. Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena May 7-10, Zakopane.

MARSHALL, A., (1925). Zasady ekonomiki [Principles of Economics], Warsaw, M. Arct, polski przekład publikacji z roku 1890 [Polish translation of the publication from 1890].

KAWA, P., (2007). Rola wiedzy i innowacji w stymulowaniu wzrostu gospodarczego [The role of knowledge and innovation in stimulating economic growth], [in:] K. Piech. E. Skrzypek (eds) „Wiedza w gospodarce. społeczeństwie i przedsiębiorstwach: pomiary, charakterystyka, zarządzanie" ["Knowledge in the economy, society and enterprises: measurement, characteristics, management], Instytut Wiedzy i Innowacji, Warszawa, pp. 16–29.

OECD, (2011). Perspectives on Global Development 2012: Social Cohesion in a Shifting World. OECD Publishing. Paris 2011, DOI: http://dx.doi.org/10.1787/persp_glob_dev-2012-en

PORTER, M. E., (1990). The Competitive Advantage of Nations. Harvard Business Review.

ROMER, P., (1990). Endogenous technological change", Journal of political Economy", No 5, pp. 71–102.

ROMER, P., (1986). Increasing returns and long-run growth, "Journal of political Economy", October 1986, pp. 1002–1037.

STANLEY, D., (2003). What Do We Know about Social Cohesion: The Research Perspective of the Federal Government's Social Cohesion Research Network. The Canadian Journal of Sociology/Cahiers canadiens de sociologie, Vol. 28, No. 1, Special Issue on Social Cohesion in Canada (Winter 2003), pp. 5–17.

STRAHL, D., (2010). Innowacyjność europejskiej przestrzeni regionalnej a dynamika rozwoju gospodarczego [Innovation in the European regional area and the dynamics of economic development], Uniwersytet Ekonomiczny, Wrocław.

WALESIAK, M., (2006). Uogólniona miara odległości w statystycznej analizie wielowymiarowej [Generalised distance measure in statistical multivariate analysis], Wrocław University of Economics Publishing House, Wrocław.

WOOLDRIDGE, J. M., (2002). Econometric analyses of cross section and panel data, Massachusetts Institute of Technology.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 265–280

HETEROSCEDASTIC DISCRIMINANT ANALYSIS COMBINED WITH FEATURE SELECTION FOR CREDIT SCORING

Katarzyna Stąpor[1], **Tomasz Smolarczyk**[2], **Piotr Fabian**[3]

ABSTRACT

Credit granting is a fundamental question and one of the most complex tasks that every credit institution is faced with. Typically, credit scoring databases are often large and characterized by redundant and irrelevant features. An effective classification model will objectively help managers instead of intuitive experience. This study proposes an approach for building a credit scoring model based on the combination of heteroscedastic extension (Loog, Duin, 2002) of classical Fisher Linear Discriminant Analysis (Fisher, 1936, Krzyśko, 1990) and a feature selection algorithm that retains sufficient information for classification purpose. We have tested five feature subset selection algorithms: two filters and three wrappers. To evaluate the accuracy of the proposed credit scoring model and to compare it with the existing approaches we have used the German credit data set from the study (Chen, Li, 2010). The results of our study suggest that the proposed hybrid approach is an effective and promising method for building credit scoring models.

Key words: heteroscedastic discriminant analysis, feature subset selection, variable importance, credit scoring model.

1. Introduction

Credit scoring models are the basis for financial institutions like retail and consumer credit banks. The purpose of these models is to evaluate the likelihood of credit defaulting applicants in order to decide whether to grant them credit. The set of decision models and their underlying methods that serve lenders in granting consumer credits are called credit scoring (**CS**) (Zhang et al. 2010).

[1] Institute of Computer Science, Silesian University of Technology, Gliwice, Poland. E-mail: katarzyna.stapor@polsl.pl.
[2] Institute of Computer Science, Silesian University of Technology, Gliwice, Poland. E-mail: tomasmo356@student.polsl.pl.
[3] Institute of Computer Science, Silesian University of Technology, Gliwice, Poland. E-mail: piotr.fabian@polsl.pl.

Since customer demand for personal loans has increased in the last decades, the consumer credit market evolved to become an important sector in the financial field and today represents a high-volume business. These developments in the retail credit market requires automatic, fast and consistent decisions and processes to handle the huge amount of applications. The use of credit scoring models is now a key component in retail banking. The development of the so-called scorecards therefore represents the core competence of a retail bank's risk management when assessing the creditworthiness of an individual. Since the market is changing rapidly, new statistical and mathematical methods are required to optimize the scoring problem to decide on the question of whom to offer credit to.

Discriminant analysis, linear regression, logistic regression, neural networks, k-nearest neighbours, support vector machines and classification trees cover the range of different surveys on CS models (Thomas et al., 2005). An overview of publications is given in Thomas (2000) and Crook et al. (2007).

Many credit scoring models have been widely developed by reducing redundant features through feature selection to improve the accuracy of credit scoring models during the past few years. The detailed survey of the existing methods for feature selection is given in (Dash, 1997), for example. A feature subset selection algorithm can be divided into two categories: the filter approach and the wrapper approach (Dash, 1997). The filter relies on various measures like distance, information, dependency on feature evaluation which are then used for their ranking. The wrapper model usually uses the predictive accuracy of the pre-determined learning algorithm to determine the goodness of the selected feature subsets.

The use of feature selection in the construction of credit scoring models has already been reported, for example in (Chen, Li, 2010, Somol, 2005), but there are no references on the selection of variables for their use in discriminant analysis in building credit scoring models. In (Chen, Li, 2010), classical Fisher Discriminant Analysis (**FDA**) (Fisher, 1936, Fukunaga, 1990, Krzyśko, 1990) is used, but with all the input features to generate discriminators for their use in SVM classifier. No feature selection is applied to this model.

This work proposes a new method for constructing a credit scoring model which is based on the feature selection in Heteroscedastic Discriminant Analysis (**HDA**) (Loog, Duin, 2002). HDA is the extension of FDA for dealing with the case of unequal covariance matrices in populations, the situation that occurs very often in practice and in our experiments, too.

In our experiments, for the evaluation of the accuracy of our proposed credit scoring model, we have used the German credit data set, the same that was used in the (Chen, Li, 2010) study. Using classical FDA for feature extraction, we have obtained very poor results (prediction accuracy defined as the number of correct classifications divided by the total number of classifications was about 30%), suggesting that probably the covariance matrices in the two classes are not equal. This was the main reason for the usage of heteroscedastic extension of FDA in

our proposed model. Using HDA as feature extraction combined with the input feature subset selection causes the prediction accuracy to improve up to 76%. This proves that the proposed model is better fitted to the data.

The prediction accuracy of the credit scoring model based on FDA from study (Chen, Li, 2010) is the same as in our case (i.e. 75%). However, this accuracy was achieved in (Chen, Li, 2010) by using nonlinear SVM classifier (with Gaussian kernel), making the learning process more complex – one should estimate the parameters of the SVM classifier in the separate validation procedure which requires the additional data set and is a computationally intensive process (the grid method). Moreover, their credit scoring model uses all the input variables which makes its usage less economic and less intuitive for the interpretation. Additionally, the necessity of specifying the values of the parameters of the SVM classifier in the separate validation procedure will cause worse generalizability of the model.

Thus, our proposed credit scoring model, by using feature selection, proper model for feature extraction as well as the simpler classifier, does not have the above mentioned disadvantages of the model from (Chen, Li, 2010) study.

The valuable step in our proposed credit scoring model is the variable importance analysis, a very useful process of analysing attributes in the context of their significance in the discrimination of good and bad credit consumers.

This paper is organized as follows. Section 2 and 3 shortly present the heteroscedastic extension of the classical FDA which is based on the notion of distance directed matrices (Loog, Duin, 2002) and the feature subset selection algorithms used in the construction of our credit scoring model, respectively. Section 4 describes the proposed methodology for building credit scoring models, while section 5 – experimental results together with the variable importance analysis. Section 6 presents the conclusions and suggestions for future research and practice of our new credit scoring model based on the heteroscedastic extension of FDA.

2. Two-class Heteroscedastic Disriminant Analysis

Fisher Discriminant Analysis (FDA) (Fisher 1936; Krzyśko 1990; Fukunaga 1990) is a multivariate technique to classify study instances into groups and/or describe group differences. Discriminant analysis is widely used in many areas such as biomedical studies, banking environment (for credit evaluation), financial management, bankruptcy prediction, marketing, and many others.

There are many formulations of FDA, a typical one for pattern recognition community is given below (according to (Fukunaga, 1990)).

FDA is concerned with the search for a linear transformation that reduces the dimension of a given n-dimensional statistical model to d ($d<n$) dimensions, while maximally preserving the discriminatory information for the several classes

within the model. It determines a linear mapping *A*, a $d \times n$ matrix *A*, that maximizes the so-called Fisher criterion J_F:

$$J_F(A) = tr\left(\left(AS_W A^T\right)^{-1}\left(AS_B A^T\right)\right) \tag{1}$$

Here, $S_B = \sum_{i=1}^{c} \frac{n_i}{n}\left(m_i - \overline{m}\right)\left(m_i - \overline{m}\right)^T$ and $S_W = \sum_{i=1}^{c} \frac{n_i}{n} S_i$ are the between-class and the average within-class scatter matrices, respectively; *c* is the number of classes, m_i is the mean vector of class *i*, n_i is a number of samples in class *i*, $n = \sum_{i=1}^{c} n_i$, and the estimated overall mean equals $\overline{m} = \sum_{i=1}^{c} \frac{n_i}{n} m_i$, $S_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(X_{ij} - m_i\right)\left(X_{ij} - m_i\right)^T$ is the within-class covariance matrix of class *i*.

Optimizing (1) comes down to determining an eigenvalue decomposition of $S_W^{-1}S_B$, and taking the rows of *A* equal to *d* eigenvectors corresponding to *d* largest eigenvalues (Fukunaga, 1990).

For the two-class case we have:

$$S_B = (m_1 - m_2)(m_1 - m_2)^T \text{ and } S_W = p_1 S_1 + p_2 S_2,\ p_2 = 1 - p_1.$$

A limitation of FDA is that it merely tries to separate class means as good as possible and it does not take the discriminatory information, which is present in the difference of the covariance matrices, into account. It is incapable of dealing explicitly with heteroscedastic data, i.e., data in which classes do not have equal covariance matrices.

For building our credit scoring model we have used one of the existing heteroscedastic generalizations of the Fisher criterion (1), namely that based on the Chernoff criterion (Loog, Duin, 2002). The heteroscedastic extension in (Loog, Duin, 2002) is based on the notion of Distance Directed Matrices (DDM) which capture not only the difference in means between two classes, but also describe their difference in covariance in a certain way. (Loog, Duin, 2002) proposed DDM based on the Chernoff distance between two probability density functions d_1, d_2:

$$\partial_C = -\log \int d_1^{\alpha}(x) d_2^{1-\alpha}(x) dx \tag{2}$$

where $\alpha \in (0,1)$.

Another interesting approach to heteroscedastic linear discriminant analysis can be found in (Krzyśko, Wołyński, 1996), where authors proposed the optimal classification rules based on linear functions which maximize probabilistic distances: the Chernoff or the Morisita or the Kullback-Leibler ones.

For two normally distributed densities, the DDM is a positive semi-definite matrix S_C:

$$S_C = S^{-\frac{1}{2}}(m_1 - m_2)(m_1 - m_2)^T S^{-\frac{1}{2}} + \frac{1}{p_1 p_2}(\log S - p_1 \log S_1 - p_2 \log S_2) \qquad (3)$$

where $\alpha = p_1$, $S = p_1 S_1 + p_2 S_2$. The trace of S_C is the Chernoff distance ∂_C between those two densities. Determining transformation *A* by an eigenvalue decomposition of S_C means that we determine a transform which preserves as much of the Chernoff distance in the lower dimensional space as possible. The heteroscedastic two-class Chernoff criterion J_C is defined as:

$$J_C(A) = tr\left(\left(AS_W A^T\right)^{-1} A(m_1 - m_2)(m_1 - m_2)^T A^T \right.$$
$$\left. - AS_W^{\frac{1}{2}} \frac{p_1 \log\left(S_W^{-\frac{1}{2}} S_1 S_W^{-\frac{1}{2}}\right) + p_2 \log\left(S_W^{-\frac{1}{2}} S_2 S_W^{-\frac{1}{2}}\right)}{p_1 p_2} S_W^{\frac{1}{2}} A^T \right) \qquad (4)$$

This is maximized by determining an eigenvalue decomposition of:

$$S_W^{-1}\left(S_B - S_W^{\frac{1}{2}} \frac{p_1 \log\left(S_W^{-\frac{1}{2}} S_1 S_W^{-\frac{1}{2}}\right) + p_2 \log\left(S_W^{-\frac{1}{2}} S_2 S_W^{-\frac{1}{2}}\right)}{p_1 p_2} S_W^{\frac{1}{2}} \right) \qquad (5)$$

and taking the rows of the transform *A* equal to *d* eigenvectors corresponding to the *d* largest eigenvalues.

3. Feature subset selection methods

In the proposed methodology for building credit scoring models, we have used five feature selection methods. Three of them are wrapper-based and use different search strategies for finding a suboptimal set of features: the Sequential Floating Forward Search (SFFS) method (Pudil, et al., 1994), the method using Memetic Algorithms (MA) (Moscato, 2002) and the method that utilizes the Greedy Randomised Adaptive Search Procedure (GRASP) (Feo, Resende, 1989). The two filter-based methods use different techniques for scoring individual features which are then used for their ranking and selecting the top best features: Correlation-based Filter Selection (CFS) (Hall, 1997) and Fisher Score (FS) (Duda, Hart, Stork, 2001).

3.1. SFFS

SFFS is an enhanced version of the Sequential Forward Search (SFS) algorithm (Pudil, et al. 1994). Besides adding the most significant feature in each step, SFFS searches for the least significant feature in the current subset and checks whether removing it will result in the increased performance of the classifier. If so – the feature is removed and the algorithm repeats the procedure of searching and removing unnecessary features. The stopping criterion is the number of added features that did not increase the performance (set to 2 in this research).

3.2. GRASP

GRASP constructs solutions – feature subsets - based on the greedy algorithm and the controlled randomization. It starts with an empty initial solution and in each iteration a list of candidate variables with the best performance is generated from which the algorithm selects at random one variable and add it to the current solution. The level of randomization is controlled by the α parameter $(0 \le \alpha \le 1)$. Each solution is improved by a simple local search procedure in which the current solution is replaced by a better properly defined neighbouring solution. The stopping criterion is defined as the maximum number of iterations – 30 in our study, and parameter α was set to 0.8.

3.3. MA

MA (sometimes called hybrid Genetic Algorithms) are a class of stochastic global search heuristics in which evolutionary algorithm-based approaches (Goldberg, 1989) are combined with problem-specific solvers. The later might be implemented, for example, as a local search heuristics techniques. The hybridization is meant to either accelerate the discovery of good solutions, or to reach solutions that would otherwise be unreachable by evolution or a local method alone. A single solution (a chromosome) is the vector with length equal to the number of all features composed of zeros and ones (zero means that the feature is not present in the subset). During selection phase, 50% of the best chromosomes are selected for later breeding. Population size was set to 30, crossover rate to 0.05, mutation rate to 0.05, and the number of iterations – 20 in our case – was the stopping criteria.

3.4. CFS

In CFS the goodness of a given feature is measured by the degree of association between a feature and a class, and is estimated based on the information theory (Cover, Thomas, 1991) as:

$$\text{symmetrical uncertainty} = 2.0 \times \left[\frac{H(Y) + H(X) - H(X,Y)}{H(Y) + H(X)} \right] \tag{6}$$

where $H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y)$ is the entropy of Y before observing X and $H(Y \mid X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x) \log_2 p(y \mid x)$ is the entropy of Y after observing X (X and Y are discrete random variables). Necessary preprocessing was accomplished as in (Hall, Smith, 1997).

3.5. FS

The FS is a measure of how a given feature is efficient for discrimination. It is defined by between-class and within-class scatter matrices S_B and S_W:

$$FisherScore = \frac{|S_B|}{|S_W|} \tag{7}$$

where |.| is a determinant. The larger the FisherScore value the more likely for the feature to be discriminative.

4. The proposed methodology for building Credit Scoring model

Figure 1 presents the proposed methodology for building the CS model which is then evaluated in this research. Feature selection is conducted as the wrapper or filter-based approach. Then, based on the selected features, the extraction methods are applied: the classical Fisher Discriminant Analysis (FDA) and heteroscedastic extension of FDA – i.e. FDA with Chernoff Criterion (FDA_Cher). For the classification of the samples in the new discriminant space (i.e. the space spanned by the extracted features - eigenvectors), the Fisher classifier is used, which is the nearest centroid method (Duda, Hart, Stork, 2011, Stąpor, 2011) but in the new discriminant space. The prediction accuracy is calculated on the test data as the ratio of correct predictions to the number of all test cases (Duda, Hart, Stork, 2011).

For the filter-based methods, the importance of each feature is evaluated individually for each feature by determining the value of the criterion function (which is specific to a particular filter – see formulas (6) and (7)). Features are then ranked in order of descending values of this criterion function. The number of the top best features is the one which gives the best classification performance. In the wrapper approach, variable importance is calculated as the frequency of the selection of each feature in 10 iterations (using FDA_Cher variant of feature extraction).

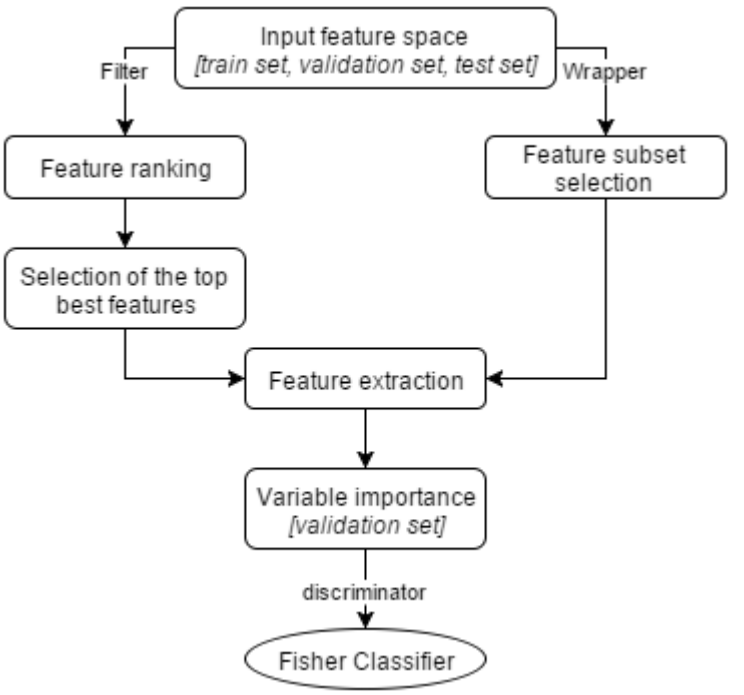


Figure 1. The methodology for building CS model (for learning stage)

5. Experimental analysis

5.1. Data set description

For the evaluation of the prediction accuracy of our proposed CS model we have used the real-world data set, the German credit data set which was also used in Chen and Li research published in Expert Systems with Applications (Chen, Li, 2010). The German data set consists of 700 instances of creditworthy borrowers and 300 of bad borrowers. It is composed of 20 numeric and nominal features containing information about credit duration, history, purpose, amount, savings, age, job and other personal information (a detailed structure of this data set is given in the Appendix 1). Nominal features were replaced by binary features, each one representing one of its possible states. Preprocessed data set contained 59 attributes.

5.2. Experimental results

The obtained results are summarized in Table 1. For the German data set, the classification accuracy of the two extraction methods (with all 59 features) achieved 30.00% ± 0.00% (FDA), 59.40% ± 10.43% (FDA_Cher) for extraction on one directions and 59.70% ± 11.18% for extraction on 3 dimensions. For FDA

extraction, SFFS feature selection was the best approach with average 58.50% ± 3.06% classification accuracy, and the median selected attributes were equal to 2. The FDA_Cher extraction achieved significantly better results. The best feature selection algorithm (Fisher Score) was able to achieve 75.10% ± 3.38% accuracy rate with 18 attributes selected and 3 directions.

Table 1. Results summary with 10-fold cross validation for German data set

Algorithm \ data set	FDA		
	Accuracy rate (%)		Number of selected features
	Avg.	Std.	Median
All features	30.00%	0.00%	59
CFS	34.00%	12.65%	1
FS	55.50%	18.77%	23
SFFS	58.50%	3.06%	2
GRASP	57.90%	7.37%	2
MA	30.00%	0.00%	27

Algorithm \ data set	FDA_Cher				FDA_Cher (1 direction)		
	Accuracy rate (%)		Number of selected features	Number of directions	Accuracy rate (%)		Number of selected features
	Avg.	Std.	Median		Avg.	Std.	Median
All features	59.70%	11.18%	59	3	59.40%	10.43%	59
CFS	73.90%	4.95%	28	2	73.10%	5.13%	24
FS	75.10%	3.38%	18	3	74.60%	2.99%	18
SFFS	67.00%	6.04%	6	3	66.60%	5.23%	6
GRASP	67.90%	5.43%	17	3	65.70%	6.06%	12
MA	65.80%	8.68%	28	3	61.20%	22.01%	26

Using FDA_Cher model, which does not require the homoscedasticity of the data, increased the average accuracy of prediction. The feature selection algorithm helped to decrease the number of features taken into the model and in some cases significantly increased the accuracy.

In a two-class problem, FDA reduces dimensionality to one direction, since only one eigenvalue is different than 0 and there is no discriminatory information in other directions. In heteroscedastic extension of FDA, DDMs have more than one nonzero eigenvalue. Those extra directions capture, in general, the heteroscedasticity in the data. In our research, we have examined between 1 to 5 dimensions/directions to which features could be extracted by FDA_Cher. Table 1 in the "Number of directions" column presents the best results and dimensions for which this result is obtained. In all cases, the best result was achieved in more than one dimension, which demonstrates that higher dimensions also contain discriminatory information that was not captured in the first direction.

5.3. Attribute importance analysis

In credit scoring, it is very important to know which attributes (features) characterizing a consumer introduced to the model are relevant, i.e. more significant in the classification task, and which are of less importance. Generally, variable importance measures can be divided into two groups: those that use the model information and those that do not. Our proposed method for the analysis of importance (i.e. the classification effectiveness) of the attributes belongs to the second group.

Table 2. Feature rankings

	Top 10																Last 10			
CFS	11	15	29	50	42	16	18	48	28	59	…	46	31	58	12	9	45	2	25	1	8
FS	11	8	1	15	25	2	50	12	29	45	…	54	14	44	32	19	43	52	7	4	39
Importance	Most important ⟷ Least important																				

Table 2 shows the results of filter-based variable importance analysis: rankings created by CFS and FS measures on the entire data set. Both algorithms selected as the most important feature 11 indicate that the customer does not have a checking account. Attribute 8, the status of the existing checking account, was selected by FS as the second most important attribute. However, CFS moved this attribute to the last place. Delay in paying off in the past is the second most important feature (attribute 15) for CFS and fourth for FS. For FS, the third most important feature was duration in month (attribute 1), which was ranked 58th for CFS.

Table 3. Feature selection frequency – attribute number (frequency [%])

	Top 7								Last 3		
SFFS	1 (100%)	8 (80%)	11 (60%)	25 (50%)	50 (50%)	15 (40%)	9 (30%)	…	57 (0%)	58 (0%)	59 (0%)
GRASP	12 (70%)	8 (60%)	35 (60%)	24 (50%)	40 (50%)	3 (40%)	5 (40%)	…	36 (0%)	51 (0%)	57 (0%)
MA	12 (80%)	13 (80%)	52 (80%)	8 (70%)	21 (70%)	44 (70%)	1 (60%)	…	46 (20%)	58 (20%)	7 (10%)
Importance	Most important							↔	Least important		

Table 3 shows the results of wrapper-based variable importance analysis (in percentage). The more frequently the feature was selected, the better evaluation it got during the feature selection step. Attribute 1 (duration in month) was selected in each case by SFFS algorithm and in 60% of the cases in MA. The most frequently selected attribute by GRASP and MA was attribute 12 describing a customer's credit history (no credits taken/all credits paid back duly). Attribute 8 was selected in 80% by SFFS, 70% by MA and 60% by GRASP, and it was also highly ranked by Fisher Score. Attribute 11 (the most important for filter features) was selected in 60% of cases by SFFS, 40% by MA and 10% by GRASP. In 10 iterations, SFFS algorithm was selected only from a subset of 14 attributes. One of the least frequently selected attribute was attribute 7 – the number of people being liable to provide maintenance for (0% by SFFS, 10% by GRASP, 10% by MA).

