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
Beata Witek: Secretary
b.witek@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, Autumn 2013
Vol. 14, No. 3, pp. 349—356

EDITOR'S NOTE AND ACKNOWLEDGEMENTS

This issue concludes the journal's activities for the year 2013. Therefore, besides of a short preview of the articles included in it (as usual), several points on related matters that have took place over the last year period are being mentioned as important to the journal itself or to the community of statisticians at large.

First of all, the journal has obtained a new mark of recognition, this time from the Index Copernicus International which included it into its system of citation, 'IC Journals Master List 2012'. On the community side, it was possible to deliver the Jerzy Splawa-Neyman Medal - that was established on the occasion of the 100th anniversary of the Polish Statistical Association - to Professor Graham Kalton, who was awarded the Medal for his extraordinary contribution to the contemporary statistics and to the survey methodology in particular. Since Graham Kalton could not attend the Association's meeting on April 2012 in Poznan during which the Medal was presented to its recipients, it was a great honor and pleasure for writing these words to hand it over to him personally; a brief note on Professor Kalton's main professional activities is on page 521. This issue provides us also with an opportunity to express our gratefulness, both on behalf of the Editorial Office and of all the journal's partners - especially of authors and readers - to our key collaborators, peer-reviewers, who kindly shared their knowledge and expertise with us and with authors in order to improve the quality of the published papers. We feel truly indebted to them and list their names in the 'Acknowledgements' below.

The present issue contains a set of ten papers divided into three categories: sampling methods and estimation; research articles; and other articles. While the first and the second constitute a regular part of the journal, the last one is composed of papers based on presentations given to the Multivariate Statistical Analysis Conference that was held on November 2013 at the University of Lodz. As the product of an event that took place within a series of such meetings being organized since 1981 on an annual basis, the papers were considered for publication jointly with Professor Czeslaw Domanski who kindly acted as a guest co-editor of (this part of) the issue being especially instrumental in arranging for them and making pre-selection among the conference presentations for this purpose.

The first group of articles begins with **Alina Jędrzejczak's** and **Jan Kubacki's** paper entitled ***Estimation of Income Inequality and the Poverty Rate in Poland, by Region and Family Type.*** Inspired by observation that growth may not be shared equally and economic crises may further widen the gap between the wealthy and the poor, they concentrate on inequality and poverty analysis of households and groups of populations distinguished by family type. They aimed at presenting some income inequality and poverty estimates based on data from the Polish Household Budget Survey using both direct estimation methods and a model-based approach. **Prayas Sharma** and **Rajesh Singh** propose ***Improved Estimators for Simple Random Sampling and Stratified Random Sampling Under Second Order of Approximation*** taking for a point of departure two types of estimators offered recently in the literature for estimating population mean $\overline{Y}$: one by Singh and Solanki (2012) and the other by Koyuncu (2012). Up to the first order of approximation and under optimum conditions, the minimum mean squared error of both the estimators is equal to the MSE of the regression estimator. The authors have tried to find out the second order biases and mean square errors of these estimators using information on auxiliary variable based on simple random sampling. In consequence of comparison of the performance of these estimators using some numerical illustration, the authors conclude that the behavior of the estimators changes dramatically when we consider the terms up to the second order of approximation. In the paper ***A Ratio-Cum-Product Estimator of Finite Population Mean in Systematic Sampling*** by **Rajesh Tailor, Narendra K. Jatwa, Housila P. Singh** the problem of estimation of population mean is considered using information on two auxiliary variables in systematic sampling. The Singh (1967) estimator for estimation of population mean in systematic sampling is being extended along with derivation of the expressions for the bias and mean squared error of the suggested estimator. The suggested estimator is compared with existing estimators and the conditions under which it is more efficient are discussed.

A set of three articles constitutes the second part of this issue, *research articles.* **Marek Cierpial-Wolan** in the paper ***Processes in Transborder Areas – Significant Impact on the Economic Growth*** discusses some methodological issues related to the specificity of the socio-economic processes that are taking place across the state borders and in the so-called transborder areas. The needs for a coherent system of information on these processes are discussed and an example of the required type of multi-method research is being presented, encompassing household survey, border traffic survey, entrepreneurship survey, etc. Some results of the employed information generating system turned out to be unexpected, especially in terms of the estimates of the selected items in Balance

of Payment (BoP). In conclusion it is suggested that the appropriate changes should be made in the calculation of Gross Domestic Product (GDP).

Gurprit Grover, V. Sreenivas, Sudeep Khanna, Divya Seth in the paper ***Multi-State Markov Model: an Application to Liver Cirrhosis*** present a frame to estimate survival and death probabilities of the patients suffering from liver cirrhosis and HCC in the presence of competing risks. Using database of a Delhi hospital, authors employed a stochastic illness-death model to the process of two liver illness states (Cirrhosis and HCC) and two death states (death due to liver disease and death due to competing risk), for individuals being observed for one year. The survival and death probabilities of the individuals suffering from liver cirrhosis and HCC have been estimated using the method of maximum likelihood. The probability of staying in the cirrhotic state is estimated to be threefold higher than that of developing HCC (0.64/0.21). The probability of cirrhotic patient moving to HCC state is twice (0.21/0.11) the probability of dying due to liver disease. Markov model proves to be a useful tool for analysing chronic degenerative disease like liver cirrhosis, providing insight for both the researchers and policy makers to issues related to this complex problem.