6. Conclusions

This work proposes a new method for constructing credit scoring models which is based on the feature selection in Heteroscedastic Discriminant Analysis, which is the extension of the classical linear Fisher Discriminant Analysis for dealing with the case of unequal covariance matrices in populations.

The prediction accuracy of our proposed credit scoring model is the same as in the best models currently proposed in the literature, but this accuracy is achieved using a linear (i.e. simpler) model, which implies better generalization properties.

We have proved that using heteroscedastic extension of the classical linear Fisher Discriminant Analysis results in better prediction accuracy. Moreover, this accuracy can be further improved by feature selection algorithms.

Not all information stored in the databases is relevant to predict customer behaviour and feature selection methods together with the feature extraction are crucial in reducing the dimensionality of the feature space, which is important from computational and economical point of view as well as because of the curse of dimensionality phenomenon.

Furthermore, thanks to the applied variable importance analysis, we can specify the most relevant variables for the classification task, which could be useful for the analysis of a given customer and for a better understanding of the credit scoring problem.

REFERENCES

CHEN, F., LI, F., (2010). Combination of feature selection approaches with SVM in credit scoring, Expert Systems with Applications, Vol. 37, pp. 4902–4909.

COVER, T., THOMAS, J., (1991). Elements of information theory. John Wiley & Sons, New York, NY.

CROOK, J. N., EDELMAN, D. B., THOMAS, L. C., (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research 183 (3), pp. 1447–1465.

DASH, M., LIU, H., (1997). Feature selection for classification. Intelligent Data Analysis, 1, pp. 131–156.

DUDA, R., HART, P., STORK, D., (2001). Pattern Classification. John Wiley & Sons, New York, 2 ed.

FEO, T. A., RESENDE, M. G. C., (1995). Greedy randomized adaptive search procedures. J. Global Optim. 2, pp. 1–27

FISHER, R. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, pp. 179–188.

FUKUNAGA, K., (1990). Introduction to statistical pattern recognition. New York: Academic Press.

GOLDBERG, D., (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley Professional.

HALL, M., SMITH, L., (1997). Feature subset selection: a correlation based filter approach, in International Conference on Neural Information Processing and Intelligent Information Systems, Berlin.

KRZYŚKO, M., (1990). Discriminant analysis, WNT, Warszawa (in Polish).

KRZYŚKO, M., WOŁYŃSKI, W., (1996). Discriminant rules based on distances, Tatra Mountains Math. Publ. 7(1996), pp. 289–296.

LOOG, M., DUIN, R., (2002). Non-iterative heteroscedastic linear dimension reduction for two-class data: from Fisher to Chernoff. Proc. 4th Int. Workshop S+SSPR, pp. 508–517.

MATUSZCZYK, A., (2012). Credit scoring. Warszawa: CeDeWu Sp. z o.o.

MOSCATO, P., (2002). Memetic algorithms. In Pardalos, P.M., Resende, M. (eds.): Handbook of Applied Optimization. Oxford: Oxford University Press, pp. 157–167.

PACHECO, J., et al., (2006). Analysis of new variable selection methods in discriminant analysis, Computational Statistics & Data Analysis, Vol. 51, 3, pp. 1463–1478.

PUDIL, P., et al., (1994). Floating search methods in feature selection, Pattern Recognition Letters, Vol. 15, 11, pp. 1119–1125.

SOMOL, P., et al., (2005). Filter- versus Wrapper-based Feature Selection For Credit Scoring, International Journal of Intelligent Systems, Vol. 20 (10), pp. 985–999.

SPENCE, C., SAJDA, P., (1998). Role of feature selection in building pattern recognizers for computer-aided diagnosis, in Medical Imaging 1998: Image Processing, San Diego.

STĄPOR, K., (2011). Classification methods in computer vision. PWN, Warszawa (in Polish).

STĄPOR, K., (2015) Better alternatives for stepwise discriminant analysis. Acta Universitatis Lodziensis, Folia Oeconomica, Multivariate Statistical Analysis in Theory and Practice, nr 1(311), Lodziensis University Press, pp. 9–15.

THOMAS, L. C., (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting 16 (2), pp. 149–172.

THOMAS, L. C., OLIVER, R. W., HAND, D. J., (2005). A survey of the issues in consumer credit modelling research. Journal of the Operational Research Society 56 (9), pp. 1006–1015.

ZHANG, D., X., ZHOU, S., LEUNG, C. H., ZHENG, J., (2010). Vertical bagging decision trees model for credit scoring. Expert Systems with Applications 37 (12), pp. 7838–7843.

APPENDIX

The structure of the German credit data set

Attribute	Description	Values
1.	Status of existing checking account (qualitative)	A11 : ... < 0 DM A12 : 0 <= ... < 200 DM A13 : ... >= 200 DM /salary assignments for at least 1 year A14 : no checking account
2.	Duration in month (numerical)	
3.	Credit history (qualitative)	A30 : no credits granted/all credits paid back duly A31 : all credits at this bank paid back duly A32 : existing credits paid back duly until now A33 : delay in paying off in the past A34 : critical account/other credits existing (not at this bank)
4.	Purpose (qualitative)	A40 : car (new) A41 : car (used) A42 : furniture/equipment A43 : radio/television A44 : domestic appliances A45 : repairs A46 : education A47 : (vacation - does not exist?) A48 : retraining A49 : business A410 : others
5.	Credit amount (numerical)	
6.	Savings account/bonds (qualitative)	A61 : ... < 100 DM A62 : 100 <= ... < 500 DM A63 : 500 <= ... < 1000 DM A64 : .. >= 1000 DM A65 : unknown/ no savings account

7.	Present employment since (qualitative)	A71 : unemployed A72 : ... < 1 year A73 : 1 <= ... < 4 years A74 : 4 <= ... < 7 years A75 : .. >= 7 years
8.	Instalment rate in percentage of disposable income (numerical)	
9.	Personal status and sex (qualitative)	A91 : male : divorced/separated A92 : female : divorced/separated/married A93 : male : single A94 : male : married/widowed A95 : female : single
10.	Other debtors / guarantors (qualitative)	A101 : none A102 : co-applicant A103 : guarantor
11.	Present residence since (numerical)	
12.	Property (qualitative)	A121 : real estate A122 : if not A121 : building society savings agreement/life insurance A123 : if not A121/A122 : car or other, not in attribute 6 A124 : unknown / no property
13.	Age in years (numerical)	
14.	Other instalment plans (qualitative)	A141 : bank A142 : stores A143 : none
15.	Housing (qualitative)	A151 : rent A152 : own A153 : for free
16.	Number of existing credits at this bank (numerical)	

17.	Job (qualitative)	A171 : unemployed/ unskilled - non-resident A172 : unskilled - resident A173 : skilled employee / official A174 : management/ self-employed/highly qualified employee/ officer
18.	Number of people being liable to provide maintenance (numerical)	
19.	Telephone (qualitative)	A191 : none A192 : yes, registered under the customer's name
20.	Foreign worker (qualitative)	A201 : yes A202 : no

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 281–294

KNOWLEDGE-BASED ECONOMY IN THE EUROPEAN UNION – CROSS-COUNTRY ANALYSIS

Iwona Skrodzka[1]

ABSTRACT

The knowledge-based economy is an economy where knowledge is created, acquired, transmitted and used effectively by businesses, organizations, individuals and communities. It is not narrowly focused on the industries of advanced technology or ICT, but provides a framework for analysing the range of policy options in education, information infrastructure and systems of innovation, which could help contribute to the knowledge economy. The aim of the paper is to analyse spatial differences in the level of development of the knowledge-based economy in the European Union countries. The study uses a soft modelling method, which enables the estimation of a synthetic measure of KBE as well as the arrangement and classification of the UE-27 countries into typological groups. The research covers the years 2000 and 2013.

Key words: knowledge-based economy, knowledge assessment methodology, economic development, soft modelling.

1. Introduction

On the one hand, the knowledge-based economy (KBE) is perceived in a narrow sense as a part of economy dealing with knowledge industry, mainly science. However, in a broader sense, it is understood as the economy whose one production factor is knowledge (Piech, 2009, pp. 214). The classical definition of KBE is the one proposed by Organization for Economic Co-operation and Development(OECD), which defines it as an economy directly depending on knowledge and information production, distribution and use (OECD, 1996, pp. 7). The Asia-Pacific Economic Co-operation Economic Committee defined KBE as an economy in which the production, distribution, and the use of knowledge is the main driver of growth, wealth creation and employment across all industries (APEC Economic Committee, 2000, p. vii). According to the definition coined by the OECD and the World Bank Institute, KBE is an economy where knowledge is created, acquired, transmitted and used effectively by enterprises, organizations,

[1] Faculty of Economics and Management University of Bialystok. E-mail: i.skrodzka@uwb.edu.pl.

individuals and communities. It does not focus narrowly on high-technology industries or on information and communications technologies, but rather provides a framework for analysing a range of policy options in education, information infrastructure and innovation systems that can help usher in the knowledge economy (OECD, World Bank, 2001, pp. 3).

The vital work on KBE was the OECD report published in 1996, in which the notion of the 'knowledge economy' was used for the first time. Although during the last 20 years multiple studies have been conducted and numerous works have been written on KBE, one widely accepted measurement method has not been arrived at. We can only list a few dominant measurement methods, such as the Knowledge Assessment Methodology (KAM), drawn up by the World Bank, or the methodology proposed by the OECD. The work on them is still in progress, and each methodology is subject to constant criticism (Piech, 2009, pp. 315).

The paper focuses on the issue of measuring KBE in the European Union countries. KBE is difficult to measure due to its complexity, multidimensionality and unobservability. Its measurement requires prior solution to various problems such as: the imprecise and unquantifiable definition of KBE, the choice of the method, the choice of indicators referring to different aspects of KBE, the choice of an optimal set of indicators, data availability.

The aim of the paper is to analyse spatial differences in the KBE development level in the European Union countries (UE-27) in two periods of time – the years 2000 and 2013.In this study the concept of KBE measurement is based on KAM methodology and the soft modelling method. The following research hypotheses have been formulated:

H1: Observable variables (indicators) do not play equally important roles in reflecting the KBE development level in the European Union countries.

H2: Positive correlations between the pillars of KBE and the KBE development level in the European Union countries exist.

H3: A positive correlation between the KBE development level and the economic development level in the European Union countries exist.

2. Research method

In the literature the description of the soft modelling method can be found in (Wold, 1980), its generalization in (Rogowski, 1990) and examples of application in (Perło, 2004), (Skrodzka, 2015).

A soft model enables conducting the research of unobserved variables (latent variables). The values of these variables cannot be directly measured due to the lack of a generally accepted definition or the absence of a clear way of measuring them. A soft model consists of two sub-models:

- an internal sub-model – a system of relationships among latent variables, which describes the relationship arising from the theory,

- an external sub-model – defines the latent variables based on observed variables, known as indicators.

The indicators enable indirect observation of latent variables and are selected following the chosen theory or the researcher's intuition. In soft modelling, a latent variable can be defined by indicators in two ways: inductively – this approach is based on the assumption that indicators create latent variables (formative indicators) or deductively – this approach is based on the assumption that indicators reflect their theoretical notions (reflective indicators). In both approaches, latent variables are estimated as weighted sums of their indicators.

A soft model is constructed similarly to classical econometric models, with the following stages:

-specification of an internal sub-model (describing relationships among latent variables),

- specification of an external sub-model (describing latent variables by indicators),

- estimating model parameters with the Partial Least Square (PLS method), and

- statistical verification of a model (Stone-Geisser test and "2s" rule).

The Stone-Geisser test measures the prognostic property of a soft model. Its values are in the range from $-\infty$ to 1. A positive (negative) value of this test indicates high (poor) quality of the model."2s" rule says that if the doubled standard deviation, calculated based on the Tukey cut method, is lower than the absolute value of the parameter estimator, the parameter is statistically significant.

As a result of using the PLS method, we obtain estimates of latent variables, which can be regarded as synthetic measures. These quantities depend not only on external relations but also on relations among latent variables assumed in the internal model. It means that the cognition depends not only on the definition of a given notion but also on the theoretical description. Soft modelling makes full use of theoretical and empirical knowledge. This is one of the things which distinguishes the presented method from most of the commonly applied methods of multidimensional comparative analysis (this is also characteristic of structural models),

In this study the concept of KBE measurement is also based on the KAM methodology, which was developed within the framework of "The Knowledge for Development" (K4D) programme. The KAM methodology is regarded as the most efficient way of measuring KBE. It specifies four key pillars:

- Economic Incentive and Institutional Regime, responsible for developing economic policy and the work of institutions. The extension, dissemination and the use of knowledge by these entities is supposed to ensure effectiveness by an adequate division of resources and by boosting creativity.
- Education and Human Resources, which means personnel who can adapt to constantly developing technological solutions thanks to upgrading their skills.

- Innovation System, which involves the activities of economic entities, research centres, universities, advisory bodies and other organizations whose operations are adjusted to preferences of more and more demanding customers.
- Information Infrastructure, which ensures effective communication and faster transfer of data. All these aspects influence the transfer of information and knowledge (Chen, Dahlman, 2005, pp. 5–9).

The pillars are used to construct two global indexes:

- Knowledge Index (KI), which determines the knowledge potential of a country; this indicator is calculated as an arithmetic average of the three subindexes, which represent the three pillars of KAM (except the Economic Incentive and Institutional Regime);
- Knowledge Economy Index (KEI), which determines a general development level of the knowledge-based economy; this indicator is calculated as an arithmetic average of the four subindexes, which represent the four pillars of KAM (Chen, Dahlman, 2005, pp. 9–13).

The advantages of this method are: simplicity, clarity and versatility. It enables the comparison of the KI and KEI indicators and their components in both dimensions: intertemporal and international. The method is criticised, inter alia, for: insufficient theoretical background, the tendency to repeat information by indicators, the lack of differentiated weights for indicators, insufficient information about many of the analysed economies, inaccessibility of indicators in the systems of international statistics, incomparability of data due to a variety of data sources (Becla, 2010, pp. 56–70).

3. Specification of soft model

Figure 1 presents the concept of the internal sub-model. The concept assumes the relationship between two unobserved variables: the level of development of KBE and the level of economic development. KBE is defined by four pillars (according to KAM methodology): economic regime, education and human resources, innovation system and information infrastructure. They are also unobserved. Hence, KBE is the second-order latent variable.

The estimated model consists of two following equations:

$$KBE = \alpha_1 REG + \alpha_2 EDU + \alpha_3 INN + \alpha_4 ICT + \alpha_0 + \varepsilon \qquad (1)$$

$$ED = \beta_1 KBE + \beta_0 + \xi \qquad (2)$$

where

KBE –level of development of knowledge-based economy,

REG – economic regime,

EDU – education and human recourses,

INN – innovation system,

ICT – information infrastructure,

ED –level of economic development,

α_0, $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, β_0, β_1 – structural parameters,

ε, ξ– error terms.

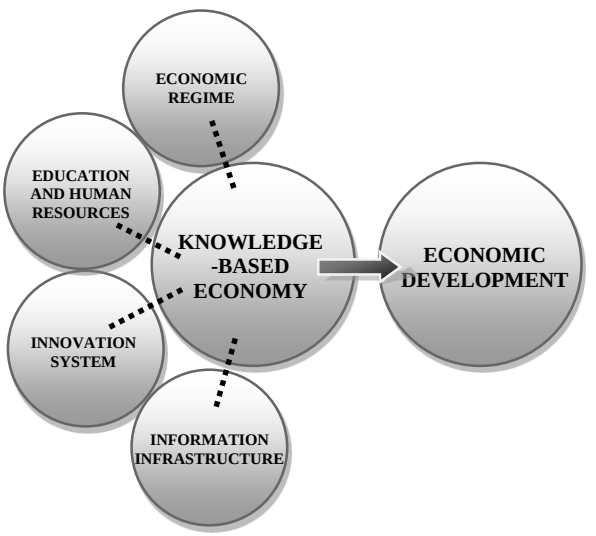


Figure 1. The concept of the internal sub-model
Source: own elaboration.

Each of the latent variables is defined by a set of indicators (see Table 1) based on a deductive approach. Data used to specify the model are taken from Eurostat and refer to 27 countries. Croatia was excluded from the research because of the large amount of missing data. The research focuses on the years 2000 and 2013,which is also related to the availability of data.

The following items were measured statistically: the variability of indicators (the coefficient of variation above 10%), the correlation level (depending on the way a latent variable is defined by indicators, an inductive or a deductive approach, indicators should show low or high correlation respectively). Missing data were complemented using native forecasting – complemented by adjacent values.

Table 1. Indicators of latent variables

Latent variable		Indicator	Meaning	Type of indicator
KBE	*REG*	*REG01*	Direct investment flows (% of GDP)	stimulant
		REG02	Exports of goods and services (% of GDP)	stimulant
		REG03	Business enterprise R&D expenditure (% of GDP)	stimulant
	EDU	*EDU04*	Persons with tertiary education attainment (%)	stimulant
		EDU05	Employees with tertiary education attainment (%)	stimulant
		EDU06	Life-long learning of persons aged 25-64 (%)	stimulant
		EDU07	Graduates (ISCED 5-6) in mathematics, science and technology (% of all fields)	stimulant
	INN	*INN08*	Persons employed in science and technology (% of total population)	stimulant
		INN09	Researchers in business enterprise sector (per 10 000 employees)	stimulant
		INN10	Total intramural R&D expenditure (% of GDP)	stimulant
	ICT	*ICT11*	Individuals who used computer in last 3 months (% of total population)	stimulant
		ICT12	Households with Internet access (%)	stimulant
		ICT13	Persons employed using computers with access to World Wide Web (% of total employment)	stimulant
ED		*ED01*	Gross domestic product per capita (PPS)	stimulant
		ED02	Gross value added per employee (PPS)	stimulant
		ED03	The share of agriculture in gross value added (%)	destimulant
		ED04	The share of professional, scientific and technical activities in gross value added (%)	stimulant
		ED05	Gini coefficient	destimulant

Source: own elaboration.

4. Diversity of knowledge-based economy in the European Union countries in 2000 – results of soft model estimation

The model presented in Figure 1 was estimated using the PLS software (created by J. Rogowski) based on data which refer to 2000. Table 2 contains estimates of the parameters of the external sub-model(weights, loadings) and standard deviations calculated based on the Tukey cut method. Indicators are ordered in decreasing order with regard to the absolute values of loadings (if the deductive approach is used to define the latent variable, we should interpret loadings).

Table 2. Estimates of the parameters of the external sub-model MM2000

Latent variable	Indicator	Weight	Standard deviation	Loading	Standard deviation
REG	*REG03*	0.8567	0.0118	0.9118	0.0091
	REG01	0.2547	0.0119	0.5340	0.0174
	REG02	0.1977	0.0112	0.4188	0.0185
EDU	*EDU04*	0.3647	0.0008	0.8940	0.0010
	EDU05	0.2806	0.0013	0.7967	0.0014
	EDU06	0.4829	0.0014	0.7566	0.0010
	EDU07	0.1984	0.0025	0.4282	0.0023
INN	*INN10*	0.3608	0.0012	0.9592	0.0001
	INN09	0.3510	0.0014	0.9278	0.0002
	INN08	0.3602	0.0003	0.9114	0.0001
ICT	*ICT11*	0.3633	0.0001	0.9608	0.0000
	ICT13	0.3486	0.0006	0.9528	0.0001
	ICT12	0.3351	0.0006	0.9511	0.0001
KBE	*ICT11*	0.1346	0.0010	0.9543	0.0005
	REG03	0.1289	0.0012	0.9177	0.0010
	ICT13	0.1285	0.0026	0.9157	0.0022
	INN10	0.1237	0.0014	0.9024	0.0032
	INN08	0.1251	0.0006	0.9009	0.0004
	ICT12	0.1342	0.0012	0.8802	0.0013
	INN09	0.1225	0.0034	0.8778	0.0035
	EDU06	0.0993	0.0033	0.7937	0.0060
	EDU04	0.0689	0.0025	0.5994	0.0064
	EDU05	0.0489	0.0027	0.4612	0.0060
	EDU07	0.0542	0.0012	0.3261	0.0015
	REG01	0.0568	0.0102	0.2728	0.0162
	REG02	0.0457	0.0069	0.2118	0.0146
ED	*ED01*	0.3176	0.0408	0.9393	0.1415
	ED02	0.2708	0.0383	0.9173	0.0998
	ED03	-0.2812	0.0271	-0.8437	0.0523
	ED04	0.2047	0.0460	0.7108	0.1292
	ED05	-0.1853	0.0414	-0.3810	0.0913

Source: own calculation.

All parameters are statistically significant ("2s" rule). Moreover, the results are consistent with expectations. The stimulants have positive weights and loadings and destimulants have negative ones.

Indicators have a different strength of impact on latent variables. *REG* variable is very strongly correlated with*REG03* indicator and moderately correlated with *REG01* and *REG02* indicators. *EDU* variable is strongly reflected by *EDU04*, *EDU05*, *EDU06* indicators and moderately reflected by*EDU07* indicator.*INN* and *ICT* variables are very strongly correlated with all indicators that define them. *KBE* variable is very strongly reflected by *ICT11*, *REG03*, *ICT13*, *INN10*, *INN08* indicators, while indicators *REG01*, *REG02* are weakly correlated with this variable. *ED* variable is very strongly correlated with *ED01* and *ED02* indicators, strongly correlated with *ED03* and *ED04* indicators, and weakly correlated with *ED05* indicator.

Equations (3) and (4) present estimations of the parameters of the internal sub-model. Standard deviations calculated based on the Tukey cut method are given in brackets.

$$\hat{KBE} = \underset{(0.0286)}{0.2213}REG + \underset{(0.0050)}{0.1967}EDU + \underset{(0.0196)}{0.2844}INN + \underset{(0.0055)}{0.3869}ICT + \underset{(0.0156)}{0.0553} \tag{3}$$

$$\hat{ED} = \underset{(0.0439)}{0,7612}KBE - \underset{(0.3840)}{1,7854} \tag{4}$$

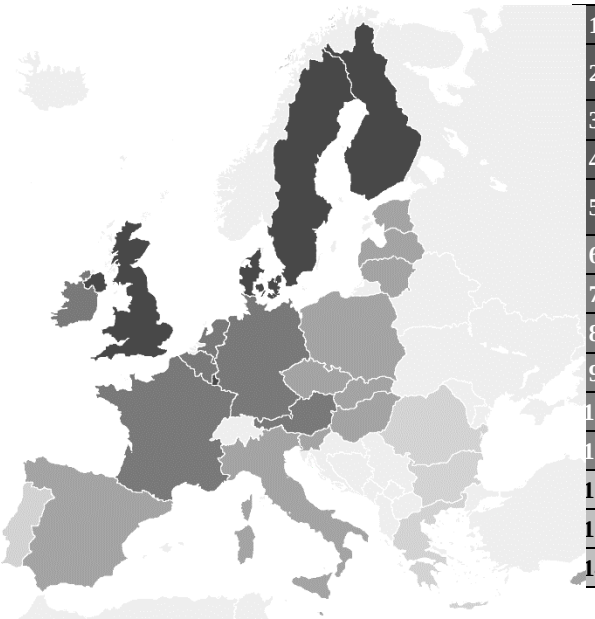
The signs of estimators are consistent with expectations. Furthermore, all latent variables are statistically significant ("2s" rule). The coefficient of determination (R^2) has the value of 1.0 for the equation (3) and the value of 0.6 for the equation (4). The general Stone-Geisser test is equal to 0.31. The model can be verified positively.

All four pillars have a positive influence on the level of KBE development (see equation 3). The pillar "information infrastructure" has the strongest impact (0.3869) and "education and human recourses" has the lowest (0.1967). The equation (4) shows that the relationship between the level of KBE development and the level of economic development is positive and strong (compare with (Dworak, 2010)).

Estimates of the values of latent variables were used to order the UE-27 countries according to the level of KBE development and to classify countries into four typological groups. Groups were constructed based on the parameters of a synthetic measure: average and standard deviation (Nowak, 1990, pp. 92–93):

- I group – very high level of KBE development,
- II group – high level of KBE development,
- III group – medium and low level of KBE development,
- IV group – very low level of KBE development.

Figure 2 presents the results of the classification.



1.	Sweden	**15.**	Cyprus
2.	Finland	**16.**	Czech Republic
3.	Denmark	**17.**	Malta
4.	Luxembourg	**18.**	Slovakia
5.	United Kingdom	**19.**	Lithuania
6.	Netherlands	**20.**	Italy
7.	Germany	**21.**	Latvia
8.	Belgium	**22.**	Hungary
9.	Austria	**23.**	Poland
10.	France	**24.**	Greece
11.	Ireland	**25.**	Portugal
12.	Slovenia	**26.**	Bulgaria
13.	Estonia	**27.**	Romania
14.	Spain		

Figure 2. The level of development of the knowledge-based economy in the UE-27 countries in 2000

Source: own elaboration.

A very high level of KBE development was achieved in 2000 by: Sweden, Finland, Denmark, Luxemburg and United Kingdom. Six countries: Netherlands, Germany, Belgium, Austria, France and Ireland a had high level of KBE development. The group of countries with a medium and a low level of KBE development included Slovenia, Estonia, Spain, Cyprus, Czech Republic, Malta, Slovakia, Lithuania, Italy, Latvia, Hungary and Poland. Greece, Portugal, Bulgaria and Romania were in the last group with very low level of KBE development. Poland was 23rd in the ranking and was classified in the third group.

5. Diversity of knowledge-based economy in the European Union countries in 2013 – results of soft model estimation

Table 3 contains estimates of the parameters of the external sub-model and standard deviations. All parameters are statistically significant. Furthermore, the results are consistent with expectations – stimulants have positive weights and loadings and destimulants have negative ones.

Table 3. Estimates of the parameters of the external sub-model

Latent variable	Indicator	Loading	Standard deviation	Weight	Standard deviation
REG	REG03	0.9480	0.0123	0.8817	0.0175
	REG01	0.3190	0.0119	0.3511	0.0347
	REG02	0.1883	0.0192	0.2767	0.0353
EDU	EDU04	0.3573	0.0013	0.8957	0.0016
	EDU05	0.2803	0.0021	0.8085	0.0022
	EDU06	0.4945	0.0038	0.7704	0.0026
	EDU07	0.1822	0.0010	0.3975	0.0010
INN	INN09	0.3644	0.0009	0.9646	0.0004
	INN10	0.3360	0.0022	0.9289	0.0008
	INN08	0.3824	0.0033	0.8795	0.0012
ICT	ICT11	0.3465	0.0004	0.9738	0.0001
	ICT12	0.3413	0.0008	0.9686	0.0001
	ICT13	0.3515	0.0013	0.9447	0.0002
KBE	ICT13	0.1302	0.0018	0.9484	0.0027
	ICT11	0.1336	0.0009	0.9349	0.0019
	INN08	0.1350	0.0028	0.9243	0.0044
	ICT12	0.1385	0.0012	0.9209	0.0028
	INN09	0.1197	0.0017	0.8809	0.0055
	EDU06	0.1110	0.0022	0.8614	0.0047
	INN10	0.1088	0.0026	0.8123	0.0083
	REG03	0.1095	0.0024	0.7978	0.0077
	EDU04	0.0835	0.0012	0.6223	0.0036
	EDU05	0.0656	0.0015	0.4883	0.0047
	EDU07	0.0510	0.0023	0.3174	0.0024
	REG01	0.0592	0.0087	0.2685	0.0168
	REG02	0.0413	0.0075	0.1584	0.0166
ED	ED01	0.2673	0.0316	0.8982	0.1093
	ED03	-0.2742	0.0342	-0.8860	0.0734
	ED02	0.2192	0.0578	0.8497	0.0870
	ED04	0.2247	0.0505	0.8121	0.0788
	ED05	-0.2332	0.0404	-0.6356	0.1540

Source: own calculation.

REG variable is strongly correlated with *REG03* indicator and weakly correlated with *REG01* and *REG02* indicators. *EDU* variable is strongly reflected by *EDU04*, *EDU05*, *EDU06* indicators and weakly reflected by *EDU07* indicator. *INN* and *ICT* variables are very strongly correlated with all indicators that define them. *KBE* variable is very strongly reflected by *ICT13*, *ICT11*,*INN08*, *ICT12* indicators, while indicators:*EDU07*, *REG01*, *REG02* are weakly correlated with this variable. *ED* variable is strongly correlated with all indicators except for one – ED05.

Equations (5) and (6) present estimations of the parameters of the internal sub-model.

$$\underset{(0.0219)}{\hat{KBE}} = \underset{(0.0219)}{0.1945}REG + \underset{(0.0034)}{0.2374}EDU + \underset{(0.0322)}{0.2221}INN + \underset{(0.0104)}{0.4312}ICT + \underset{(0.0157)}{0.0523} \quad (5)$$

$$\underset{(0.0402)}{\hat{ED}} = \underset{(0.0402)}{0.8053}KBE - \underset{(0.8033)}{3.8929} \quad (6)$$

The signs of estimators are consistent with expectations. Moreover, all latent variables are statistically significant ("2s" rule). The coefficient of determination (R^2) has the value of 1.0 for the equation (5) and the value of 0.65 for the equation (6). The general Stone-Geisser test is equal to 0.27. The model can be verified positively.

All four pillars have a positive influence on the level of KBE development. The pillar "information infrastructure" has the strongest impact (0.4312) and "economic regime" has the lowest (0.1945). The equation (6) shows that the relationship between the level of KBE development and the level of economic development is positive and strong.

Figure 3 presents the results of classification of the UE-27 countries according to the level of KBE development in 2013. Countries are divided into four groups. The first group– countries with the highest level of KBE development – consists of: Sweden, Finland, Denmark, Finland and Luxemburg. Countries: Netherlands, United Kingdom, Germany, France, Ireland, Belgium, Austria, Slovenia and Estonia are in the second group and have a high level of KBE development. The third group includes: Czech Republic, Spain, Malta, Hungary, Lithuania, Cyprus, Latvia, Slovakia, Poland, Portugal and Italy. They have medium and low level of KBE development. Very low level of KBE development is characteristic for: Greece, Bulgaria and Romania. Poland was 22nd in the ranking and was classified in the third group.

1.	Sweden	**15.**	Spain
2.	Denmark	**16.**	Malta
3.	Finland	**17.**	Hungary
4.	Luxembourg	**18.**	Lithuania
5.	Netherlands	**19.**	Cyprus
6.	United Kingdom	**20.**	Latvia
7.	Germany	**21.**	Slovakia
8.	France	**22.**	Poland
9.	Ireland	**23.**	Portugal
10.	Belgium	**24.**	Italy
11.	Austria	**25.**	Greece
12.	Slovenia	**26.**	Bulgaria
13.	Estonia	**27.**	Romania
14.	Czech Republic		

Figure 3. The level of development of the knowledge-based economy in the UE-27 countries in 2013

Source: own elaboration.

6. Conclusions

The studies presented in the paper concerned the analysis of spatial differences in the KBE development level in the EU-27 countries. The soft modelling method used in research enabled:

- the investigation into the relationships between indicators and the KBE latent variable,
- the investigation into the relationships between the pillars of KBE and the KBE development level as well as between the KBE development level and the economic development level in the European Union countries,
- the estimation of the values of KBE synthetic measure and the arrangement of countries according to the KBE development level as well as the division of counties into typological groups.

In both estimated models (2000 and 2013) indicators had a different strength of impact on the KBE latent variable (from very strong to weak). Moreover, both estimated models indicated positive influence of the KBE pillars on the KBE development level. Furthermore, in both estimated models the relationship between the KBE development level and the economic development level was

positive and strong. Hence, the hypotheses which were formulated in the introduction can be positively verified.

The highest level of development of the knowledge-based economy both in 2000 and in 2013 was characteristic for Sweden, Denmark, Finland and Luxembourg, whereas the lowest one for Greece, Bulgaria and Romania. Four of the 27 countries were classified into other typological groups in 2013 compared to 2000. The United Kingdom was classified into the group with a lower level of KBE development, while Slovenia, Estonia and Portugal to the group with a higher level of KBE development. Eleven countries, including Poland, improved their ranking in 2013 compared to 2000, while nine countries reduced their positions. The highest increase was in Hungary (22nd position in 2000 and 17th position in 2013) and the largest fall in Italy (22nd position in 2000 and 24th position in 2013).

REFERENCES

APEC Economic Committee, (2000). Towards Knowledge-based Economies in APEC.

BECLA, A., (2010). Wady i zalety metody KAM (Knowledge Assessment Methodology) służącej do identyfikacji poziomu zaawansowania gospodarki opartej na wiedzy [Advantages and disadvantages of KAM method (Knowledge Assessment Methodology) used for the identification of the level of advancement of the knowledge-based economy], In: Prace Naukowe UE we Wrocławiu, Wrocław: Wydawnictwo UE we Wrocławiu, No. 139, pp. 56–70.

CHEN, D. H. C., DAHLMAN, C. J., (2005). The Knowledge Economy, the KAM Methodology and World Bank Operations [online].Washington: World Bank Institute, D.C. 20433, http://siteresources.worldbank.org/INTUNIKAM/Resources/2012.pdf.

DWORAK, E., (2010). Analysis of knowledge-based economy impact on economic development in the European Union countries, Comparative Economic Research. Central and Eastern Europe Vol. 13, No. 4, pp. 5–25.

OECD, (1996). The Knowledge-based Economy. Paris.

OECD, World Bank, (2001). Korea and Knowledge-based Economy. Making the Transition. Paris.

NOWAK, E., (1990). Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych [Taxonomic methods in the classification of socio-economic subjects], Warsaw: PWE.

PERŁO, D., (2004). Źródła finansowania rozwoju regionalnego [Sources of funding for regional development], Białystok: Wydawnictwo Wyższej Szkoły Ekonomicznej w Białymstoku.

PIECH, K., (2009). Wiedza i innowacje w rozwoju gospodarczym: w kierunku pomiaru i współczesnej roli państwa [Knowledge and innovation in economic development: towards the measurement and contemporary role of the state], Warsaw: Instytut Wiedzy i Innowacji.

ROGOWSKI, J., (1990). Modele miękkie. Teoria i zastosowanie w badaniach ekonomicznych [Soft models. Theory and application in economic studies], Białystok: Wydawnictwo Filii UW w Białymstoku.

SKRODZKA, I., (2015). Kapitał ludzki polskich województw – koncepcja pomiaru [Human capital of Polish voivodships - the concept of measurement], Białystok: Wydawnictwo UwB.

WOLD, H., (1980). Soft Modelling: Intermediate between Traditional Model Building and Data Analysis, Banach Centre Publication 6, Mathematical Statistics.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 295–304

NEW METHOD OF VARIABLE SELECTION FOR BINARY DATA CLUSTER ANALYSIS

Jerzy Korzeniewski[1]

ABSTRACT

Cluster analysis of binary data is a relatively poorly developed task in comparison with cluster analysis for data measured on stronger scales. For example, at the stage of variable selection one can use many methods arranged for arbitrary measurement scales but the results are usually of poor quality. In practice, the only methods dedicated for variable selection for binary data are the ones proposed by Brusco (2004), Dash et al. (2000) and Talavera (2000). In this paper the efficiency of these methods will be discussed with reference to the marketing type data of Dimitriadou et al. (2002). Moreover, the primary objective is a new proposal of variable selection method based on connecting the filtering of the input set of all variables with grouping of sets of variables similar with respect to similar groupings of objects. The new method is an attempt to link good features of two entirely different approaches to variable selection in cluster analysis, i.e. *filtering* methods and *wrapper* methods. The new method of variable selection returns best results when the classical *k*-means method of objects grouping is slightly modified.

Key words: cluster analysis, market segmentation, selection of variables, binary data, k-means grouping.

1. Introduction

Feature selection is probably the most important stage of cluster analysis just like in many other parts of statistics. The results of variable selection determine significantly the final results of cluster identification and if received incorrectly may render it impossible to identify any clusters. The task of variable selection in cluster analysis has probably been highlighted by the well-known article by Carmone et al. (1999) in which the HINoV method was proposed. Although in this article the authors mention some earlier attempts to approach the task of variable selection, they assess them as absolutely infeasible in application to empirical data. After 1999 several methods or algorithms for variable selection were proposed, however, most of them were meant rather for strong scales on

[1] University of Lodz, Poland. E-mail: jurkor@wp.pl.

which the variables are measured. A good evaluation of some of them is given in Steinley et al. (2008) and in Korzeniewski (2012). Some of these methods allow even for a form of statistical inference like, e.g. Raftery and Dean (2006) method. As far as weaker measurement scales are concerned, e.g. binary data sets, it is not easy to find any well-performing methods. The methods developed by Brusco (2004), Dash et al. (2000) and Talavera (2000) should be mentioned as the ones to be investigated and assessed. A particular problem of cluster analysis of binary data arises when one is confronted with the task of market segmentation. A characteristic feature of this type of data is the existence of a couple of groups of binary variables with possible pairwise correlation within the groups. Such data usually are confronted with when carrying out statistical research of large numbers of clients on the market. This kind of research is very often performed in the form of a questionnaire comprising several questions with possible binary answers. Dimitriadou et al. (2002) proposed a way of simulating this kind of binary data for the task of determining the number of clusters. Their *bindata* package as recent as 2015 is freely available in R language. One very characteristic feature of the marketing type of binary data is its relatively big size – the number of possible objects is at least several thousands. The conclusion, therefore, is that one rather has to use partitioning methods of objects grouping, one cannot use, e.g. agglomerative methods.

The objective of this article is to propose a new and efficient method for the earlier stage of cluster analysis, i.e. variable selection on the marketing type of binary data and to assess the efficiency of this method in comparison with other existing methods. The article is organized as follows. In the next part an overview of the methodology of three methods is presented with possible hints of their applicability to the task in the context of marketing data. Part three presents a proposal of the new method. Part four includes an empirical evaluation of the new method and other methods. The final fifth part contains conclusions and prospects for future research.

2. Overview of variable selection methods

The number of variable selection methods in cluster analysis is quite large comprising several proposals, however, methods which were constructed for strong measurement scales (predominantly continuous variables) do not perform well for binary variables or cannot be applied at all. This phenomenon is quite common across all statistical methods. Therefore, we limit our examinations to the three methods described in this chapter which were constructed for nominal scales, some of them especially for a binary scale.

The Brusco method of variable selection dedicated for binary variable consists of the following steps:

1) For arbitrary subset of the set of all variables group all data set objects 500 times in the predetermined number of clusters and remember the sum of average distances inside the clusters.

2) Add a single new variable to this subset if the new sum (with the new variable) of average distances is smaller than the previous one (the very subset), i.e. $Z_{\min} < Z_{B2}$.
3) Stop the process of variable adding if

$$Z_{B2} > Z_{\min} + \delta \frac{M}{4} \tag{1}$$

where M is the number of data set numbers and $\delta \in (0,1)$ is a parameter to be fixed intuitively.

There are some major doubts which can be raised about this method. Firstly, Brusco uses the classical form of *k*-means grouping stating that it renders good results. The results depend on the type of data used in the experiment. If the data were slightly more obscure (clusters less distinct) the results could be much worse because the classical *k*-means is not well suited for binary variables as in the first loop there are many draws (equal Sokal-Michener distances) and it is impossible to say into which direction a given object should "go". Secondly, the number of clusters into which the sets have to be partitioned is the proper one and Brusco advocates that there are "excellent" ways of determining the number of clusters, mentioning and recommending the Ratkowsky-Lance index. The Ratkowsky-Lance index was investigated by Dimitriadou et al. (2002) on the marketing binary data (as well as by other authors) and the results are quite clear – it finds the proper number of clusters in about 70% of cases, sometimes going wrong by more than two clusters. If the number of clusters was established erroneously then the results would probably be worse. Thirdly, there is a question of the level of cluster separation which is discussed below.

Table 1. The pattern of Brusco binary data.

Number of clusters	Number of variables		
	4	6	8
4	1 0 0 1 1 1 1 0 0 0 1 1 0 1 0 1	1 0 0 1 1 0 1 1 1 0 0 0 0 0 1 1 0 0 0 1 0 1 0 1	1 0 0 1 1 0 1 0 1 1 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0 1 0 1 1 0
6	1 0 0 1 1 1 1 1 1 0 1 0 0 1 0 1 0 0 0 1 0 1 1 0	1 0 0 0 1 1 1 1 0 1 1 0 1 1 1 0 0 0 0 1 0 0 0 1 0 1 1 1 1 0 0 0 0 1 1 0	1 0 0 0 1 1 0 1 1 1 0 1 1 0 1 0 1 1 1 0 0 0 0 1 0 1 0 0 0 1 1 1 0 1 1 1 1 0 1 1 0 0 0 1 1 0 0 1
8	1 0 1 1 1 0 0 0 1 1 1 0 1 1 0 1 0 1 0 1 0 1 0 0 0 0 1 1 0 0 0 1	1 0 0 1 1 1 1 0 1 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1 0 1 0 0 1 0 0 1 1 0 0 1 0 0 1 1 1 0 0 0 1 0 0 1	1 0 0 1 1 1 0 1 1 0 1 0 0 0 1 1 1 1 1 1 1 1 0 0 1 1 0 0 0 1 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0 0 1 0 1

Source: Brusco (2004).

The outlined above method was evaluated by Brusco on the data sets the skeleton of which is given in Table 1. Obviously, the data was varied with respect to the size of clusters, the number of objects in data sets, the level of cluster separation, etc. A natural question arises: what is the difference between data sets of this type and the ones generated by Dimitriadou et al. (see section 4)? The answer seems to be that the major difference lies in the level of cluster separation. Brusco allows only for 4% (at the worst case) of 1 being changed into 0 or vice versa. The way of defining cluster separation is entirely different in the work of Dimitriadou et al. (2002). It seems, however, that their levels of, e.g. 0.8 for 1 and 0.8 for 0, allow for much less clear cluster structure, to say nothing of the levels of 0.7 and 0.3, respectively.

The method proposed by Talavera consists in using the formula

$$Kor(v_M) = \frac{\sum_{v}\sum_{j} P(x_v = a_{vj}) \sum_{j_M} \left(P^2\left(x_{v_M} = a_{vj_M} \middle| x_v = a_{vj}\right) - P^2\left(x_{v_M} = a_{vj_M}\right)\right)}{\left|\{v | v \neq v_M\}\right|} \tag{2}$$

for arranging all variables in descending order with respect to the strength of correlation between variable v_M and the remaining variables. In formula (2) the symbol a_{vj} stands for the *j*-th variant of *v*-th variable and the formula was derived with the use of the Bayes theorem starting from the maximization of a measure of the quality of the division of the data set into a predetermined number of clusters. It seems that we can assess this method at this stage basing our judgement on the evaluations which can be found in the literature. In order to apply the Talavera method to a particular data set one has to use the COBWEB algorithm (or similar based on a hierarchical tree). All applications to be found (e.g. Devaney et al. (1997)) analyse small data sets of not more than a couple of hundred objects (e.g. heart disease UCI data set and LED UCI data set). It is not feasible to apply this method to the whole data sets of the marketing type (one rather has to use partitioning methods) unless one tries to draw small samples and somehow unify the results. Besides, the number of clusters has to be known.

The Dash and Liu method is a very general method which can be applied to any measurement scale because it is based on the analysis of the data set entropy. The smaller the entropy (for different combinations of variables used) the better it is for the considered set of the variables with respect to the strength of evidence on possible cluster structure. The entropy of the data set is measured with the formula

$$E = -\sum_{x_1, x_2} \left[S(x_1, x_2) \log S(x_1, x_2) + (1 - S(x_1, x_2)) \log(1 - S(x_1, x_2)) \right] \tag{3}$$

where $S(x_1, x_2)$ stands for the similarity of two objects being a simple transformation of the distance between these two objects. It seems that we can assess the applicability of this method at this stage taking into consideration its basic characteristics. Firstly, the entropy-based method only allows for the ordering of all variables with respect to their importance to a possible cluster structure. If this ordering is done incorrectly there are no chances of the correct selection of variables. This seems to be a major drawback. Secondly, one needs some kind of criterion as to where to divide the sequence of ordered variables. The authors suggest a criterion based on the results of objects grouping. However, as in the case the Brusco method, the number of clusters has to be predetermined. It is possible to assess the entropy of the data set on any chosen subset of the set of all variables and pick up the best one. However, this leads to the necessity of examining all possible subsets of variables.

3. New method formulation

To make the presentation of our proposal as clear as possible let us start from dividing this method into two stages:

- Stage 1. Filtering stage which consists in grouping all variables into classes of similar variables with respect to some kind of correlation measure.
- Stage 2. Wrapper stage which consists in possible grouping of the classes of variables received in stage 1 with respect to the similarities of grouping of the data set objects.

Any method consisting of the two steps given above is not going to work properly if one uses classical techniques like, e. g. coefficient of linear correlation in stage 1 or

k-means clustering in stage 2 due to well known drawbacks of these methods when applied to binary data. However, if we use more versatile measures the method is going to work very well.

Thus, in stage 1 we will use the *distance based correlation* (Korzeniewski, 2012) between two sets *A*, *B* of variables given by the formula:

$$DBC(A,B,l) = \frac{\frac{1}{l}\sum_{i=1}^{l} d_i^A d_i^B - \bar{d}^A \bar{d}^B}{s^A s^B}, \tag{4}$$

where $1 \le l \le n$ denotes the number of observation pairs drawn without replacement from all pairs of observations; d_i^A, d_i^B denote distances for *i*-th pair of objects based on the variables from sets *A*, *B*, respectively; $\bar{d}^A$, $\bar{d}^B$, s^A, s^B denote arithmetic means and standard deviations computed from all *l* distances on both sets of variables, respectively. This kind of correlation measure is extremely useful when applied in cluster analysis (Korzeniewski,

2012) because if there is a cluster structure and both sets of variables *A* and *B* participate in creating it, then any substantial changes in distances between objects on set *A* should cause changes of distances on set *B*. To fix all technicalities let us establish that we will apply formula (4) only to sets *A* and *B* consisting of single variables and *l*=20 with the value of $DBC(u,v)$ (*l* is skipped) being the arithmetic mean from 100 repetitions.

In stage 2 we have to use some kind of partitioning algorithm because the data sets are too big for agglomerative algorithms. The most popular and applied in virtually any comparative simulation study is the *k*-means clustering. However, in the case of binary data sets, it cannot be used under no pretence whatever. Firstly, as in the case of the methods described in section 2, we would have to specify the number of clusters *k* which would give no advantage over the other methods. Secondly, one of the basic drawbacks of *k*-means partitioning of binary data is its ambiguity caused by huge percentage of draws in the first loop of the *k*-means partitioning. We propose the following partitioning process based on multiple *k*-means partitioning for *k*=2. We partition the whole data set into two clusters, then each of the two clusters is partitioned into two clusters and so on. Such a way of partitioning gives much better results being a cure for almost all *k*-means ailments. We only have to specify a stopping rule. It can be, e.g. the minimum cluster size, however this would be a new parameter, nowhere to be found. A better way is to specify a threshold of a reasonable quality of partitioning into 2 clusters. There are many measures of data division quality, e.g. the ones based on replication techniques. Another stopping rule can refer simply to the depth of the partitioning process. In the case of the binary marketing data this way is absolutely sufficient since due to multiple partitioning of the same (or very similar) data sets being subsets of the whole data set and the random character of the initialization of *k*-means partitioning, as well as a small number of clusters into which we want to segment the market, it is enough if we establish the depth to be 3. To be more precise, we partition the whole data set into 2 clusters, each of which is partitioned into two clusters, each of which is likewise partitioned. There is one another justification for this relatively small depth of partitioning, namely the value of the threshold from which we will decide that two divisions made on two different sets of variables are similar. The measure of the quality of division will be the adjusted Rand index (Hubert, 1985) which usually assumes values from interval (0,1). We set the threshold value to be 0.15. This is a small number as far as demanding high similarity of divisions is considered. However, it is sufficient, as two different binary variables with equal and random distributions of their variants never returned at least one value of the Rand index greater than 0.15 in 1000 random simulations (*k*-means, *k*=2, random starting points). Therefore, the threshold of 0.15 seems to be a very mild one and at the same time a rigorous one. As the threshold is very small we do not have to seek very intently for two very similar partitions, that is why we can stop our multiple partitioning at the depth of 3.

Summing up the above considerations we propose the following steps:

1. Group all variables into classes of variables such that in each class, for every variable *v* there is variable *u* such that $DBC(u,v) > 0.1$.
2. Merge two different groups of variables resulting from step 1 if the value of the adjusted Rand index between any of the 8 divisions of the data set made on one group of variables and any of the 8 divisions made on the other group of variables exceeds 0.15.
3. Repeat step 2 until no merges can be made.
4. Consider all single variables to be noisy variables, i.e. not participating in creating cluster structure and discard them.
5. Consider each class of variables consisting of more than one variable to be important for cluster structure. If there is more than one such class, it suggests the existence of multiple cluster structures.

4. Simulation experiment

In order to assess the efficiency of the new method on binary marketing data 162 data sets were generated. We followed the pattern suggested by Dimitraidou et al. [2002] in which every data set is described by twelve binary variables composed into four groups of different or equal numbers of variables. An example of such data pattern is presented in Table 2. The idea of this example is to present connections between groups of

Table 2. An example of binary marketing data pattern, twelve variables in four groups.

	Group1			Group2			Group3			Group4		
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
Cluster1	H	H	H	H	H	H	L	L	L	L	L	L
Cluster2	L	L	L	L	L	L	H	H	H	H	H	H
Cluster3	L	L	L	H	H	H	H	H	H	L	L	L
Cluster4	H	H	H	L	L	L	L	L	L	H	H	H
Cluster5	L	L	L	H	H	H	L	L	L	H	H	H
Cluster6	H	H	H	L	L	L	H	H	H	L	L	L

Source: Dimitriadou et al. [2002].

respondents and groups of questions in a questionnaire. The symbol H stands for the high probability of value 1 on a given variable and the symbol L stands for the low probability of 1. Obviously, the number of variables in each group, their correlation within the group, the level of H and L will vary. We used the very recent *bindata* (Leisch et al., 2015) package available in R language. The data sets generated were diversified with respect to the following parameters.

- Probability; for H there are 3 variants: 0.9, 0.8, 0.7 and for each variant, respectively, for L there are 3 variants: 0.1, 0.2, 0.3.
- Correlation inside groups of variables; there are 3 variants: uncorrelated variables, variables correlated with moderate strength (0.4), variables correlated with great strength (0.8).
- Number of clusters; 3 variants: 4, 5, 6.
- Numbers of objects in the clusters: 3 variants: (1000, 1000, 1000, 1000, 1000, 1000), (2000, 500, 1000, 700, 700, 1100), (3000, 300, 1000, 500, 700, 500).
- Number of variables within groups; 2 variants: (3, 3, 3, 3), (5, 4, 2, 1).

If there are less than 6 clusters we take into account only the initial clusters, i.e. the ones from the top of Table 2. All combinations of variants result in 162 data sets. In order to assess the efficiency of the new method in proper variable selection a similar number of noisy variables were added to each of 162 data sets. The noisy variables resulted in adding 8 uniformly distributed sets of observations coming from pairwise independent binary variables (equal probabilities for 1 and 0).

Every *k*-means partitioning in the new algorithm described in section 2 follows the classical form of this method, i.e. the two starting objects are randomly chosen, the procedure is repeated 100 times and we pick up the variant with the smallest sum of squared distances. The distance measure used is the Sokal-Michener distance.

5. Results and conclusions

The new method performed very well because it was almost perfect in 89% of data sets and absolutely wrong (returning 20 separate variables, i.e. discovering no cluster structure) in 11% of data sets. What is more, the wrong decisions comprised all 18 data sets (and none else set) with no correlation in the groups of variables creating cluster structure and the weakest variant of cluster separation, i.e. probabilities of 0.7 and 0.3 for high (H) and low (L) probability of 1, respectively. In other words, if there is at least a small hint of cluster structure existence (i.e. correlation between variables or decent levels of cluster separation), the new method is very likely to detect it. Other numerical characteristics of the results are as follows. In two cases (1.2% of data sets) the new method incorporated more than 4 noisy variables into the set of variables true for cluster

structure. In 9 cases (5.6% of data sets) the new method found 2 separate cluster structures, usually one "major" created by 8 or 9 variables and one "minor" created by 2 or 3 variables. It seems, however, that in these cases one could pursue the detecting similarities of divisions of data set objects in some other way than the techniques used, because one has more options to detect such similarities when there are a few variables in each of the two sets of variables. In 21 cases (13% of data sets), with the vast majority from the second type of data sets, i.e. the second (5, 4, 2, 1) case of the numbers of variables in each group, the new method missed one single variable (properly detecting 11 others). The new method allows for perfect variable selection (selecting 12 true variables and discarding 8 single remaining variables) in 65% of data sets.

REFERENCES

CARMONE, F., KARA, A., MAXWELL, S., (1999). HINoV: A New Model to Improve Market Segment Definition by Identifying Noisy Variables, Journal of Marketing Research, Vol. 36, No. 4, 501–510.

DASH, M., LIU, H., (2000). Feature selection for clustering, Proceedings of Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, (PAKDD), 110–121.

DEVANEY, M., RAM, A., (1997). Efficient feature selection in conceptual clustering, Proceedings of the Fourteenth International Conference on Machine Learning, Nashville, 92–97.

DIMITRIADOU, E., DOLNICAR, S., WEINGESSEL, A., (2002). An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets, Psychometrika 67(1), 137–160.

HUBERT, L., ARABIE, P., (1985). Comparing Partitions, Journal of Classification 2.

LEISCH, F., WEINGESSEL, A., HORNIK, K., (2015). Bindata package manual.

KORZENIEWSKI, J., (2012). Selekcja zmiennych w analizie skupień. [The selection of variables in cluster analysis]. Nowe procedury, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

RAFTERY, E., DEAN, N., (2006). Variable selection for model-based clustering, Journal of the American Statistical Association , 101(473): 168–178.

STEINLEY, D., BRUSCO, M., (2007). Initializing k-means batch clustering: A critical evaluation of several techniques, Journal of Classification 24, 99–121.

STEINLEY, D., BRUSCO, M., (2008). Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures, Psychometrika 73, 125–144.

TALAVERA, L., (2000). Dependency-Based Feature Selection for Clustering Symbolic Data, Intelligent Data Analysis 4, 19–28.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 305–316

THE GLUEVAR RISK MEASURE AND INVESTOR'S ATTITUDES TO RISK– AN APPLICATION TO THE NON-FERROUS METALS MARKET

Dominik Krężołek[1]

ABSTRACT

Investing in the economic world, characterized by a high level of uncertainty and volatility, entails a higher level of risk related to investment. One of the most commonly used risk measure is Value-at-Risk. However, despite the ease of calculation and interpretation, this measure suffers from a significant drawback – it is not subadditive. This property is the key issue in terms of portfolio diversification. Another risk measure, which meets this assumption, has been proposed – Conditional Value-at-Risk, defined as a conditional loss beyond Value-at-Risk. However, the choice of a risk measure is an individual decision of an investor and it is directly related to his attitudes to risk.
In this paper the new risk measure is proposed – the GlueVaR risk measure, which can be defined as a linear combination of VaR and GlueVaR. It allows for calculating the level of investment loss depending on investment's attitudes to risk. Moreover, GlueVaR meets the subadditivity property, therefore it may be used in portfolio risk assessment. The application of the GlueVaR risk measure is presented for the non-ferrous metals market.

Key words: risk, metal market, subadditivity, VaR, GlueVaR

1. Introduction

In the economic and financial world any disturbances observed in the market as well as additional (non-market) factors affect significantly the level of risk taken. Given the market risk, its level often derives from the investor's behaviour and the way how he assesses the reality around him. The reality is usually different from the assumptions of statistical models, where the most common assumption is the normality of empirical distribution of returns. In the area of financial time series, it is possible to mention certain specific characteristics like significant level of autocorrelation, leptokurtosis, clustering, heavy tails in empirical distributions, etc. These features do not allow for using models based

[1] University of Economics in Katowice. E-mail: dominik.krezolek@ue.katowice.pl.

on the normality assumption, therefore scientists have to seek for new theoretical solutions to cope with this problem.

The main goal of the paper is the application of the new family of risk measures, called GlueVaR risk measures, to investment risk assessment. The specific type of risk is considered – extreme risk (catastrophic risk), which is related to events with low probability of occurrence, but if they do take place, they can produce large losses (Jajuga 2009). This type of risk is often defined as Low Frequency, High Severity (LFHS), but the precise definition may be represented as in Table 1.

Table 1. Location of risks

Loss	Low probability	High probability
Small	-	regular risk
Large	extreme risk	-

This definition explains that extreme risk is related to its negative perception, where the result of investment generates losses. Theoretical methods used for modelling and examining extreme risk include two popular approaches. The first one is based on the analysis of the distribution of maxima described by the Generalized Extreme Value Theory, and the second one is based on the peaks over threshold (Generalized Pareto Distribution). Extreme risk analysed in this article should be understood more generally. Such risk is considered as related to the event whose probability of occurrence is significantly different from the expected value of the empirical distribution (such models covering these kind of phenomena are within the family of heavy-tailed distributions).

2. Properties of the risk measure

At the very beginning it is necessary to define the measure of risk. Let $\mathbb{X}$ be the set of all random variables defined for a given probability space $(\Omega, \mathcal{A}, P)$. A risk measure ρ is a mapping from $\mathbb{X}$ to $\mathbb{R}$:

$$\mathbb{X} \to \rho(X) \in \mathbb{R}$$

Therefore, $\rho(X)$ is defined as a real value for each $X \in \mathbb{X}$. If the risk measure is defined, certain properties of this measure have to be shown. In 1999 Artzner *et al*. (Artzner *et al*., 1999) presented some axioms describing appropriate risk measure:

- positive homogeneity: $\rho(\lambda X) = \lambda\rho(X)$
- subadditivity: $\rho(X + Y) \leq \rho(X) + \rho(Y)$
- monotonicity: $X \leq Y \Rightarrow \rho(X) \leq \rho(Y)$
- translation invariance: $\rho\left(X + \alpha R_{free}\right) = \rho(X) - \alpha$

These axioms define a coherent risk measure. The assumptions of positive homogeneity and subadditivity are often replaced by the assumption of convexity:

$$\rho[\lambda X + (1-\lambda)Y] \leq \lambda\rho(X) + (1-\lambda)\rho(Y) \text{for}\ ,\ 0 \leq \lambda \leq 1.$$

Taking into account the investor's point of view, all these axioms are of great importance, but the assumption of subadditivity deserves particular attention. Subadditivity means that the risk of portfolio is equal or lower than the sum of its individual risks. Considering the definition of subadditivity one may link it with diversification, which means that the cumulated risks of individual portfolios cannot be greater than the total risk of the investments. Therefore, the good risk measure should hold these four axioms together.

One of the most popular tools for calculating risk are Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR). The advantage of CVaR compared to VaR is that the first one holds the subadditivity assumption and measures the average level of loss in the most adverse cases whereas VaR shows only the minimum loss. The value of CVaR is usually higher than the value of VaR and the selection of risk measure depends on the investor's attitude towards risk.

3. GlueVaR risk measure

The selection of adequate measure of risk and, consequently, the level of risk, is based on underlying investor's attitude to risk. Belles-Sempera*et al*. (Belles-Sempera *et al*., 2014) introduced a new family of risk measures based on Value-at-Risk and Conditional Value-at-Risk namely the GlueVaR risk measure. For a fixed confidence level, the family of the GlueVaR risk measures contains risk measures which lies between the values of VaR and CVaR. Thus, they reflect a particular investor's attitude towards risk. The family of the GlueVaR risk measures is expressed in terms of distortion function and Choquet integral[2]. The distortion function of the GlueVaR risk measure is defined by four-parameter function of the form:

$$\eta_{\gamma_2,\gamma_1}^{m_1,m_2} = \begin{cases} \dfrac{m_1}{1-\gamma_2}u & if \quad 0 \leq u < 1-\gamma_2 \\ m_1 + \dfrac{m_2 - m_1}{\gamma_2 - \gamma_1}[u - (1-\gamma_1)] & if \quad 1-\gamma_2 \leq u < 1-\gamma_1 \\ 1 & if \quad 1-\gamma_1 \leq u \leq 1 \end{cases}$$

where γ_1, γ_2 define confidence levels such that $\gamma_1, \gamma_2 \in [0,1]$ and $\gamma_1 \leq \gamma_2$. Two additional parameters m_1 and m_2 are defined as hits of distortion function such that $m_1 \in [0,1]$and $m_2 \in [m_1, 0]$.

[2] For more details see Yaari (1987), Choquet (1954), Denneberg (1994).

The GlueVaR risk measure can be expressed in terms of the Choquet integral using the formula:

$$GlueVaR_{\gamma_2,\gamma_1}^{m_1,m_2}(X) = \int X\, d\mu = \int X\, d(\eta_{\gamma_2,\gamma_1}^{m_1,m_2} \circ P)$$

An interesting feature of the GlueVaR risk measure is that it can be expressed as a linear combination of standard risk measures: VaR at the level γ_1, CVaR at the level γ_1 and CVaR at the level γ_2 under the assumption that $0 < \gamma_1 \leq \gamma_2 < 1$):

$$GlueVaR_{\gamma_1,\gamma_2}^{m_1,m_2}(X) = w_1 CVaR_{\gamma_2} + w_2 CVaR_{\gamma_1} + w_3 VaR_{\gamma_1}$$

where weights w_1, w_2 and w_3 are calculated as below:

$$\begin{cases} w_1 = m_1 - \dfrac{(m_2 - m_1)(1-\gamma_2)}{\gamma_2 - \gamma_1} \\ w_2 = \dfrac{m_2 - m_1}{\gamma_2 - \gamma_1}(1-\gamma_1) \\ w_3 = 1 - w_1 - w_2 = 1 - m_2 \end{cases}$$

As discussed by Belles-Sempera *et al*., the pairs (m_1, m_2) representing hits of a distortion function of GlueVaR and (w_1, w_2) representing weights given to CVaR at the levels γ_2 and γ_1 respectively are linearly related to each other. The relationship can be expressed in terms of the theory of matrices. Therefore, the relation is as follow:

$$\begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \mathrm{H} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \text{ and } \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \mathrm{H}^{-1} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

where matrices H and H^{-1} are of the form:

$$\mathrm{H} = \begin{bmatrix} 1 & \frac{1-\gamma_2}{1-\gamma_1} \\ 1 & 1 \end{bmatrix} \text{ and } \mathrm{H}^{-1} = \begin{bmatrix} \frac{1-\gamma_1}{\gamma_2-\gamma_1} & \frac{\gamma_2-1}{\gamma_2-\gamma_1} \\ \frac{\gamma_1-1}{\gamma_2-\gamma_1} & \frac{1-\gamma_1}{\gamma_2-\gamma_1} \end{bmatrix}$$

For a given parameters w_1 and w_2 we can show special cases of the GlueVaR risk measures:

- if $w_1 = 0$ and $w_2 = 0$ then the GlueVaR risk measures reduce to Value-at-Risk at the level γ_1;
- if $w_1 = 0$ and $w_2 = 1$ then the GlueVaR risk measures reduce to Conditional Value-at-Risk at the level γ_1;
- if $w_1 = 1$ and $w_2 = 0$ then the GlueVaR risk measures reduce to Conditional Value-at-Risk at the level γ_2.

The linear combination of the GlueVaR risk measure allows for defining a particular investor in terms on his attitude towards risk (Belles-Sampera *et al*. 2015). If an investor selects weights $(w_1, w_2) = (1,0)$ then he represents highly conservative attitude towards risk. For the pair $(w_1, w_2) = (0,1)$ he can be defined as conservative. And finally, if he selects weights $(w_1, w_2) = (0,0)$ he is less conservative towards risk. Hence, for given confidence levels γ_1 and γ_2, and for certain levels of weights w_1 and w_2 reflecting the investor's attitude towards risk, the appropriate risk measure within the new family of the GlueVaR risk measures can be selected.

An interesting and attractive feature of the GlueVaR risk measure is that there exist explicit formulas for the most popular probability distributions describing returns. For a normally distributed random variable X, any GlueVaR risk measure can be calculated as:

$$GlueVaR_{\gamma_2,\gamma_1}^{m_1,m_2}(X) = \mu + \sigma q_{\gamma_1}[1-m_2] + \sigma\frac{m_2-m_1}{\gamma_2-\gamma_1}\left[\phi\left(q_{\gamma_1}\right)-\phi\left(q_{\gamma_2}\right)\right] + \sigma\frac{m_1}{1-\gamma_2}\phi(q_{\gamma_2})$$

where $X \sim N(\mu, \sigma)$, q_{γ_1}, q_{γ_2} represent γ_1 −quantile and γ_2 −quantile of standard normal distribution respectively, and $\phi(\cdot)$ represents the density of standard normal distribution.

If a random variable X is described by t −Student distribution, the expression for the GlueVaR risk measure is of the form:

$$GlueVaR_{\gamma_2,\gamma_1}^{m_1,m_2}(X) = \mu + \sigma[\left(\frac{m_1}{1-\gamma_2}-\frac{m_2-m_1}{\gamma_2-\gamma_1}\right)f\left(t_{\gamma_2}\right)\left(\frac{k+t_{\gamma_2}^2}{k-1}\right) + \frac{m_2-m_1}{\gamma_2-\gamma_1}f\left(t_{\gamma_1}\right)\left(\frac{k+t_{\gamma_1}^2}{k-1}\right)+(1-m_2)t_{\gamma_1}]$$

where $t_{\gamma_1}, t_{\gamma_2}$ represent γ_1 −quantile and γ_2 −quantile of t −Student distribution respectively, k represents degrees of freedom and $f(\cdot)$ represents the density function of t −Student distribution.

4. Empirical analysis on the non-ferrous metals market

The GlueVaR risk measure is applied to assess the risk of investments on the non-ferrous metals market. Due to financial and economic crises observed in the first decade of the 21st century, investors have been forced to search other

possibilities to invest capital, which would generate positive returns (Krężołek 2012). The analysis is based on a daily log-returns of spot closing prices of certain non-ferrous metals quoted on the London Metal Exchange from January 2008 to June 2015. The set of assets includes ALUMINIUM, COPPER, LEAD, NICKEL, TIN and ZINC. The quantile-based risk measures such as VaR, CVaR and GlueVaR have been calculated for quantile 0.952 and 0.996, using empirical and theoretical distributions: normal, t −Student and α −stable. All parameters for theoretical distributions have been calculated using Maximum Likelihood Method. Figures 1-2 present the levels of prices and log-returns for COPPER and ZINC.

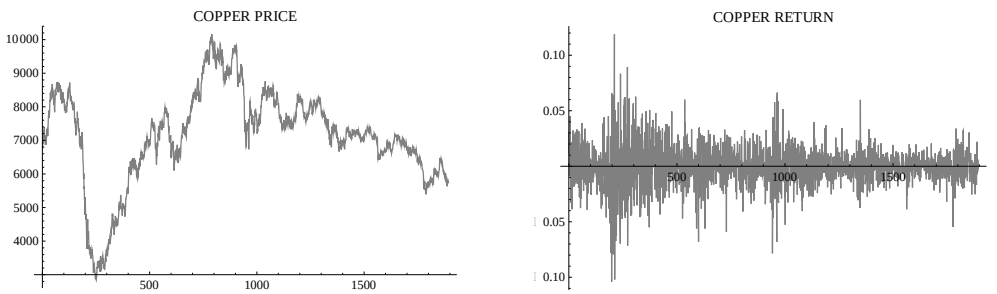


Figure 1. Time series of prices (left) and log returns (right) –COPPER

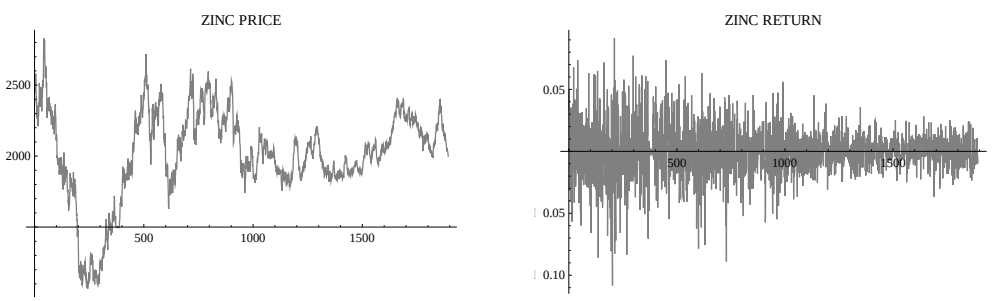


Figure 2. Time series of prices (left) and log returns (right) –ZINC

Figures 1-2 show significant disturbances in price levels, which affect the volatility in log-returns. If log-returns are considered, we can find some specific characteristics of time series, which are very typical for financial assets: clustering of variance, high volatility, long memory effect, etc. In Table 2 certain descriptive statistics of log-returns are presented.

Table 1. Descriptive statistics of log-returns – all metals

Metal/ Statistics	MEAN	STANDARD DEVIATION	KURTOSIS	SKEWNESS	MIN	MAX
ALUMINIUM	-0.00019	0.01492	1.51218	-0.14286	-0.07437	0.05913
COPPER	-0.00008	0.01877	3.80788	-0.11059	-0.10400	0.11880
LEAD	-0.00020	0.02333	3.25932	-0.15558	-0.12850	0.12675
NICKEL	-0.00042	0.02388	3.31503	0.03299	-0.13605	0.13060
TIN	-0.00009	0.01999	4.94503	-0.09549	-0.11435	0.14253
ZINC	-0.00009	0.02047	2.26455	-0.12307	-0.10832	0.09135

The results shown in Table 1 indicate that investments in all analysed metals generate losses. The lowest values of standard deviation are for ALUMINIUM and COPPER. All metals, except NICKEL, are negatively skewed. Moreover, all analysed assets are leptokurtic. This may lead one to assume that empirical distributions are not normal. Goodness-of-fit tests (Anderson-Darling and Cramer-von Misses) have confirmed this hypothesis of non-normality[3]. As an alternative, the t −Student and α −stable distributions have been fitted to the data. The results of estimated parameters are shown in Tables 2-3.

Table 2. Estimated parameters of t −Student distribution*

Metal/ Parameters	$\hat{\mu}$	$\hat{\sigma}$	$\hat{k}$**
ALUMINIUM	-0.00017	0.01250	6.60991
COPPER	0.00007	0.01260	3.28272
LEAD	0.00010	0.01657	3.75685
NICKEL	-0.00058	0.01758	4.18096
TIN	0.00067	0.01253	2.88158
ZINC	-0.00005	0.01497	3.95568

*Maximum Likelihood estimates
**Degrees of freedom

Table 3. Estimated parameters of α −stable distribution*

Metal/ Parameters	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\mu}$	$\hat{\sigma}$
ALUMINIUM	1.84137	-0.00641	-0.00009	0.00957
COPPER	1.62745	-0.03151	-0.00006	0.01023
LEAD	1.66834	-0.07549	-0.00030	0.01315
NICKEL	1.71689	0.05195	-0.00036	0.01387
TIN	1.56374	-0.21853	-0.00063	0.01027
ZINC	1.68205	0.02343	0.00002	0.01187

*Maximum Likelihood estimates

[3] Due to the length of the paper, some results which are not directly related to the topic are omitted.

The parameters of $\alpha-$stable distribution allow for indicating additional characteristics of empirical distributions which are not exhibited if normal distribution is considered. The parameter α describes the thickness of tails in empirical distribution, and $\alpha \in (0,2]$. If $alpha < 2$ then variance of the distribution is infinite. The lower values of $alpha$, the thicker tails of empirical distributions. The heaviest tails are for TIN and COPPER, which means that the probability of occurrence of extreme returns is higher than for other metals. If $\alpha = 2$, then the variance is infinite. If $\alpha < 1$ even the mean is infinite. Remaining parameters describe asymmetry ($\beta \in [-1,1]$), location ($\mu \in \mathbb{R}$) and scale of the distribution ($\sigma > 0$).

The main goal of this analysis is to assess the risk using quantile-based risk measures. Assuming two confidence levels $\gamma_1 = 0.952$ and $\gamma_2 = 0.996$ and the set of weights $w = \{w_1, w_2, w_3\}$, three risk measures VaR_α, $CVaR_\alpha$ and $CVaR_\beta$, have been calculated. The confidence level $\gamma_1 = 0.952$ denotes that the extreme event appears twelve times per year4, and $\gamma_1 = 0.996$ denotes that the extreme event appears one time per year. Taking into account the set of weights, seven scenarios have been discussed: $S = \{s_1, s_2, \dots, s_7\}$. The results for ALUMINIUM and ZINC are shown in tables 4-5.

Table 4. Estimated GlueVaR risk measure for ALUMINIUM

	Scenarios						
	s_1	s_2	s_3	s_4	s_5	s_6	s_7
w_1	100.00%	0.00%	0.00%	33.33%	33.33%	66.67%	50.00%
w_2	0.00%	100.00%	0.00%	33.33%	66.67%	33.33%	50.00%
w_3	100.00%	0.00%	0.00%	33.33%	0.00%	0.00%	0.00%
	Distributions						
Empirical distribution	0.04987	0.03253	0.02469	0.03570	0.03831	0.04409	0.04120
Normal distribution	0.03982	0.02972	**0.02480**	0.03145	0.03309	0.03645	0.03477
t-Student distribution	**0.05924**	**0.03249**	0.02397	**0.03857**	**0.04141**	**0.05032**	**0.04586**
$\alpha-$stable distribution	0.08113	0.03458	0.02317	0.04629	0.05010	0.06561	0.05786

[4] One year is understood as 250 days of trading.

Table 5. Estimated GlueVaR risk measure for ZINC

Scenarios							
	s_1	s_2	s_3	s_4	s_5	s_6	s_7
w_1	100.00%	0.00%	0.00%	33.33%	33.33%	66.67%	50.00%
w_2	0.00%	100.00%	0.00%	33.33%	66.67%	33.33%	50.00%
w_3	100.00%	0.00%	0.00%	33.33%	0.00%	0.00%	0.00%
Distributions							
Empirical distribution	0.07289	0.04775	0.03551	0.05205	0.05613	0.06451	0.06032
Normal distribution	**0.05831**	0.04292	**0.03487**	0.04537	0.04805	**0.05318**	0.05061
t-Student distribution	0.09685	**0.04628**	0.03124	0.05812	0.06314	0.07999	0.07156
α −stable distribution	0.09452	0.04502	0.03145	**0.05699**	**0.06152**	0.07802	**0.06977**

The values in bold in Tables 4-5 represent the estimates of theoretical risk measures closest to empirical ones. As we can find, the closest values are mainly for the heavy-tailed distributions. This finding covers all analysed metals. The results obtained for scenarios s_1, s_2, and s_3 represent the GlueVaR risk measure equal to $CVaR_{\gamma_2}$, $CVaR_{\gamma_1}$ and VaR_{γ_1}, respectively. In scenario s_4 we give equal weights to all components of GlueVaR. Scenario s_5 gives higher weight to $CVaR_{\gamma_1}$ and lower to $CVaR_{\gamma_2}$. On the contrary, scenario s_6 - higher weight to $CVaR_{\gamma_2}$ and lower to $CVaR_1$. And finally, scenario s_7 gives equal weights to $CVaR_{\gamma_1}$ and $CVaR_{\gamma_2}$. The location of scenarios in a two-dimensional space of weights is presented in Figure 3.

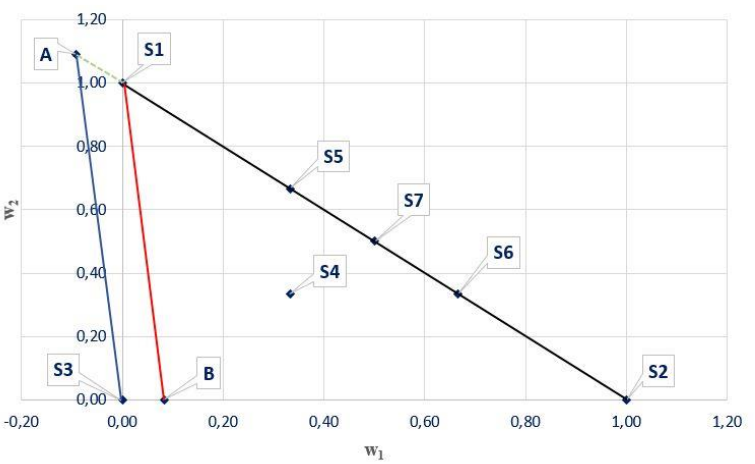


Figure 3. Location of scenarios within the area of feasible weights for GlueVaR risk measures

As discussed in the theoretical part of this paper, the GlueVaR risk measure is associated with confidence levels and weights given to VaR and CVaR. The confidence levels reflect the probability of the occurrence of some extreme events, and weights reflect how much these events are important for a particular investor. To hold the assumption of subadditivity for the GlueVaR risk measure, the weight corresponding to non-subadditive risk component of GlueVaR (i.e. VaR) should meet the relation that$w_3 = 0$. Belles-Sampera *et al.* showed that GlueVaR is subadditive if both weights w_1 and $w_2$belong to the area delimited by the triangle $S1BS2$, especially if they lie on the line segment in a coordinate system described by points: $A = (w_1, w_2) = (\frac{\beta-1}{\beta-\alpha}, \frac{1-\alpha}{\beta-\alpha})$ and $S2 = (w_1, w_2) = (1,0)$ for fixed values of γ_1 and γ_2 $(0 < \gamma_1 \leq \gamma_2 < 1)$. Moreover, the position of a particular point on this line represents the investor's attitude towards risk. The nearer to the point A, the less conservative attitude towards risk. For example, if scenarios s_5 and s_6 are of interest, the values of the GlueVaR risk measure are presented in figure 4.

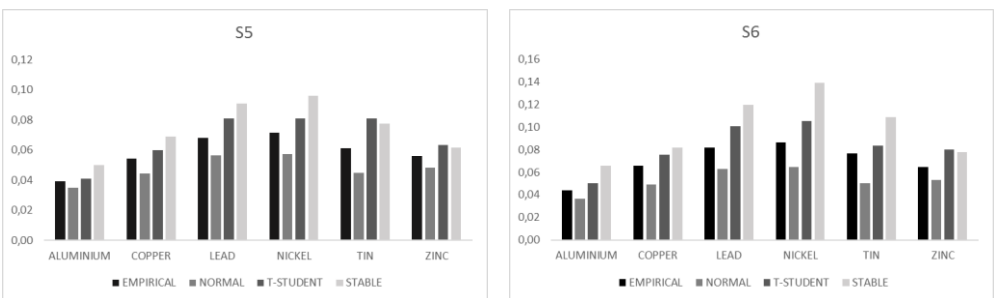


Figure 4. GlueVaR risk measure for scenario s_5 and s_6

The analysis of risk measure provided an interesting conclusion in terms of the use of theoretical distributions. Taking into account fixed confidence levels, despite the fact that the weights are given for each risk measure, the relationship between risk measures and theoretical distributions are presented in Table 6.

Table 6. Relationship between risk measures and theoretical distributions

Theoretical distribution/Risk measure	VaR_{γ_1}	$CVaR_{\gamma_1}$	$CVaR_{\gamma_2}$	$GlueVaR_{\gamma_2,\gamma_1}$
Normal distribution	overestimated	underestimated	underestimated	underestimated
t −Student distribution	underestimated	underestimated	overestimated	overestimated
α −stable distribution	underestimated	underestimated	overestimated	overestimated

The information contained in Table 6 indicates that the normal distribution usually overestimates the value of VaR at the level γ_1 and underestimates remaining risk measures at the levels γ_1 and γ_2. On the other hand, if fat-tailed distributions are considered, the values of VaR and CVaR at the level γ_1 are usually underestimated, while the remaining risk measures at the levels γ_1 and γ_2 are overestimated.

5. Conclusions

In this paper the new family of risk measures, called GlueVaR, has been applied to risk measurement on the non-ferrous metals metal market. This area of investment is not very popular within researchers, although it is a very attractive alternative to classical investments areas (i.e. stocks, exchange rates, etc.). As presented in this paper, some tools for risk assessment used on financial markets can also be used effectively on alternative markets. The methodology of the GlueVaR risk measure is directly related to popular quantile-based risk measures: Value-at-Risk and Conditional Value-at-Risk. These two risk measures determine the value of loss of extreme events. The use of VaR as a risk measure has been impaired due to the failure to meet the assumption of subadditivity. As mentioned before, risk measures such as CVaR and GlueVaR do not suffer such disadvantage. An important feature of the family of the GlueVaR risk measures is that it can be defined as a linear combination of standard risk measures VaR and CVaR for a given confidence levels and for given weights. Taking into account the investor's point of view, the confidence level corresponds to the probability of occurrence of some catastrophic event whereas the weights indicate how such an event is important for the investor. Therefore, a particular investor is able to decide consciously about the acceptable level of risk.

The analysis conducted in this paper is based on both empirical and theoretical distributions (normal, $t-$Student and $\alpha-$stable). The selection of distribution was based on the characteristics of log-returns of the analysed prices of metals. The results show that if the probability of unwanted event is not very low, then the corresponding risk measure should be calculated using normal distribution. Otherwise the fat-tailed distributions are more appropriate. In conclusion, one can say that the family of the GlueVaR risk measures is an attractive and effective tool for risk assessment. This feature results from the subadditivity assumption held for the GlueVaR and from the possibility of considering an individual investor's attitude towards risk. Compared to classical measures, the most useful feature of the proposed new risk measures is that for a particular investor it is possible to implicitly define the set of adverse events and determine the importance of such events. This advantage enables taking into account an individual investor's attitudes towards risk.

REFERENCES

ARTZNER, P., DELBAEN, F., EBER, J-M., HEAT, D., (1999). Coherent Measures of Risk, Mathematical Finance, Vol. 9, No. 3, pp. 203–228.

BELLES-SAMPERA, J., GUILLÉN, M., SANTOLINO, M., (2014). Beyond Value-at-Risk: GlueVaR Distortion Risk Measures, Risk Analysis, Vol. 34, No. 1, pp. 121–134.

BELLES-SAMPERA, J., GUILLÉN, M., SANTOLINO, M., (2015). What attitudes to risk underlie distortion risk measure choice?, UB Riskcenter Working Paper Series, Working paper 2015/05, Research Group on Risk in Insurance and Finance, University of Barcelona.

CHOQUET, G., (1954). Theory of Capatities, Annales de l'Institute Fourier, No. 5, pp. 131–295.

DENNENBERG, D., (1994). Non-Additive Measure and Integral, Dordrecht: Kluwer Academic Publischer.

JAJUGA, K., (2009). Zarządzanie ryzykiem [Risk management], Polskie Wydawnictwo Naukowe PWN.

KRĘŻOŁEK, D., (2012). Non-Classical Measures of Investment Risk on the Market of Precious Non-Ferrous Metals Using the Methodology of Stable Distributions, Dynamic Econometric Models, Vol. 12/2012, pp. 89–104.

MCNEIL, A., FREY, R., EMBRECHTS, P., (2005). Quantitative Risk Management: Concepts, Techniques and Tools, New York: Princeton Series in Finance, Princeton University Press.

ROCKAFELLAR, R. T., URYASEV, S., (2002). Optimization of Conditional Value-at-Risk, Journal of Risk, No. 2, pp. 21–41.

SZEGÖ, G., (2002). Measures of risk, Journal of Banking & Finance, No. 26, pp. 1253–1272.

WANG, S. S., (1996). Premium Calculations by Transforming the Layer Premium Density, „ASTIN Bull", Vol. 26, No. 1, pp. 71–92.

YAARI, M. E., (1987). The Dual Theory of Choice under Risk, Econometrica, Vol. 55, Issue 1, pp. 95–115.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 317–330

EXAMINING SIMILARITIES IN TIME ALLOCATION AMONGST EUROPEAN COUNTRIES

Marta Hozer-Koćmiel[1], **Christian Lis**[2]

ABSTRACT

The aim of the article is to analyse the similarities between the selected European countries in terms of time allocation. Time allocation has been defined as the daily distribution of time to various activities. Professional work time, domestic work time and leisure time are the most important for the economic approach. It has been proved that there are coherent groups of countries with similar structure of time allocation. The taxonomic methods used in order to verify the thesis included: cluster analysis, k-means method, generalised distance measure GDM and interval taxonomic method TMI. The analysis was performed on the basis of HETUS data.

Key words: time allocation, cluster analysis, k-means method, generalised distance measure GDM, interval taxonomic method TMI, HETUS survey.

1. Introduction

The previous study of the authors was dedicated to the classification of the Baltic Sea Region countries by the time spent on particular household production activities. It was shown that there was a similarity in the quantity and quality of time used on household work in the Scandinavian countries and Germany, and in the Baltic republics and Poland (Hozer-Koćmiel, Lis 2015). This article extends the research area from a single activity to all activities of the time budget, extends the research group including other European countries and uses not one but several taxonomic methods.

The issue of time allocation is not very popular, but there are economic theories that take it into account. The most important of these is the Becker's Theory of the Allocation of Time (1965), which introduced the concept of the total income significantly exceeding the monetary income. The difference is

[1] University of Szczecin, Faculty of Economics and Management, Department of Statistics. E-mail: mhk@wneiz.pl.
[2] University of Szczecin, Faculty of Economics and Management, Department of Statistics. E-mail: chrislis@wneiz.pl.

constituted by the foregone earnings directly related to the use of the time factor. Becker recommends the economists to devote more attention to time allocation and its efficiency.

The Commission on the Measurement of Economic Performance and Social Progress by A. Sen, J. Stiglitz and J.P. Fitoussi (2010) is another, broader approach to production, which takes into account non-market area. It has been noted that there is a need to introduce new methods of measurement for such categories as welfare, quality of life and sustainability of economic development. It is recommended to take greater account of the household perspective and extend the measurement of income to include the value of unpaid work of households.

Levy Economics Institute in the USA has proposed a two-dimensional approach to the study of poverty, taking into account its material and time aspect. They introduced an adjusted poverty threshold measure LIMTIP (the Levy Institute's Measure of Time and Income Poverty), which differs from the standard indicator by including the assessed deficit of time. Empirical studies show that the official poverty is significantly deeper than the one indicated by the standard measure, which does not take into account the aspect of time. One can distinguish the individuals and households that are not classified as poor, but their time allocation leads them to poverty (Antonopoulos et al. 2012).

The theory of the economics of happiness by Easterlin (2008) assumes that satisfaction in life is derived from three main areas, namely professional work, family life and health. It notes that happiness depends on the material situation, but it is not a linear relation. The increase in wealth does not result in the proportionate increase in happiness. For those who want to increase the level of happiness Easterlin recommends devoting more time to family life and health.

Other economic concepts regarding time allocation of the population include the theory of household satellite accounts (EC 2003), the theory of the triangle of the human economy by Pietila (1997), the theory of the economics of care (Folbre 1994) and the concept of the distribution of working time DCP by Hozer-Koćmiel (2008).

The aim of this study is to investigate the similarity of the selected European countries in terms of time allocation of the population. A thesis is formulated that there are groups of countries that show strong similarity in setting the time budget of the population.

2. Methodology and data

The taxonomic methods used for the classification of countries in terms of time allocation include cluster analysis methods such as the agglomerative procedure based on single linkage and Euclidean distances or the *k*-means method (Pociecha et al. 1988, Jajuga 1993), and other taxonomic methods such as interval taxonomic method TMI (Strahl, Walesiak 1997; Lis 2013), and generalised

distance measure GDM (Walesiak 2000, 2011). The normalization of variables for determining TMI was performed using zeroed unitarization.

The analysis included the working population by gender from 15 European countries: Belgium, Estonia, Finland, France, Spain, Lithuania, Latvia, Germany, Norway, Poland, Slovenia, Sweden, Hungary, United Kingdom and Italy. They were investigated in respect of the following variables of time allocation of the population in minutes per day: 1. Professional work 2. Domestic work 3. Leisure 4. Study, 5. Travel, 6. Sleep 7. Other physiological functions.

The data are derived from a survey on the population time budget HETUS - Harmonised European Time Use Survey 2004 (see Table 1).

Table 1. The average duration of particular activities in minutes per day in the selected European countries in 2004

WOMEN

	Belgium	Estonia	Finland	France	Germany	Hungary	Italy	Latvia	Lithuania	Norway	Poland	Slovenia	Spain	Sweden	UK
Professional work	228	248	247	273	213	275	275	337	349	208	277	250	285	235	234
Study	5	5	13	2	19	8	4	9	6	18	9	13	12	10	12
Domestic work	232	244	201	212	191	234	231	188	204	206	238	264	209	212	208
Leisure	231	240	278	190	289	223	197	193	185	322	223	231	214	267	261
Travel	90	75	76	66	87	62	88	86	67	77	70	69	82	88	93
Sleep	496	503	502	521	491	498	480	501	493	487	488	492	491	485	505
Other physiological functions	156	126	122	176	151	141	164	126	136	122	134	122	148	143	127

MEN

	Belgium	Estonia	Finland	France	Germany	Hungary	Italy	Latvia	Lithuania	Norway	Poland	Slovenia	Spain	Sweden	UK
Professional work	298	295	324	339	294	320	370	396	388	286	362	311	363	310	333
Study	5	5	8	3	11	5	3	5	3	10	8	9	8	7	9
Domestic work	135	140	119	110	112	129	70	86	99	132	113	144	80	143	114
Leisure	263	287	306	234	311	277	246	238	242	337	269	292	260	291	281
Travel	103	80	77	72	91	70	100	91	77	83	75	74	83	92	96
Sleep	481	502	492	506	480	488	478	496	488	473	479	486	495	472	491
Other physiological functions	155	131	115	176	141	150	172	128	143	118	134	127	151	125	115

Source: own study based on the data from HETUS survey.

3. The results of grouping the European countries by time allocation of the population

3.1. Tree diagrams

For the active part of the day, most of the time is spent on professional work, domestic work and leisure. The study was conducted for the employed people, therefore the duration of professional work was relatively long. If the analysis had been performed for all people, both employed and unemployed, most of the daily time would have been spent on domestic work, then professional work and leisure.

Some significant differences in time allocation for men and women were visible. Women spent less time than men on professional work and more time on domestic work on average. They had less free time and studied more than men.

On the basis of the tree diagram it can be stated that the countries that are most similar to each other in terms of time allocation for women were Poland and Hungary. The linkage based on the Euclidean distance was the shortest in this case. The next countries that joined the group were Spain, Italy, Slovenia and Estonia. For this group, professional work time and domestic work time was clearly longer than in other countries, whereas leisure time was reduced. These are the countries with relatively traditional division of responsibilities in the household compared to other European countries.

A strong similarity of the time use structure for women was also observed in Sweden, Finland, United Kingdom and Germany. For Germany, the distance of the linkage was significantly longer, and therefore the similarity of time allocation in relation to the Scandinavian countries was smaller. The group, unlike the previously described one, was characterized by relatively short professional work time and domestic work time, and considerable leisure time. The second group was comprised of more economically developed countries, the so-called 'old EU', including countries that were the world leaders in terms of equality between women and men (Scandinavia).

The countries that showed a clear distinction in relation to all other countries included Lithuania and Latvia. They represent a separate small group with exceptionally long professional work time, short domestic work time and short leisure time. They are less economically developed compared to the aforementioned representatives of the 'old EU'. A few decades ago they underwent the economic transformation. The situation on their labour market is more difficult than in western countries, which is probably the reason why the average working time for women is so long there.

In order to observe similarities of time allocation in selected EU countries and the process of agglomeration, the agglomerative procedure (joining) based on the single linkage and Euclidean distances was used and presented as a tree diagram below.

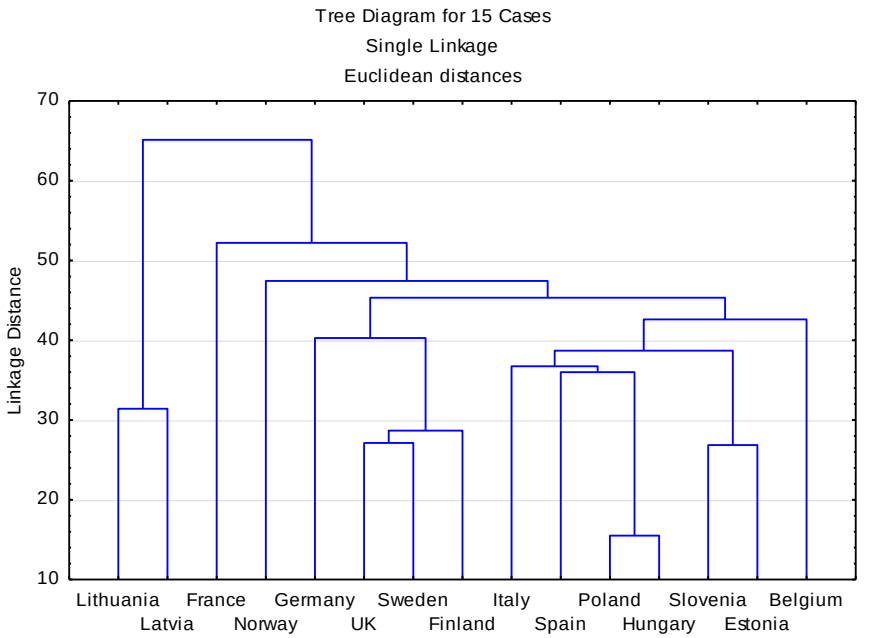


Figure. 1. Tree diagram of time allocation for women in the selected EU countries
Source: own study based on the data from HETUS survey.

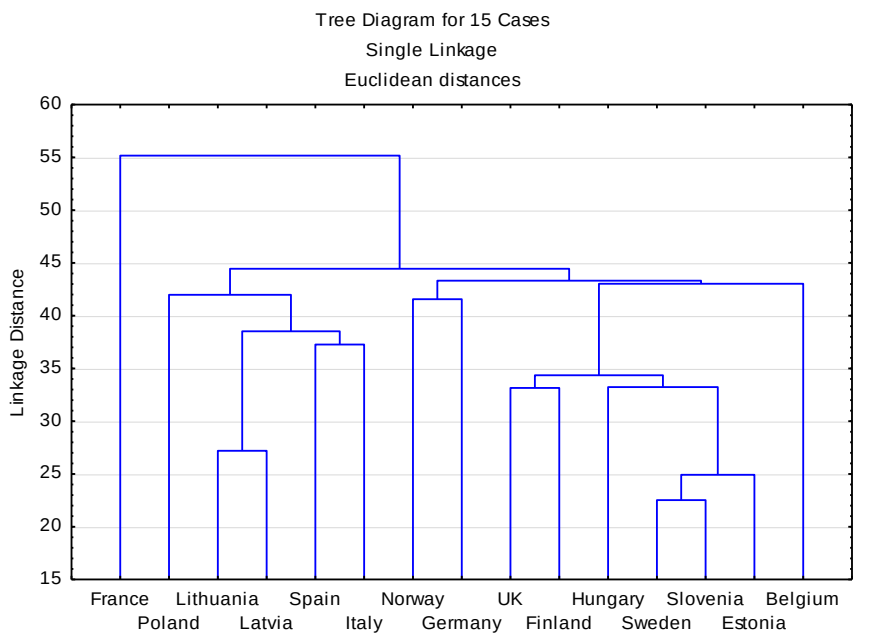


Figure 2. Tree diagram of time allocation for men in the selected EU countries
Source: own study based on the data from HETUS survey.

For men, Sweden, Slovenia and Estonia were the most similar in terms of time allocation were Sweden, Slovenia and Estonia, followed by the United Kingdom, Finland and Hungary. They were characterized by a relatively short professional work time, long domestic work time and long leisure time. In the previous work (Hozer-Koćmiel, Lis 2015) it was observed that men in the Scandinavian countries, which are part of this group, spend significantly more time on household activities, particularly on childcare.

Lithuania and Latvia, followed by Spain, Italy and Poland, also formed a distinct group for men. These are the countries that have undergone a transformation towards market economy (Lithuania, Latvia, Poland) and countries firmly embedded in the Mediterranean culture (Spain and Italy). The common denominator for them is the popularity of traditional roles in the household. As a consequence, men in this group worked a long time, had relatively long leisure time, but did not engage considerably in housework.

3.2. Cluster analysis

In the next stage of the research, the *k*-mean method, as a method of cluster analysis, was used. As in the previous section of the paper, Euclidean distances between clusters were calculated. Three clusters were proposed a priori to distinguish between women and men.

The results of grouping of the EU countries in terms of time allocation for employed men and women are presented in Table 2.

Table 2. Cluster analysis of the EU countries in terms of time allocation for women and men in 2004

WOMEN		
Cluster 1	Cluster 2	Cluster 3
Latvia Lithuania	Finland Sweden Norway Germany UK	Estonia Hungary Poland Slovenia Spain Italy France Belgium
Prof.work = 343.2	Prof.work = 227.3	Prof.work = 264.0
Study = 7.7	Study = 14.4	Study = 7.2
Dom.work = 195.8	Dom.work = 203.6	Dom.work = 233.0
Sleep = 497.1	Sleep = 494.0	Sleep = 496.2
Leisure = 189.0	Leisure = 283.5	Leisure = 218.6
Other physio. = 131.0	Other physio.= 132.8	Other physio. = 145.9
Travel = 76.3	Travel = 84.1	Travel = 75.3

Table 2. Cluster analysis of the EU countries in terms of time allocation for women and men in 2004 (cont.)

MEN		
Cluster 1	Cluster 2	Cluster 3
Finland Sweden UK Belgium Estonia Hungary Slovenia	Germany Norway	Latvia Lithuania Poland Spain Italy France
Prof.work = 312.9	Prof.work = 290.1	Prof.work = 369.6
Study = 6.9	Study = 10.5	Study = 5.0
Dom.work = 132.1	Dom.work = 122.2	Dom.work = 93.0
Sleep = 487.6	Sleep = 476.5	Sleep = 490.3
Leisure = 285.3	Leisure = 324.1	Leisure = 248.1
Other physio. =131.0	Other physio. = 129.3	Other physio = 150.7
Travel = 84.5	Travel = 86.8	Travel = 83.0

Source: own study based on the data from HETUS survey.

The grouping of the countries was carried out differently for men and women due to the fact that the division of daily time differs significantly by gender. The countries which were similar to each other were grouped together, not only according to time allocation, but also economically and culturally. For women, the first cluster consisted of Lithuania and Latvia, two of the three Baltic republics seized after World War II by the Soviet Union. The difficult economic situation, including the labour market, caused that on average employed women from this group worked longer than in other countries. This is also related to the fact that less people worked part-time there. Long professional work time contributed to the reduction of leisure time devoted to cultural life, sport, social activities or hobbies. Women living in Lithuania and Latvia devoted relatively small amount of time for domestic work.

The second cluster included the Scandinavian countries and two more developed countries of the 'old EU': Germany and the United Kingdom. As for the main features of time allocation, this group was characterized by short professional work time, average domestic work time and relatively long leisure time. The countries of the second cluster show high economic activity of women, but the average work time is shorter here than in other parts of Europe. This results from the availability of part-time and flexible forms employment, among others. It is also worth noting that the study time for women was on average twice as long as in the other two groups.

The third cluster for women gathered less wealthy countries after economic transformation and the countries of the south-west Europe. The latter are the countries of Roman culture, characterized by the Latin-based language and, in the case of Spain and Italy, more traditional division of work in the family, among others. The duration of particular daily activities turned out average there: time spent on professional work, domestic work and leisure.

For men, the clusters of the European countries were formed differently and it was harder to find a common denominator for their components. The first cluster included the Scandinavian countries, the more economically developed United Kingdom and Belgium, and three less developed eastern countries: Estonia, Hungary and Slovenia. Basic variables of the time budget for men were at the average level (professional work time and leisure time), whereas domestic work time was relatively long. The latter regularity is associated with the presence of the Scandinavian countries in the cluster in which the equality policy is the most advanced in the world.

The second small group consisted of Germany and Norway. Professional work time for men was short there, domestic work time was average (affected by the situation in Germany), whereas the leisure time was significantly longer in comparison with other countries. This structure of time allocation is optimal according to Bergmann (2014), who recommends shorter professional work time, more domestic work and work for the local community, and engaging only in activities that bring satisfaction.

The third cluster for men covered Latvia, Lithuania and Poland – Eastern Europe, and Spain, Italy, France – countries of Latin roots. This group was characterized by long professional work time, short domestic work time and short leisure time. This is related to the traditional division of roles in the household, where men focus on market activities and women on household work.

3.3. Interval taxonomic method TMI

The next part of the analysis was the classification of the countries by the interval taxonomic method TMI.

The first stage of TMI calculation is splitting the set of variables into three possible sets: stimulants, destimulants and nominants. Then, a specific measure d_{ij} for each variable from the subsets has to be calculated. The measure represents the distance to the pattern. An object that is defined as the pattern has got the best possible values for each j-th variable, which denotes the maximal value observed in the economy for stimulants (x_{max}), the minimal value for destimulants (x_{min}) and finally a specific value that is needed for nominants (x_N).

Next, distances to the best possible values are calculated according to formulae as follows:

1. Stimulants:

$$d_{ij} = \frac{x_{max\ j} - x_{ij}}{x_{max\ j} - x_{min\ j}}; \tag{1}$$

2. Destimulants:

$$d_{ij} = \frac{x_{ij} - x_{min\,j}}{x_{max\,j} - x_{min\,j}}; \tag{2}$$

3. Nominants:
   a) determined at the point x_{Nj}

$$d_{ij} = \begin{cases} 0 & if\ x_{ij} = x_{Nj}; \\ \dfrac{x_{Nj} - x_{ij}}{x_{Nj} - x_{min\,j}} & if\ x_{ij} < x_{Nj}; \\ \dfrac{x_{ij} - x_{Nj}}{x_{max\,j} - x_{Nj}} & if\ x_{ij} > x_{Nj}; \end{cases} \tag{3}$$

   b) determined in the interval $<x_{NLj}, x_{NUj}>$

$$d_{ij} = \begin{cases} 0 & if\ x_{ij} \in < x_{NLj}, x_{NUj} >; \\ \dfrac{x_{NLj} - x_{ij}}{x_{NLj} - x_{min\,j}} & if\ x_{ij} < x_{NLj}; \\ \dfrac{x_{ij} - x_{NUj}}{x_{max\,j} - x_{NUj}} & if\ x_{ij} > x_{NUj}. \end{cases} \tag{4}$$

Finally, TMI measure is calculated for each *i*-th object with the use of aggregation formula as follows:

$$TMI_i = d_i = \frac{\sum_{j=1}^{k} d_{ij}}{k}. \tag{5}$$

For each variable of time allocation, the following standards for a pattern were adopted:

- Professional work – the minimum value, 208 min. per day (W) and 286 min. per day (M)
- Study – the maximum value, 19 min. per day (W) and 11 min. per day (M),
- Domestic work – the average value, 218 min. per day (W) and 115 min. per day (M),

- Leisure – the maximum value: 322 min. per day (W) and 337 min. per day (M),
- Travel – the minimum value: 62 min. per day (W) and 70 min. per day (M),
- Sleep – the optimal level of sleep: 8 hours per day, 480 min. per day (W, M),
- Other physiological functions – the average value:140 min. per day (W) and 139 min. per day (M).

Table 3. The results of the classification of the countries based on TMI in 2004

Rank	Country	TMI women	Rank	Country	TMI men	Rank	Country	TMI total
1.	Norway	0.168	1.	Germany	0.146	1.	Norway	0.176
2.	Germany	0.274	2.	Norway	0.185	2.	Germany	0.210
3.	Sweden	0.312	3.	Slovenia	0.256	3.	Finland	0.322
4.	Finland	0.354	4.	Poland	0.281	4.	Slovenia	0.324
5.	Hungary	0.362	5.	Finland	0.290	5.	Poland	0.325
6.	Poland	0.369	6.	Hungary	0.315	6.	Hungary	0.339
7.	Slovenia	0.392	7.	Sweden	0.385	7.	Sweden	0.349
8.	Spain	0.418	8.	Estonia	0.392	8.	UK	0.411
9.	UK	0.428	9.	UK	0.395	9.	Estonia	0.430
10.	Estonia	0.468	10.	Belgium	0.443	10.	Spain	0.447
11.	Belgium	0.490	11.	Spain	0.477	11.	Belgium	0.467
12.	Lithuania	0.499	12.	Lithuania	0.513	12.	Lithuania	0.506
13.	Italy	0.533	13.	France	0.571	13.	France	0.594
14.	France	0.616	14.	Latvia	0.631	14.	Italy	0.608
15.	Latvia	0.621	15.	Italy	0.683	15.	Latvia	0.626

Source: own study based on the data from HETUS survey.

Germany and Norway bore the closet similarity with the adopted standard. For men, they were followed by: Slovenia, Poland, Finland, Hungary and Sweden. For women, the countries were exactly the same but in different order: Sweden, Finland, Hungary, Poland and Slovenia. France, Italy and Latvia showed the lowest similarity to the pattern.

TMI was also used to compare the similarity in time allocation for the selected countries in relation to each other. Time allocation for women in Poland was the most similar to that of Hungary, Slovenia, Estonia and Spain. For those countries, TMI values were the lowest, i.e. lower than 0.2. Time allocation structure for men in Poland showed the closest similarity to Slovenia, Spain, Finland and Hungary.

3.4. Generalized distance measure (GDM)

The last step of the analysis involved the generalized distance measure. GDM combines different variables in one synthetic indicator. Such an approach makes it possible to classify the selected countries by the similarity of time allocation. Unfortunately, the variables used in the analysis had different ranges. In order to make variables comparable, transformations that make them similar are performed. Such procedures are called variable normalizations. One of them, used in this study, is known as standardization.

After the process of standardization the vector of distances from the specific object (a virtual, artificial country) that is characterized by the best possible values of each variable has to be calculated. GDM is based on the idea of the generalized correlation coefficient. The generalized distance measure is calculated by the following equation (Walesiak 2000):

$$GDM = d_{ik} = \frac{1 - s_{ik}}{2} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{l=1}^{n} a_{ilj} b_{klj}}{\sqrt{\sum_{j=1}^{m} \sum_{l=1}^{n} a_{ilj}^2 \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} b_{klj}^2}} \qquad [6]$$

where:

d_{ik} (s_{ik}) – distance (similarity) measure;
$i, k, l = 1, 2, \ldots, n$ – the number of objects;
$j = 1, 2, \ldots, m$ – the number of variables.

If variables are measured on ratio or interval scale, a_{ipj}, b_{krj} are defined as:

$$a_{ipj} = x_{ij} - x_{pj} \quad \text{for } p = k, l \qquad [7]$$

$$b_{krj} = x_{kj} - x_{rj} \quad \text{for } r = i, l \qquad [8]$$

If variables are measured on ordinal scale, then a_{ipj}, b_{krj} are defined as follows:

$$a_{ipj}\left(b_{krj}\right) = \begin{cases} 1 \ \ when\ x_{ij} > x_{pj} \ \left(x_{kj} > x_{rj}\right) \\ 0 \ \ when\ x_{ij} = x_{pj} \ \left(x_{kj} = x_{rj}\right) \\ -1 \ \ when\ x_{ij} < x_{pj} \ \left(x_{kj} < x_{rj}\right) \end{cases}, \ for\ p = k, l; r = i, l \qquad [9]$$

where:
x_{ij} (x_{pj}, x_{rj}) - the i^{th} (p^{th}, r^{th}) observation of the j^{th} variable.

Similarly, as in TMI procedure, the selected countries were finally ranked for women, men and total population using the presented method. The results are listed in the table below.

Table 4. Classification of the selected countries by GDM

Rank	Country	GDM women	Rank	Country	GDM men	Rank	Country	GDM total
1.	Norway	0.015	1.	Norway	0.029	1.	Norway	0.022
2.	Germany	0.058	2.	Germany	0.032	2.	Germany	0.045
3.	Sweden	0.097	3.	Finland	0.113	3.	Finland	0.108
4.	Finland	0.102	4.	Slovenia	0.148	4.	Sweden	0.129
5.	UK	0.152	5.	Estonia	0.154	5.	Estonia	0.212
6.	Estonia	0.270	6.	Sweden	0.161	6.	UK	0.225
7.	Belgium	0.289	7.	Hungary	0.227	7.	Slovenia	0.244
8.	Slovenia	0.341	8.	UK	0.299	8.	Belgium	0.295
9.	Hungary	0.439	9.	Belgium	0.301	9.	Hungary	0.333
10.	Poland	0.450	10.	Poland	0.460	10.	Poland	0.455
11.	Spain	0.545	11.	Spain	0.521	11.	Spain	0.533
12.	France	0.598	12.	France	0.561	12.	France	0.580
13.	Italy	0.606	13.	Lithuania	0.660	13.	Italy	0.646
14.	Latvia	0.712	14.	Italy	0.687	14.	Lithuania	0.696
15.	Lithuania	0.732	15.	Latvia	0.693	15.	Latvia	0.702

Source: own calculations.

Norway and Germany ranked first and second in the classification both for women and men. It means that these countries were the most similar to the invented artificial object (country) that had the best possible structure of time allocation. Poland was ranked 10th. The last two countries were Latvia and Lithuania for women, and Italy and Latvia for men.

TMI and GDM classifications were compared using the product-moment coefficient of correlation. It is worth noting that the achieved results were coherent. The coefficient of correlation between TMI and GDM for women was equal to 0.79 and for men to 0.87, which indicates a positive and relatively strong association between the two rankings. The scatter plots of the two classifications for women and men are presented in Fig. 3.

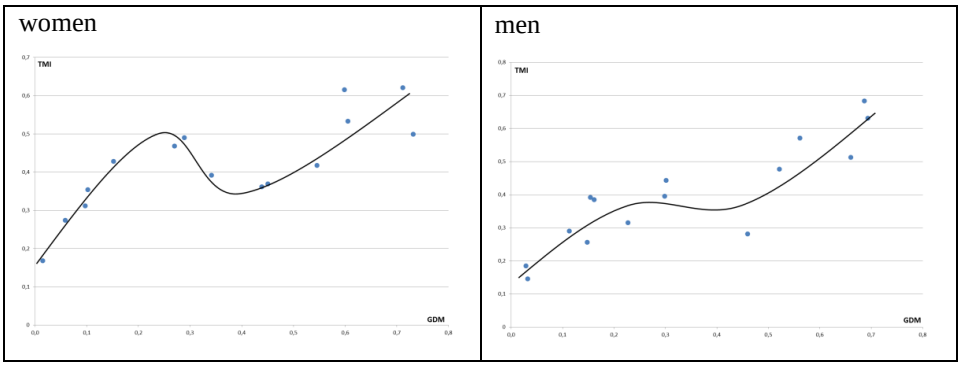


Figure 3. Relation between TMI and GDM results

Source: own work.

Interestingly, the method of calculating TMI and GDM makes those measures associated in a non-linear way. Scatter plots show that there are slight differences between TMI and GDM classification for objects (countries) in the middle of the ranking table. Summing up, despite the nonlinearity of the correlation between the results of the two aforementioned methods, they were still highly coherent.

4. Summary and directions for further research

The inclusion of the issue of time allocation in economic analyses is recommended by J. E. Stiglitz, A. Sen, J. P. Fitoussi and G. Becker. The implementation of the quantitative approach is possible based on the data from the study of time budgets of the population.

The analysis of the similarities between the selected European countries in terms of time allocation led to the conclusion that there are groups of countries that show strong similarity in setting the time budget of the population. The first group includes the 'new' European countries that has undergone economic transformation. They were characterized by a distinctly longer professional work time and shorter leisure time. The second coherent group consists of the Scandinavian countries and the more developed countries of Western Europe. In these regions, the basic variables of time allocation were opposite: relatively short professional work time and long leisure time.

The next step in the research will be to design the LIMTIP indicator for Poland, i.e. the measure which takes into account the material and time aspect of poverty.

REFERENCES

ANTONOPOULOS, R., MASTERSON, T., ZACHRIAS, A., (2012). It's about 'time': Why time deficits matter for poverty. Levy Economics Institute of Bard College Public Policy Brief, No. 126.

BECKER, G. S., (1965). A theory of the Allocation of Time. Economic Journal, September 1965.

BERGMANN, F., (2014). New Work Life: New Work New Culture: New Business Enterprise. Flow Zone EDITION.

EASTERLIN, R. A., (2004). The Economics of Happiness. Daedalus, Vol. 133, No. 2, Spring 2004.

FOLBRE, N., (1994). Who Takes Care of the Kids? Gender and the structures of constraint. Routledge, London.

HOZER-KOĆMIEL, M., (2008). Gender Mainstreaming in economics. Woman work time and value distribution. US, IADiPG, Szczecin.

HOZER-KOĆMIEL, M., LIS, CH., (2015). Klasyfikacja krajów nadbałtyckich ze względu na czas prac wykonywanych w gospodarstwie domowym [Classification of the Baltic states in terms of duration of works performed in a household], Taksonomia 25, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, No. 385.

JAJUGA, K., (1993). Statystyczna analiza wielowymiarowa [Statistical multidimensional analysis], PWN, Warsaw.

LIS, CH., (2013). Wartość dodana brutto i jej znaczenie w procesie akumulacji kapitału w świetle teorii wzrostu i konwergencji. Podejście taksonomiczne [Gross Value Added and Its Significance in the Capital Formation with Reference to Growth and Convergence Theories, A Taxonomic Approach]. Publishing House Volumina.pl, Szczecin.

PIETILA, H., (1997).The triangle of human economy: household - cultivation - industrial production. An attempt at making the human economy visible in toto, Ecological Economics, No 20.

POCIECHA, J., PODOLEC, B., SOKOŁOWSKI, A., ZAJĄC, K., (1988). Metody taksonomiczne w badaniach społeczno-ekonomicznych [Taxonomic methods in socio-economic studies], PWN, Warsaw.

STIGLIZ, J. E., SEN, A., FITOUSSI, J. P., (2010). Report by the Commission on the Measurement of Economic Performance and Social Progress. <http://www.stiglitz-sen-fitoussi.fr> [Accessed 12.10.2015].

STRAHL, D., WALESIAK, M., (1997). Normalizacja zmiennych w skali przedziałowej i ilorazowej w referencyjnym systemie granicznym [Normalization of variables in interval and ratio scales in the border reference system]. Przegląd Statystyczny, No. 1, Warsaw.

WALESIAK, M., (2000). Propozycja uogólnionej miary odległości w statystycznej analizie wielowymiarowej [The Proposal of the Generalised Distance Measure in Multivariate Statistical Analysis], In J. Paradysz (ed.), Statystyka regionalna w służbie samorządu lokalnego i biznesu [Regional statistics in the service of local self-government and business]. Wydawnictwo AE, Poznań.

WALESIAK, M., (2011). Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R [GDM Generalized distance measure in multivariate statistical analysis using R software]. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu. Wrocław.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 331–348

DECOMPOSITION OF DIFFERENCES IN INCOME DISTRIBUTIONS USING QUANTILE REGRESSION

Joanna Małgorzata Landmesser[1]

ABSTRACT

The paper deals with microeconometric techniques useful for the study of differences between groups of objects, methods that go beyond simple comparison of average values. Techniques for the decomposition of differences in distributions by constructing counterfactual distributions were considered. Using the Machado-Mata quantile regression approach the empirical decomposition of the inequalities in income distributions of one-person households in urban and rural areas was performed. We employed data from the Household Budget Survey for Poland in 2012. It was found that the tendency towards increased income inequalities between urban and rural residents when moving to the right of the income distribution can be observed. The rural residents are at a disadvantage. The decomposition of the inequalities revealed a growing share of the part explained by different characteristics of people and a declining share of the unexplained part, associated with the evaluation of those characteristics.

Key words: decomposition of differences, quantile regression, counterfactual distribution.

1. Introduction

Recent years have witnessed the rapid development of microeconometric techniques useful in the context of studying the differences between groups of objects. Various inequality decomposition methods are becoming more popular. Since the seminal works of Oaxaca (1973) and Blinder (1973) many procedures that go beyond simple decomposition of differences between the average values have been proposed. These are the variance decomposition techniques and the decomposition allowing the analysis of the differences with respect to the entire distribution of the outcome variable.

The main advantage of modern decomposition methods is to help to discover the factors affecting changes in the distribution of wages, for example. Studying

[1] Warsaw University of Life Sciences. E-mail: joanna_landmesser@sggw.pl.

changes in the distribution of wages has become an active area of research (see, e.g. Juhn, Murphy and Pierce, 1993; DiNardo, Fortin and Lemieux, 1996; Gosling, Machin and Meghir, 2000; Donald, Green and Paarsch, 2000; Machado and Mata, 2005; Autor, Katz and Kearney, 2005). For instance, DiNardo, Fortin and Lemieux (1996) analysed the implications of the observed changes on the labour market for specific points of the wage distribution and found that the minimum wage affects only the bottom end of this distribution. Other explanations based on de-unionization tend to affect the middle of the distribution (Card, 1992). The differences in income distributions between various groups of people, e.g. women and men, were also analysed (see Albrecht, Björklund and Vroman (2003), who look whether there is a glass ceiling in female earnings).

In particular, new techniques make it possible to carry out the decomposition of the differences in distributions, by constructing a counterfactual distribution that mixes conditional distribution for the outcome variable Y with various distributions for explanatory variables X. The most popular methods of constructing counterfactual distributions are those proposed in Juhn, Murphy and Pierce (1993), DiNardo, Fortin and Lemieux (1996), Machado and Mata (2005). Machado and Mata (2005) suggested using quantile regression in order to estimate counterfactual unconditional wage distributions.

The aim of our study was to decompose the observed inequalities in income distributions of one-person households in urban and rural areas applying the Machado-Mata technique. We employ data from the Household Budget Survey (HBS) for Poland in 2012. Applying the method to one-person households in urban and rural areas, not the typical class of the Polish households, allows interpersonal comparisons of the individual's income. An income of, say 4000 zlotys per head a month, implies a different purchasing power for a household of one and four persons. Living costs are often higher for single person households. Needs for housing space, electricity, etc. will not be four times as high for a household with four members as for a single person. Therefore, in order to properly compare the household incomes our attention has been focused only on one-person households.

The research concerning the income gap, conducted so far in Poland, was limited to decomposing mainly the average level of wage differences for men and women using the Oaxaca-Blinder method (e.g. Kot, Podolec and Uhlman, 1999; Słoczyński, 2012; Goraus, 2013; Śliwicki, Ryczkowski, 2014). Only a few studies go beyond the mean-decomposition. Grajek (2003) applied the John, Murphy and Pierce decomposition to analyse data on Polish employees from the period 1987–1996. He found that the explained component of gender pay gap is relatively small and rises slowly over the analysed period. Newell and Socha (2005), on the basis of quantile analyses using Labour Force Survey (LFS) data for 1992–2002, showed that many of the factors influencing wages, including gender, have a stronger impact in higher quantiles of wage distribution. Rokicka and Ruzik (2010) found that the inequality of earnings between women and men tends to be larger at the top of the earnings distribution (in the case of formal employees).

Nobody in Poland has made the decomposition of income inequalities for residents in urban and rural areas. In this paper we apply the Machado-Mata technique in order to move beyond estimation based on mean values. We argue that employing these techniques can provide deeper insights into the nature of income differentials.

The structure of the paper is as follows. Section 2 describes various techniques used for the decomposition of inequalities. Section 3 presents data and the results of the decomposition of differences in income densities between urban and rural inhabitants. Section 4 discusses the results and offers some concluding remarks.

2. Analysis method

This section outlines the methodology to be employed. First, we present the Oaxaca-Blinder decomposition of differences in mean wages. Then, we explain the idea of the decomposition of differences along the entire distribution. Finally, we present the conditional quantile decomposition techniques developed by Machado and Mata (2005).

2.1. Oaxaca-Blinder decomposition of differences in mean wages

There are two groups given, A and B, an outcome variable y, and a set of predictors X. The variable y may present log wages and predictors X may concern such individual socio-demographic characteristics of people as age, education level or work experience. The idea of Oaxaca-Blinder decomposition can be applied whenever we need to explain the differences between the expected values of dependent variable y in two comparison groups (Oaxaca, 1973; Blinder, 1973). The authors of the methods assume that the expected value of y conditionally on X is a linear function of X:

$$y_g = X_g \beta_g + v_g, \qquad g = A, B, \tag{1}$$

where X_g are the characteristics of people in group g and β_g are the returns to these characteristics. The idea of Oaxaca-Blinder decomposition of the difference $\Delta^{\mu} = E(y_A) - E(y_B)$ is as follows:

$$\hat{\Delta}^{\mu} = \overline{X}_A \hat{\beta}_A - \overline{X}_B \hat{\beta}_B = \underbrace{(\overline{X}_A - \overline{X}_B)\hat{\beta}_A}_{\hat{\Delta}^{\mu}_{\text{explained}}} + \underbrace{\overline{X}_B(\hat{\beta}_A - \hat{\beta}_B)}_{\hat{\Delta}^{\mu}_{\text{unexplained}}} \tag{2}$$

The above equation is based on characteristics of one group and the estimated coefficients of the equation of another group. The first term on the right-hand side of the equation gives the effect of characteristics and expresses the difference of the potentials of both groups (the so-called explained, endowments or composition effect). The second term represents the effect of coefficients,

typically interpreted as discrimination in numerous studies (the so-called unexplained, wage structure effect). This is the result of differences in the estimated parameters, and consequently in the "prices" of individual characteristics of representatives of a group. Blinder argued that "the latter sum […] exists only because the market evaluates differently the identical bundle of traits if possessed by different demographic groups" (Blinder, 1973, pp. 438-439).

One important drawback of this technique is that it focuses only on average effects, and this may lead to a misleading assessment if the effects of covariates vary across the wage distribution (Salardi, 2012).

2.2. Beyond the mean - decomposition of differences in distributions

The preceding scalar decomposition analysis may be extended to the case of differences along the entire distribution. Let $f^A(y)$ and $f^B(y)$ be the density functions for the outcome variable y in group A and B, respectively. The distribution $f^i(y)$, $i = A, B$, is the marginal distribution of the joint distribution $\varphi^i(y,X)$:

$$f^i(y) = \int_{C(X)} \ldots \int \varphi^i(y,X)dX, \tag{3}$$

where X is a vector of individual characteristics observed and $C(X)$ is the domain on which X is defined (cf. Bourguignon and Ferreira, 2005, p.28). Denoting $g^i(y|X)$, the conditional distribution of y, an equivalent expression for (3) is:

$$f^i(y) = \int_{C(X)} \ldots \int g^i(y|X)h^i(X)dX, \tag{4}$$

with $h^i(X)$ as the joint distribution of all elements of X in group i.

The observed difference between the two distributions may be decomposed into

$$f^A(y) - f^B(y) = [f^A(y) - f^C(y)] + [f^C(y) - f^B(y)], \tag{5}$$

where $f^C(y)$ represents the counterfactual distribution, which can be constructed for example as

$$f^C(y) = \int_{C(X)} \ldots \int g^A(y|X)h^B(X)dX. \tag{6}$$

The first term on the right-hand side of equation (5) gives the effect of different endowment's distributions in group A and group B. The second term describes the inequalities between two distributions of y conditional on characteristics X. The main difference with respect to the Oaxaca-Blinder decomposition is that this decomposition refers to full distributions, rather than just to their means. The formula (5) may be applied to any statistic defined on the

distribution of outcome variable *y*: mean, quantiles, summary measures of inequality such as the variance or the Gini coefficient.

Several approaches have been suggested in the literature for estimating the counterfactual distribution $f^C(y)$ (cf. Fortin, Lemieux and Firpo, 2010). An approach proposed by Juhn, Murphy and Pierce (1993) is based on the residual imputation procedure. DiNardo, Fortin and Lemieux (1996) suggested to use a reweighting factor. Donald, Green and Paarsch (2000) used a hazard model approach, Fortin and Lemieux (1998) applied an ordered probit. Machado and Mata (2005) proposed using quantile regression to transform a wage observation *y* into a counterfactual observation y^C.

2.3. Decomposition of differences in distributions using quantile regression

The standard linear regression assumes the relationship between the regressors and the outcome variable based on the conditional mean function. This, however, gives only a partial insight into the connection. The quantile regression (Koenker and Bassett, 1978) allows the description of the relationship at different points in the conditional distribution of *y*.

Let us consider the relationship between the regressors and outcome using the conditional quantile function:

$$Q_\theta(y|X) = \Phi^{-1}_{y|X}(\theta, X) = X\beta(\theta), \tag{7}$$

where $Q_\theta(y|X)$ - the θ^{th} quantile of a variable *y* conditional on covariates *X*, $\theta \in (0,1)$; $\Phi_{y|X}$ - the joint distribution function for the variable *y*.

A different quantile θ may be specified (most frequently from three to nine). For each quantile other parameters $\beta(\theta)$ are estimated. These coefficients can be interpreted as the returns to different characteristics *X* at given quantiles of the distribution of *y*. Bootstrap standard errors are often used (Gould, 1992; 1997). Quantile regression is more robust than least squares regression to non-normal errors and outliers. This method also provides a richer characterization of data, considering the impact of covariates on the entire distribution of *y*, not only its conditional mean.

Machado and Mata (2005) used quantile regression in order to estimate counterfactual unconditional wage distributions. Since the unconditional quantile is not the same as the integral of the conditional quantiles, authors provide a simulation-based estimator where the counterfactual distribution is constructed from the generation of a random sample. This estimator is widely used in various applications (cf. Albrecht, Björklund and Vroman, 2003; Melly, 2005). The idea underlying this technique is the probability integral transformation theorem. If *U* is a uniform random variable on [0,1], then $F^{-1}(U)$ has distribution *F*. Thus, if θ_1, θ_2, …, θ_m are drawn from a uniform (0,1) distribution, the corresponding *m*

estimates of the conditional quantiles of wages at X, $\left\{X\hat{\beta}(\theta_i)\right\}, i=1,...,m$, constitute a random sample from the (estimated) conditional distribution of wages given X (Machado and Mata, 2005, p.448).

The Machado-Mata approach to generate a random sample from the wage density that would prevail in group A if model (7) was true and covariates were distributed as $h^A(X)$ is as follows:

1. Generate a random sample of size m from a $U[0,1]$: $u_1, \ldots, u_m$.
2. Using the dataset for group A, estimate m different quantile regression $Q_{u_i}(y|X_A)$, obtaining coefficients $\hat{\beta}_A(u_i), i=1,...,m$.
3. Generate a random sample of size m with replacement from the rows of X_A, denoted by $\left\{X^*_{Ai}\right\}, i=1,...,m$.
4. $\left\{y^*_{Ai} \equiv X^*_{Ai}\hat{\beta}_A(u_i)\right\}, i=1,...,m$ is a random sample of size m from the unconditional distribution $f^A(y)$.

Counterfactual distributions could be estimated by drawing X from another distribution and using different coefficient vectors. For example, to generate a random sample from the wage density that would prevail in group A and covariates that were distributed as $h^B(X)$ (e.g. the men log wage density that would arise if men were given women's labour market characteristics but continued to be paid like men), we generate a random sample from the rows of X_B, denoted by $\left\{X^*_{Bi}\right\}, i=1,...,m$, and $\left\{y^{C*}_{Ai} \equiv X^*_{Bi}\hat{\beta}_A(u_i)\right\}$ is a random sample from the counterfactual distribution $f^C(y)$.

Consequently, the idea of Machado-Mata decomposition of the difference between the wage densities in two groups, $\Delta^\theta = Q_\theta(y_A) - Q_\theta(y_B), \forall\theta$, is as follows:

$$\begin{aligned}\hat{\Delta}^\theta &= Q_\theta(y^*_A|X^*_A) - Q_\theta(y^*_B|X^*_B) = \\ &= Q_\theta(y^*_A|X^*_A) - Q_\theta(y^{C*}_A|X^*_B) + Q_\theta(y^{C*}_A|X^*_B) - Q_\theta(y^*_B|X^*_B) = \\ &= \underbrace{(X^*_A - X^*_B)\hat{\beta}_A(\theta)}_{\hat{\Delta}^\theta_{\text{explained}}} + \underbrace{(\hat{\beta}_A(\theta) - \hat{\beta}_B(\theta))X^*_B}_{\hat{\Delta}^\theta_{\text{unexplained}}}\end{aligned} \tag{8}$$

We can also use the Machado-Mata approach to estimate standard errors for the estimated densities by repeating the procedure many times and generating a set of estimated densities. The standard error of the estimator diminishes as we increase the number of replications, but estimating a large number of replications is time-consuming especially when the number of observations is high (Melly, 2006).

3. Results of empirical analysis

After outlining the methodology, we now provide the results of our empirical analysis. First, we present the data. Then, we discuss the results of Oaxaca-Blinder decomposition for mean incomes in urban and rural area. In the next step, we present the quantile regressions estimates for models with various predictors. Finally, we implement the Machado and Mata quantile decomposition technique for differences in income densities between inhabitants in urban and rural area.

3.1 Empirical data

We employ data from the Household Budget Survey (HBS) for Poland in 2012. This representative survey is conducted by the Central Statistical Office, Social Surveys and Living Conditions Statistics Department. It is one of the most comprehensive sources of socio-economic information on Polish households and it plays an important role in the analysis of the living standards of the population. It is the source of information on the revenues and outgoings of such socio-economic groups like employees' households, farmers' households, households of the self-employed, households of retirees and pensioners and households living on unearned sources (GUS, 2013).

Our data consist of a sample of 7056 one-person households (5146 town residents and 1910 village residents) containing information on household's monthly available income as well as on reference persons' attributes, such as gender, age, education level, place of residence. Household's available income is defined as the sum of household's current incomes from various sources reduced by taxes, and it comprises: income from hired work, income from a private farm in agriculture, income from self-employment, income from property and social insurance benefits.

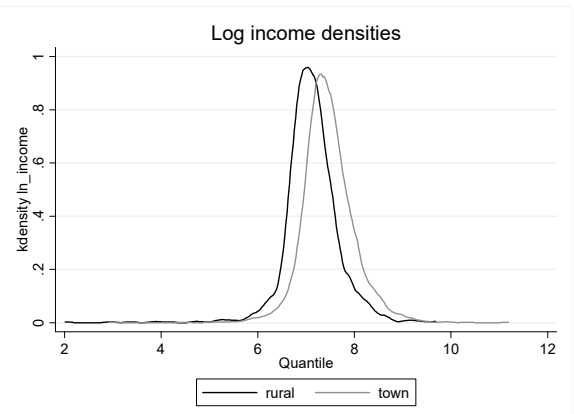


Figure 1. (Log) income densities for the residents of urban and rural areas.
Source: own elaboration based on GUS (2012).

In our empirical decomposition analysis the logarithm of the average monthly available income (*ln_income*) constitutes the outcome variable. Figure 1 illustrates the kernel estimates of the log income densities for the residents of urban and rural areas. The data indicate that the income inequality can be observed. When we look at both distributions, we find that urban residents have the level of the log income higher than the rural residents. The average monthly available income of a person in urban area was PLN 2,028.11, whereas for a person in rural area it was only PLN 1,444.47 (see Table 1).

Table 1. Descriptive statistics for the sample

	Whole sample	Urban area	Rural area
Number of observation	7056	5146	1910
Household's available	1870.13 (1602.47)	2028.11	1444.47
ln_income	7.39 (0.55)	7.46 (0.53)	7.12 (0.55)
sex (% men)	29	28	33
age	60.58 (17.54)	58.88 (18.44)	65.13 (13.86)

Sample averages; standard errors in parentheses.

Source: own elaboration based on GUS (2012).

We establish three explanatory variables in our models: *sex* – the dichotomous variable encoding gender of a person (number 1 coded the male sex), *age* – age of a person in years, *education* – an ordinal variable describing the educational level.

It is useful to look at summary statistics for some covariates (in Table 1). Among the urban residents there were less men (28%) than it was in the case of the rural residents (33% of men). The average age of a person in urban area was only 58.88 years, whereas for a person in rural area it was 65.13 years. The average educational level in the countryside was lower than in cities.

3.2. Results of Oaxaca-Blinder decomposition for differences in mean log incomes

Many authors examined the determinants of income and the income gap in the urban and rural areas. For example, Sicular et al. (2007) and Su and Heshmati (2013) analysed the urban-rural income gap in China using the Oaxaca-Blinder decomposition method. Ali et al. (2013) used this method to analyse the income gap between urban and rural Pakistan. Haisken-DeNew and Michaelsen (2011) investigated the differences in wages between rural and urban workers in the informal and formal sectors of Mexico's labour market. The set of regressors in their papers included conventional human capital characteristics (e.g. education, occupation or experience), personal characteristics (e.g. age, gender, marital status) or regional labour market conditions (Adamchik and Bedi, 2003). The demographic characteristics such as the household size, the proportion of dependents versus working-age household members may also be important for the household incomes (Knight, Song, 1999; Miles, 1997).

The first step of our analysis also included the decomposition of the income inequalities observed between residents of urban and rural areas using the Oaxaca-Blinder technique. The results on an aggregated and detailed basis are presented in Table 2.

Table 2. Oaxaca-Blinder decomposition of mean differences in log incomes for residents in urban and rural areas

Average *ln_income* in urban area			7.460
Average *ln_income* in rural area			7.124
Raw gap (differential observed)			0.336
Aggregate decomposition			
Explained effect			0.185
Unexplained effect			0.151
% explained			55
% unexplained			45
Detailed decomposition			
due to characteristics		due to returns	
sex	-0.002	*sex*	0.035
age	0.008	*age*	-0.268
education	0.179	*education*	0.039
cons	0.000	*cons*	0.344
Total	0.185	Total	0.151

Source: own elaboration using the Stata command 'decompose'.

There is a positive difference between the mean values of the log income for urban and rural residents, meaning that the inhabitants in rural area have lower average incomes than inhabitants in urban area. The inequalities examined should be assigned to a similar extent to the characteristics (55%) as well as to the coefficients (45%) of the estimated regression models. Decomposition, which was carried out, made it possible to isolate the factors explaining the inequality observed to a different extent. The strong effect of different education levels of people living in rural and urban areas can be noticed (see the value of 0.179). A different "evaluation" of personal characteristics allow the conclusion that the residents in rural area are discriminated against residents in urban area, but not because of the age of people (due to the negative value -0.268). A large part of the unexplained component lies in the intercept differences (that is, the inter-group differences in other factors were not captured in the model).

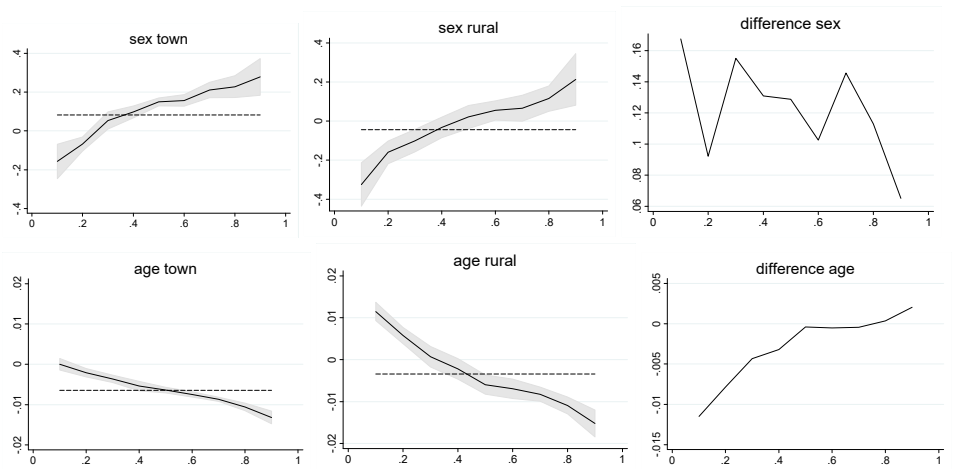
Our findings are mainly consistent with that reported elsewhere. Sicular et al. (2007) found for China that education was the only characteristic whose contribution to the income gap was significant. The contribution of education was largely due to differences in the endowments and not in the returns. According to

the results of Su and Heshmati (2013), the urban-rural income gap can be explained by attributes of individuals, especially by the level of education and the type of occupation. The educational returns were higher among urban residents. The gender income gap was evident, showing males had higher income than females. For the formal sector in Mexico, Haisken-DeNew and Michaelsen (2011) revealed that only differences in education contribute to the explanation of the wage gap and no differences in coefficients can be identified.

3.3. Estimation of quantile regressions

Some recent studies have decomposed the urban-rural income gap by focusing on the entire distribution of income and not just on the means. Nguyen et al. (2007) and Huong and Booth (2010) adopted the quantile regression method to analyse urban-rural consumption expenditure inequality in Vietnam. Shilpi (2008) and Chamarbagwala (2010) used this method to analyse income gap between urban and rural Bangladesh and India. Matita and Chirwa (2009) analysed the extent of urban-rural welfare inequalities in Malawi using the Machado and Mata decomposition technique.

Therefore, we analyse the quantile regression results for the outcome variable *ln_income* in the second step of our research. The plots in Figure 2 show the coefficient estimates $\hat{\beta}_i(\theta)$ with the associated 95% confidence intervals, obtained by the bootstrap method with 100 replications. For the variables *sex*, *age* and *education* the plots provide information on the coefficients in models estimated using the data from urban area (left column), rural area (central column) and the difference between the parameters $\hat{\beta}_i^{town}(\theta) - \hat{\beta}_i^{rural}(\theta)$ (right column). Additionally, in the first two columns the coefficients estimated by mean regression (OLS) are reported (dotted horizontal lines). The graphs illustrate what is the impact of each covariate on income inequality.



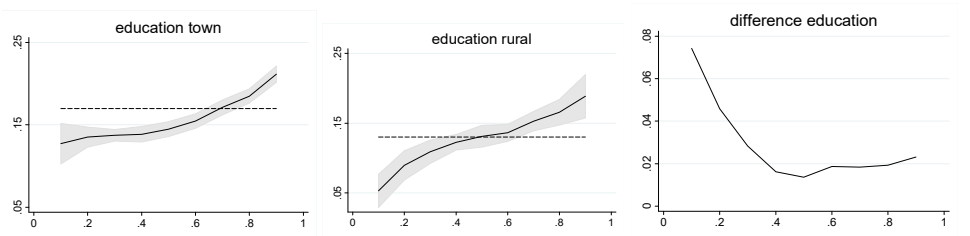


Figure 2. Quantile regression coefficients with 95% confidence intervals for the deciles; the dotted horizontal line represents the least squares (conditional mean) estimate.

Source: own elaboration using the Stata command 'sqreg'.

We can see that among the poorest, the income of men is lower than that of women (see negative coefficients for the urban and rural area), but among the richest being the man gives higher income (see positive coefficients values). The differences between parameter estimates are positive, but from the 70th quantile to the right, they are smaller. This means that the market evaluation of the gender of people is responsible for the existing but decreasing (moving to the right of the distribution) income inequalities between urban and rural inhabitants. It turns out that gender is an important unexplained contributor to the observed income gaps. It is worth mentioning that also for China the gender has greater effects on people with lower level of income in rural area (Su and Heshmati, 2013).

Age is more rewarded among the poorest, but among the richest the growing age leads to the decline in income (both in urban and rural area). We find such an influence of the role of age on the income gap at the bottom of the income distribution. The unexplained part of the gap can decrease due to the negative differences between parameter estimates (compare this with the conclusion of subsection 3.2).

Finally, we find that incomes increase with education across the whole distribution. The education level is the significant contributor to the income differences in urban and rural areas not only in endowments, which favour urban inhabitants (see Table 1), but also in returns of that individual characteristic (note that the differences between parameter estimates are positive although increasingly smaller). These results are contrary to the results obtained by Su and Heshmati (2013) for China, where the education exerts heterogeneous effects on different percentiles of the income distribution. In urban areas, education is more valued for high income earners, whereas for rural areas, specialized or tertiary education is more beneficial for poorer households.

3.4. Empirical decomposition of differences in income distributions for one-person households in urban and rural area

In the third step, we decompose the inequalities in the income distributions into differences in the covariates (individual attributes) and differences attributable to the coefficients (remuneration of individual characteristics). We follow the Machado-Mata procedure as carefully as possible, except that rather than drawing m numbers at random from $U[0,1]$ and then estimating m quantile regression coefficient vectors, we simply estimate the quantile regressions every one percentile (in Table 3 we provide decomposition results only for nine deciles). Then, we make 100 draws at random from the X matrix for each estimated coefficient vector $\hat{\beta}_t(\theta)$.

The results are summarized in Figure 3. The graph on the left plots the differences between pairs of distributions of interest (these are the raw gap, the Machado-Mata differences with the associated 95% confidence interval and the Oaxaca-Blinder mean difference for comparison purposes). The graph on the right presents the decomposition results.

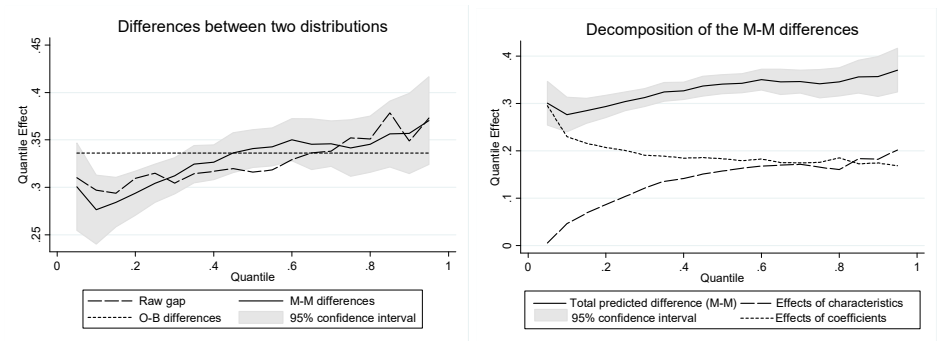


Figure 3. Decomposition of differences in income distributions for residents in urban and rural areas.

Source: own elaboration using the Stata command 'mmsel'.

These results are presented in greater detail in Table 3. The first column of this table refers to the decile number, the second column presents the raw gap between log income distributions for inhabitants in urban and rural areas. The third column contains estimates of the Machado-Mata differences with standard errors in parentheses. The next two columns decompose total inequalities into differences due to the covariates and differences due to changes in the coefficients (standard errors in parentheses). The last two columns give the respective proportions of the total inequalities explained by both kinds of differences.

The findings confirm that in urban area monthly available income of one-person household is on average greater than in rural area, and its inequality at the highest quantiles of the income distribution is also larger among the former than among the latter.

Table 3. Decomposition of differences in income distributions

Decile	Raw gap	M-M differences	Explained effect	Unexplained effect	% Ex-plained	% Unex-plained
0.10	0.2971	0.2764 (0.0185)	0.0463 (0.0175)	0.2301 (0.0193)	17%	83%
0.20	0.3096	0.2937 (0.0119)	0.0867 (0.0116)	0.2071 (0.0123)	30%	70%
0.30	0.3046	0.3124 (0.0097)	0.1216 (0.0098)	0.1908 (0.0109)	39%	61%
0.40	0.3168	0.3265 (0.0093)	0.1417 (0.0091)	0.1848 (0.0103)	43%	57%
0.50	0.3161	0.3409 (0.0102)	0.1575 (0.0111)	0.1834 (0.0112)	46%	54%
0.60	0.3292	0.3503 (0.0113)	0.1676 (0.0117)	0.1827 (0.0124)	48%	52%
0.70	0.3382	0.3461 (0.0123)	0.1716 (0.0114)	0.1745 (0.0127)	50%	50%
0.80	0.3511	0.3455 (0.0152)	0.1602 (0.0161)	0.1853 (0.0164)	46%	54%
0.90	0.3488	0.3569 (0.0215)	0.1825 (0.0197)	0.1744 (0.0198)	51%	49%

Standard errors in parentheses.

Source: own elaboration using the Stata command 'mmsel'.

Figure 3 also shows that the income gaps are wider at the top of distribution. Both covariates and coefficients, contribute to the explanation of the total inequalities sum and their effects are significantly different from zero in all of the estimated deciles (the confidence intervals are not provided because of lack of space, but they do not include zero). We can clearly see that the effect of coefficients is more important than that of covariates at the bottom of the income distribution. However, the unexplained differential shrinks as we move toward the top of the income distribution. By contrast, the percentage of the explained (due to the characteristics) income differential is considerably greater as we move to the right-side of the distribution.

Our findings can be compared with the results of Huong and Booth (2010) who found evidence of significant urban-rural expenditure inequality in Vietnam. In this case, the urban-rural gap monotonically increased across the expenditure distribution. Also, Nguyen et al. (2007) analysed the urban-rural consumption expenditure inequality in this country and showed that the returns due to the covariates were larger at the top of the distribution of household consumption expenditure per capita. Regarding the studies of urban-rural income inequalities, Shilpi (2008) and Chamarbagwala (2010) found that both the covariates and the returns were relevant in explaining the observed income gap, although their behaviours were different across the distribution of welfare.

4. Conclusions

The objective of the study was to perform the decomposition of income inequalities between one-person households in urban and rural areas. In order to extend the Oaxaca-Blinder decomposition procedure to different quantile points along the income distribution, we applied the Machado-Mata decomposition technique and constructed the counterfactual income distribution.

It is worth mentioning that the decomposition method applied was computationally intensive. The calculation could be simplified by estimating a specific number of quantile regressions (i.e. 99) instead of generating a random sample of size m from $U[0,1]$. Another limitation was the assumption of the linearity of the quantile regression model. Besides this, the Machado-Mata approach does not provide a way of performing the detailed decomposition for the endowment effect (Fortin, Lemieux and Firpo, 2010, p.61).

Turning to the main findings from the aggregate decomposition of the income disparities in urban and rural areas, we found that the income differentials tend to increase. There are higher differentials at the top of the income distribution, which are driven by the endowment effects as well as by the structure effects. However, the widening income inequalities at higher income quantiles are mainly possible due to growing differences in characteristics of people (especially in educational level) in favour of urban residents. The discrimination affecting rural residents by the higher incomes becomes less important (due to descending structure effects associated with the coefficients of the model). Moving to the bottom of the distribution the total income gap declines, owing primarily to a decline in the explained components.

REFERENCES

ADAMCHIK, V. A., BEDI, A. S., (2003). Gender pay differentials during the transition in Poland. Economics of Transition, 11(4), pp. 697–726.

ALBRECHT, J., BJÖRKLUND, A., VROMAN, S., (2003). Is There a Glass Ceiling in Sweden? Journal of Labor Economics, 21, pp. 145–177.

ALI, L. R., RAMAY, M. I., NAS, Z., (2013). Analysis of the determinants of income and income gap between urban and rural Pakistan. Interdisciplinary Journal of Contemporary Research in Business, 5(1), pp. 858–885.

AUTOR, D. H., KATZ, L. B., KEARNEY, M. S., (2005). Rising Wage Inequality: The Role of Composition and Prices. Cambridge: NBER Working Paper, No. 11628.

BLINDER, A., (1973). Wage Discrimination: Reduced Form and Structural Estimates. Journal of Human Resources, 8, pp. 436–455.

BOURGUIGNON, F., FERREIRA, F. H. G., (2005). Decomposing Changes in the Distribution of Household Incomes: Methodological Aspects. In: F. Bourguignon, F. H. G. Ferreira and N. Lustig, eds. 2005. The Microeconomics of Income Distribution Dynamics in East Asia and Latin America. Washington: World Bank and Oxford University Press. pp. 17–46.

CARD, D., (1992). The Effects of Unions on the Distribution of Wages: Redistribution or Relabelling? Cambridge: NBER Working Paper, No. 4195.

CHAMARBAGWALA, R., (2010). Economic Liberalization and Urban-Rural Inequality in India: a Quantile Regression Analysis. Empirical Economics, 39, pp. 371–394.

DINARDO, J., FORTIN, N. M., LEMIEUX, T., (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. Econometrica, 64, pp. 1001–1044.

DONALD, S. G., GREEN, D. A., PAARSCH, H. J., (2000). Differences in Wage Distributions between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates. Review of Economic Studies, 67, pp. 609–633.

FORTIN, N. M., LEMIEUX, T., (1998). Rank Regressions, Wage Distributions, and the Gender Gap. Journal of Human Resources, 33, pp. 610–643.

FORTIN, N., LEMIEUX, T., FIRPO S., (2010). Decomposition methods in economics. Cambridge: NBER Working Paper, No. 16045.

GORAUS, K., (2013). Luka płacowa miedzy kobietami a mężczyznami w Polsce. Available at:
<http://grape.uw.edu.pl/gender/wp-content/uploads/sites/35/2013/09/
Łódź_prezentacja.pdf> [Accessed 20 September 2015].

GOSLING, A., MACHIN, S., MEGHIR, C., (2000). The Changing Distribution of Male Wages in the UK. Review of Economic Studies, 67, pp. 635–686.

GOULD, W. W., (1992). Quantile Regression with Bootstrapped Standard Errors. Stata Technical Bulletin, 9, pp. 19–21.

GOULD, W. W., (1997). Interquantile and Simultaneous-Quantile Regression. Stata Technical Bulletin, 38, pp. 14–22.

GRAJEK, M., (2003). Gender Pay Gap in Poland. Economic Change and Restructuring, 36(1), pp. 23–44.

GUS, (2013). Budżety Gospodarstw Domowych w 2012 r. [Household Budget Survey 2012] Warsaw: Central Statistical Office, Social Surveys and Living Conditions Statistics Department.

HAISKEN-DENEW, J. P., MICHAELSEN, M.M., (2011). Migration Magnet: The Role of Work Experience in Rural-Urban Wage Differentials in Mexico. Bochum: Ruhr Economic Papers No. 263.

HUONG, T. L., BOOTH, A. L., (2010). Inequality in Vietnamese Urban-Rural Living Standards, 1993-2006. Bonn: IZA Discussion Paper, No. 4987.

JUHN, CH., MURPHY, K. M., PIERCE, B., (1993). Wage Inequality and the Rise in Returns to Skill. Journal of Political Economy, 101, pp. 410–442.

KNIGHT, J., SONG, L., (1999). The Urban-Rural Divide: Economic Disparities and Interactions in China. New York: Oxford University Press.

KOENKER, R., BASSETT, G., (1978). Regression Quantiles. Econometrica, 46, pp. 33–50.

KOT, S. M., PODOLEC, B., ULMAN, P., (1999). Problem dyskryminacji płacowej ze względu na płeć [Problem of earning discrimination by gender]. In: Kot, S.M., ed. 1999. Analiza ekonometryczna kształtowania się płac w Polsce w okresie transformacji. Warszawa, Kraków: Wydawnictwo Naukowe PWN.

MACHADO, J. F., MATA, J., (2005). Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression. Journal of Applied Econometrics, 20, pp. 445–465.

MATITA, M. M., CHIRWA, E. W., (2009). Rural-Urban Welfare Inequalities in Malawi: Evidence from a Decomposition Analysis. University of Malawi Chancellor College, Department of Economics, Working Paper No. 2009/05.

MELLY, B., (2005). Public-Private Sector Wage Differentials in Germany: Evidence from Quantile Regression. Empirical Economics, 30, pp. 505–520.

MELLY, B., (2006). Estimation of counterfactual distributions using quantile regression. Review of Labor Economics, 68 (4), pp. 543–572.

MILES, D., (1997). A household level study of the determinants of incomes and consumption. The Economic Journal, 107(440), pp. 1–25.

NEWELL, A., SOCHA, M., (2005). The Distribution of Wages in Poland. Bonn: IZA Discussion Paper, No. 1485.

NGUYEN, B. T., ALBRECHT, J. W., VROMAN, S. B., WESTBROOK, M. D., (2007). A Quantile Regression Decomposition of Urban-Rural Inequality in Vietnam. Journal of Development Economics, 83, pp. 466–490.

OAXACA, R., (1973). Male-Female Wage Differentials in Urban Labor Markets. International Economic Review, 14, pp. 693–709.

ROKICKA, M., RUZIK, A., (2010). The Gender Pay Gap in Informal Employment in Poland. Warszawa: CASE Network Studies and Analyses, No. 406.

SALARDI, P., (2012). Wage Disparities and Occupational Intensity by Gender and Race in Brazil: An Empirical Analysis using Quantile Decomposition Techniques. Brighton: University of Sussex, Job Market Paper. Available at: <http://www.iza.org/conference_files/worldb2012/salardi_p7646.pdf> [Accessed 20 September 2015].

SICULAR, T., YUE, X., GUSTAFSSON, B., LI, S., (2007). The Urban-Rural Income Gap and Inequality in China. Review of Income and Wealth, 53 (1), pp. 93–126.

SHIPI, F., (2008). Migration, Sorting and Regional Inequality: Evidence from Bangladesh. World Bank Policy Research Working Paper No. 4616.

SŁOCZYŃSKI, T., (2012). Wokół międzynarodowego zróżnicowania międzypłciowej luki płacowej [The issue of international differentiation of inter-gender earning gap]. International Journal of Management and Economics, 34, pp. 169–185.

SU, B., HESHMATI, A., (2013). Analysis of the Determinants of Income and Income Gap between Urban and Rural China. Bonn: IZA Discussion Paper, No. 7162.

ŚLIWICKI, D., RYCZKOWSKI, M., (2014). Gender Pay Gap in the micro level – case of Poland. Quantitative Methods in Economics, Vol. XV, No. 1, pp. 159–173.

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 349—351

REPORT

The XXXIV International Conference on Multivariate Statistical Analysis, 16–18 November 2015, Łódź, Poland

The 34th edition of the International Conference on **Multivariate Statistical Analysis** was held in **Łódź**, **Poland** on November 16-18, 2015. The MSA 2015 conference was organized by the **Department of Statistical Methods** of the University of Lodz, the **Institute of Statistics and Demography** of the University of Lodz, the **Polish Statistical Association** and the **Committee on Statistics and Econometrics of Polish Academy of Sciences**. The Organizing Committee was headed by **Professor Czesław Domański**. The scientific secretaries included Anna Jurek, M.Sc. and Elżbieta Zalewska, M.Sc. from the Department of Statistical Methods of the University of Lodz.

The Mayor of the City of Łódź, Hanna Zdanowska took the honorary patronage of the Multivariate Statistical Analysis MSA 2015 conference. Its organization was financially supported by the **National Bank of Poland**, the **Polish Academy of Sciences,** the **Łódź City Council** and **StatSoft Polska Sp. z o.o.**

The 2015 edition, as all previous Multivariate Statistical Analysis conferences, aimed at creating the opportunity for scientists and practitioners of statistics to present and discuss the latest theoretical achievements in the field of the multivariate statistical analysis, its practical aspects and applications. A number of the presented and discussed statistical issues were based on questions identified during previous MSA conferences. The scientific programme covered various statistical problems, including multivariate estimation methods, statistical tests, non-parametric inference, discrimination analysis, Monte Carlo analysis, Bayesian inference, application of statistical methods in finance and economy, especially methods used in capital market and risk management. The range of topics also included the design of experiments and survey sampling methodology, mainly for the social science purposes. The conference was attended by 93 participants from main academic centres in Poland (Białystok, Katowice, Kraków, Olsztyn, Opole, Poznań, Rzeszów, Szczecin, Toruń, Warszawa and Wrocław) and from abroad (Lithuania). The list of participants included scientists, academic tutors as well as representatives of the National Bank of Poland, local statistical offices and business. In 17 sessions 59 papers were presented, including 2 invited lectures.

The conference was opened by the Head of the Organizing Committee, **Professor Czesław Domański**. The subsequent speakers at the conference opening were **Professor Włodzimierz Nykiel**, the Rector of the University of Lodz and **Professor Paweł Starosta**, the Dean of the Faculty of Economics and Sociology of the University of Lodz.

After the opening ceremony, all participants had the opportunity to attend the invited lecture by **Professor Włodzimierz Okrasa** (Cardinal Stefan Wyszyński University in Warsaw) *Statistical Process as a Social Process of Quantification.* The second invited lecture was presented by **Professor Krzysztof Najman, Professor Kamila Migdał-Najman** and **Professor Mirosław Szreder** (University of Gdańsk), and was dedicated to *Big Data - new opportunities, new restrictions*. At the conference closing the participants attended the lecture by **Professor Czesław Domański** (University of Lodz) entitled *Meaning of models and statistics in the process of statistical inference.*

Among regular conference sessions, the historical plenary session was held, chaired by **Professor Janusz Wywiał**, dedicated to eminent Polish scientists. **Professor Tadeusz Gerstenkorn** (University of Lodz) recalled his memories of Włodzimierz Krysicki. **Professors Tadeusz Kowaleski (**University of Lodz**)** presented statistical threads in the work of Jan Długosz. **Professor Jan Kordos** (Warsaw Management University) presented his cooperation with Professor Tadeusz Walczak. **Professor Czesław Domański** (University of Lodz) presented Stanisław Staszic as a sympathizer of Łódź.

Other sessions were chaired respectively by:

SESSION II	Professor Włodzimierz Okrasa (Cardinal Stefan Wyszyński University in Warsaw)
SESSION III	Professor Andrzej Sokołowski (Cracow University of Economics)
SESSION IV A	Professor Tomasz Michalski (Warsaw School of Economics)
SESSION IV B	Professor Jerzy Korzeniewski (University of Lodz)
SESSION V	Professor Mirosław Krzyśko (Adam Mickiewicz University in Poznań)
SESSION VI A	Professor Adam Śliwiński (Warsaw School of Economics)
SESSION VI B	Professor Katarzyna Stąpor (The Silesian Technical University)
SESSION VII A	Professor Grażyna Trzpiot (University of Economics in Katowice)
SESSION VII B	Professor Marek Walesiak (Wrocław University of Economics)
SESSION VIII A	Professor Grzegorz Kończak (University of Economics in Katowice)

SESSION VIII B	Professor Elżbieta Gołata (Poznań University of Economics)
SESSION IX A	Professor Wojciech Zieliński (Warsaw University of Life Sciences)
SESSION IX B	Professor Małgorzata Markowska (Wrocław University of Economics)
SESSION X A	Professor Janusz Korol (Szczecin University)
SESSION X B	Professor Bronisław Ceranka (Poznań University of Life Sciences)
SESSION XI	Professor Józef Dziechciarz (Wrocław University of Economics)

The MSA 2015 conference was closed by the Chairman of the Organizing Committee, **Professor Czesław Domański**, who summarized the Conference as very effective and added that all discussions and doubts should become inspirations and strong motivations for further work for both scientists and practitioners. Finally, he thanked all the guests, conference partners and sponsors.

The next edition of Multivariate Statistical Analysis Conference **MSA 2016** is planned on **November 7-9, 2016** and will be held in **Łódź, Poland**. The Chairman of the Organizing Committee, **Professor Czesław Domański** informed that this will be the 35th edition of the conference and kindly invited all interested scientists, researchers and students to participate.

Prepared by:

Anna Jurek
Department of Statistical Methods, University of Lódź

Elżbieta Zalewska
Department of Statistical Methods, University of Lódź

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 353—354

REPORT

The XXIV Conference "Classification and Data Analysis – Theory and Applications" 14-16 September 2015, Gdańsk, Poland

The Section on Classification and Data Analysis of the Polish Statistical Association (SKAD) held its conference from 14th to 16th of September 2015 in Gdańsk. It was XXIV Conference "Classification and Data Analysis – Theory and Applications". The conference was jointly organized by SKAD and Department of Statistics of University of Gdańsk. The organizing committee was chaired by Professor dr hab. Mirosław Szreder and dr hab. Krzysztof Najman, supported by dr hab. Kamila Migdał Najman, dr hab. Anna Zamojska and Mrs. Anna Nowicka. The conference was financially supported by the National Bank of Poland.

The following theoretical and applied areas were covered by the conference:

a) theory – taxonomy, discriminant analysis, linear ordering methods, multivariate statistical analysis, methods of analysis of continuous, discrete and symbolic data, graphical methods;
b) applications – financial data, marketing data, spatial data, and other applications in medicine, psychology, archaeology, etc., and computer applications of statistical methods.

The main objective of the conference was the presentation of scientific results in the area of theory and applications of classification and data analysis problems. It serves as an annual forum to discuss recent developments in the above areas as well as to identify and suggest some directions for future research agenda.

The conference was attended by 81 researchers from the following universities: University of Economics in Katowice, Cracow University of Economics, Poznań University of Economics, Wrocław University of Economics, University of Gdańsk, AGH University of Science and Technology in Cracow, Łódź University of Technology, Gdańsk University of Technology, Opole University of Technology, Wrocław University of Technology, Warsaw University of Life Sciences – SGGW, Warsaw School of Economics, Adam Mickiewicz University in Poznań, Jan Kochanowski University in Kielce, University of Łódź, Nicolaus Copernicus University in Toruń, Poznań University of Life Sciences, University of Szczecin, University of Białystok. Also, representatives of the National Bank of Poland and PBS market research company, participated in the conference.

During two plenary sessions and 13 parallel sessions there were 58 papers presented covering both theoretical and applied problems of classification and data analysis. In addition, 14 posters were presented during the poster session.

Sessions were chaired by the following professors: Józef Pociecha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz and Mirosław Szreder.

On the first day of the conference an annual meeting of members of SKAD took place, chaired by Professor Józef Pociecha. During this meeting the following issues were discussed:

A. Report on the activities of the members of the Section on Classification and Data Analysis of Polish Statistical Association – there are 231 members. The most important activities included:
   - participation in the International Federation of Classification Societies conference in Bologna – 19 persons presented 15 papers at that conference;
   - participation in the European Conference on Data Analysis in Colchester – 18 persons presented 15 papers.

B. Information about upcoming conferences.

C. Organization of the next SKAD conference – it will be organized by the Department of Statistical Methods at the University of Łódź.

D. Election of SKAD representative to International Federation of Classification Societies Council. Professor Andrzej Sokołowski was elected for the 2016-2019 term.

E. Discussion about future activities of SKAD.

Prepared by:

Krzysztof Jajuga

Marek Walesiak

STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 355–358

ABOUT THE AUTHORS

Bal-Domańska Beata is an Assistant Professor at Wroclaw University of Economics, Department of Regional Economy (Poland) and as a Consultant of the Local Data Bank Centre in Statistical Office in Wrocław. Currently, her research interest embraces regional economy in particular analysis (taxonomy and econometric) in socio-economic issues, the assessment of convergence process as well as sustainable development. She has published more than 80 research papers in international or national journals and presentation at conferences; she also took a part in several scientific projects (grants). She is – among others – a member of Polish Statistical Association.

Fabian Piotr received his MSc degree in computer science from the Faculty of Automatic Control, Electronics and Computer Science of the Silesian University of Technology in Gliwice, Poland. There, in 2004, he received a PhD degree. He continues to work at the same university. His research interests include speech processing and recognition, pattern recognition, machine learning, applications for people with disabilities.

Hozer-Koćmiel Marta is an Assistant Professor at the Department of Statistics, University of Szczecin, Faculty of Economics and Management. Her main research interest is the usage of quantitative methods to study the differences and similarities of economic behaviour of women and men. She is also interested in time use surveys, methods of valuation of household work, entrepreneurship and sustainable development from gender perspective.

Hudson Irene is a mathematical statistician with an international reputation in the development of new statistical methods for climate change research and methods to analyse markers/indicators in diverse research areas of drug discovery, bio-statistics and stroke/brain research. She is Chair of Statistics in the School of Mathematical and Physical Sciences, University of Newcastle Australia. Professor Hudson has published 220 research papers in international/national journals and conferences, and published 2 books and 19 book chapters/monographs. She is a nominated member of an ARC Centre of Excellence in therapeutics/diagnostics, a collaborator with the WHO and Centre of Generational Health & Aging. Hudson is on the Editorial board of Climatic Change.

Khan Muhammad Shuaib is a lecturer in Statistics in the Victoria University Sydney Campus. Currently, he is doing PhD in Statistics in the School of Mathematical and Physical Sciences, The University of Newcastle, Australia. He has published a number of research papers in international/national peer-reviewed Journals of Statistics and conferences. His research area includes lifetime distribution theory and reliability analysis.

Khetan Mukti is a Senior Research Fellow in the Department of Applied Mathematics, Indian School of Mines, Dhanbad, India. Her research area is sampling techniques. She has published 17 research papers in reputed international journals. She has won the best paper presentation award at the 68th Annual Conference of the Indian Society of Agricultural Statistics. She has completed M.Sc. (Statistics) from the Department of Statistics, Banaras Hindu University, Varanasi, India.

King Robert is a mathematical statistician with an international reputation in the development of new methods in lifetime distributions and quantile defined distributions as well as in a number of applied statistical fields. He is a lecturer in the School of Mathematical and Physical Sciences, The University of Newcastle, Australia. He has received his PhD degree from the Queensland University of Technology Australia. He has published more than 60 research papers in international/national journals and conferences.

Korzeniewski Jerzy is an Assistant Professor at the Department of Statistical Methods, University of Lodz. His main research domain is cluster analysis and attribute selection, especially with respect to further object classification. He is also interested in all applications of cluster analysis in financial data, particularly in time series clustering.

Krężołek Dominik works as an Assistant Professor in the Department of Demography and Economics Statistics at the University of Economics in Katowice. His specialities focus on risk assessment and risk modelling, mainly in the area of the metals market. He is a co-author of three manuals related to risk analysis and the applications of statistical tools in data analysis. He is also an author of articles concerning non-classical risk measures. He actively cooperates with business area through the Research and Knowledge Transfer Centre (an internal unit of the University of Economics in Katowice). He is a member of the Polish Statistical Association and the Section on Classification and Data Analysis of the Polish Statistical Association (SKAD).

Landmesser Joanna is an Associate Professor at the Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences, and works in the Department of Econometrics and Statistics. She received the Dr. degree

in Economics in 2002 at the Bundeswehr University Munich in Germany and habilitated in 2014 in Economics ("The use of duration analysis methods for the survey of economic activity of people in Poland") at the Nicolaus Copernicus University in Toruń, Poland. Her research interests focus on evaluation of different policies on the labour market, counterfactual scenarios analysis, comparing the distributions of income, decomposition of income inequalities.

Lis Christian is an Assistant Professor at the Faculty of Economics and Management (Department of Statistics), Szczecin University, Poland. His main research interests are in economic convergence and growth theories, causes of welfare and poverty, utilisation of taxonomic methods in economics and multidimensional comparative analysis. He has published approx. 80 books and scientific papers. Dr hab. Christian Lis has carried out approx. 60 projects, for the Ministry of Infrastructure and Construction, the Ministry for Maritime Economy, Prime Minister's Office, Marshal Office of the West Pomeranian Region, Szczecin and Świnoujście Sea Port Authority, among others. In 2009 he established the Economic Research Centre in Szczecin. The company is specialized in feasibility studies, cost-benefit analysis (*CBA*) for infrastructure investment in sea ports and other analyses on maritime economy.

Maurya Shweta received a MSc in Statistics from the Department of Statistics of Banaras Hindu University in Varanasi, India, in 2012. She received a PhD in applied statistics from Department of Applied Mathematics of Indian School of Mines in Dhanbad, India, in 2016. Currently, she is working as an advisor to Maersk Global Service Centre in Advanced Analytics department. Her research interests include sampling theory.

Mussini Mauro is an Assistant Professor of Economic Statistics at the Department of Economics, University of Verona. He received his PhD degree in Statistics from the University of Milan Bicocca in 2008. His research interests include inequality and poverty measurement, statistical models for energy and environmental data analysis, statistical methods for the integration of data from different sources.

Osaulenko Oleksandr is the Rector of the National Academy of Statistics, Accounting and Audit, Doctor of Public Administration, Professor, Corresponding Member of the National Academy of Sciences of Ukraine, Honored Economist of Ukraine. During 1996-2014, he headed the national statistical office of Ukraine. He is an author of over 200 scientific works, including 20 monographs and handbooks on the problems of statistics and public administration.

Pandey Ranjita is an Assistant Professor at the Department of Statistics, University of Delhi. She received her D.Phil. degree from University of Allahabad in 2002. Her research interests include ecological modelling, demography, imputation methods and Bayesian inference.

Singh G. N. is a Professor of Statistics at the Department of Applied Mathematics, Indian School of Mines, Dhanbad, India. His research area is sample surveys. Professor Singh has guided many PhD students and published more than 120 research papers in international journals of repute. Professor Singh is a reviewer of several prestigious journals in statistics.

Skrodzka Iwona is an Assistant Professor at the Department of Econometrics and Statistics, University of Bialystok. For several years, her scientific interests have been focused on issues of human and intellectual capital (measurement and analysis of the impact on economic development). Currently, her research concerns smart growth in countries and regions of the European Union. She applies the methods of multivariate statistical analysis.

Smolarczyk Tomasz is a PhD student at the Faculty of Automatic Control, Electronics and Computer Science at Silesian University of Technology. His research interests are feature selection and extraction, machine learning and data mining. In his professional career, he uses data analysis to solve real-world problems, mostly in marketing and sales area.

Sobczak Elżbieta is an Associate Professor at Wroclaw University of Economics, Department of the Regional Economics. Her main fields of research interest include quantitative methods in economics, classification and data analysis, and multivariate statistical analysis in regional research. She is also interested in econometric regional analysis and segmentation methods of foreign markets. She is a member of Polish Classification Society (SKAD), International Federation of Classification Societies (IFCS) and Polish section of Regional Studies Association (RSA).

Stąpor Katarzyna is a Professor at the Department of Automatic Control, Electronics and Computer Science in the Silesian Technical University in Gliwice, Poland. Her research interests include statistical pattern recognition, multivariate statistical analysis, discriminant analysis in particular, computer vision and bioinformatics. Professor Stąpor has published more than 80 research papers, including one monograph on pattern recognition. She is the co-author of the integrated information management systems and the classification system supporting glaucoma diagnosis in ophthalmology.