In the next paper, ***The Properties of ATMs Development Stages - an Empirical Analysis*** by **Henryk Gurgul** and **Marcin Suder** the crucial problem of the ATMs network management is discussed, i.e. the saturation level of withdrawals or the mean level of withdrawals after dropping particular withdrawals realized in the initial time period, taking into account the length of elapsing time period necessary to reach saturation level. The paper aims to define average withdrawals after achieving saturation level and mean time necessary to stabilize withdrawals (based on historical data). Specifying some conditions (concerning similarity in terms of location and date of start) under which ATMs exhibit similar characteristics of the development effects leads to possibility to predict the size of time necessary to achieve saturation and the average withdrawal in the state of saturation.

Edyta Mazurek's and **Achille Vernizzi's** paper ***Some Considerations on Measuring the Progressive Principle Violations and the Potential Equity in Income Tax Systems*** discuss the issue of conditions (including axioms) for an equitable tax system and the consequences of their violation in terms of the distortions in evaluation of the redistributive effect of taxes. The authors calculate both the potential equity and the losses using Kakwani (1977) progressivity index and the Kakwani (1984) decomposition of the redistributive effect, focusing on the measure suggested by Kakwani and Lambert for the loss in potential equity under violations of the progressive principle. They propose a measure based on

the tax rate re-ranking index, calculated with respect to the ranking of pre-tax income distribution. Presentation of its analytical characteristics is followed by empirical results within simulated tax systems. Also, simulations compare Kakwani and Lambert's measure with the potential equity of a counterfactual tax distribution which respects the progressive principle and preserves the overall tax revenue using the approach proposed recently by Pellegrino and Vernizzi (2013).

The third part of the issue - which contains papers based on the above mentioned presentations at the Multivariate Statistical Analysis 2013 Conference in Lodz - starts with ***Empirical Evaluation of OCLUS and GenRandomClust Algorithms of Generating Cluster Structures*** by **Jerzy Korzeniewski**. Both the OCLUS algorithm (Steinley and Henson, 2005) and genRandomClust algorithm (Joe and Qiu, 2006) of generating multivariate cluster structures have the capacity of controlling cluster overlap, though they both do it in quite different ways, with indication on the former as a method having much easier and intuitive interpretation. In order to compare them at work multiple cluster structures were generated by each of them and grouped into the proper number of clusters using *k*-means. The groupings were assessed by means of divisions similarity index (modified Rand index) and the comparison criterion was the behaviour of the overlap parameters of structures. Particular attention was given to checking the existence of an overlap parameter limit for the classical grouping procedures as well as uniform nature of overlap control with respect to all clusters.

Marta Małecka's and **Dorota Pekasiewicz's** paper ***A Modification of the Probability Weighted Method of Moments and its Application to Estimate the Financial Return Distribution Tail*** discusses the issue of fitting the tail of the random variable with an unknown distribution. It plays a key role in finance statistics enabling estimation of high quantiles and subsequently offers risk measures. The parametric estimation of fat tails is based on the convergence to the generalized Pareto distribution (GPD). The paper explored the probability weighted method of moments (PWMM) applied to estimation of the GPD parameters, focusing on the tail index, commonly used to characterize the degree of tail fatness. A modification of the PWMM method is suggested due to application of the level crossing empirical distribution function. Using simulation techniques statistical properties of the GPD shape parameter estimates - with reference to the PWMM algorithm specification - were examined, and the results showed that the choice of the level crossing empirical distribution function may improve the statistical properties of the PWMM estimates.

The paper ***The Distribution of the Number of Claims in the Third Party's Motor Liability Insurance*** by **Anna Szymańska** (which concludes this section)

addresses the automobile insurance tarification problem which consists of a sequence of two questions, or stages of the assessment process. The first one concerns assessment of the net premiums on the basis of known risk factors, called *a priori* ratemaking. The second, called *a posteriori* ratemaking, accounts for the driver's claims history in the premium. Each of them requires the actuary's selection of the theoretical distribution of the number of claims in the portfolio. The paper presents methods of consistency evaluation of the empirical and theoretical distributions used in motor insurance. Some illustrations are provided using data from different European markets.

This issue is concluded by a review of the recently released by The National Academies Press report of the CNSTAT Panel (Committee on National Statistics) on Measuring Subjective Well-Being in a Policy-Relevant Framework, by Włodzimierz Okrasa. Important contributions of the Panel`s report to excelling conceptual and methodological approaches to measuring subjective well-being as a part of research activities of official statistics, that is currently increasing world-wide, is being emphasized, along with originality and usefulness of its recommendations.

Włodzimierz Okrasa
Editor

ACKNOWLEDGEMENTS TO REVIEWERS

The Editor and Editorial Board wish to thank the following persons who served from 31 December 2012 to 31 December 2013 as peer-reviewers of manuscripts for the ***Statistics in Transition new series* – Volume 14, Numbers 1–3**; the authors' work has benefited from their feedback.

Alagarajan Manoj, International Institute for Population Sciences, Maharashtra, India

Baszczyńska Aleksandra, University of Lodz, Poland

Białek Jacek, University of Lodz, Poland

Choudhury Sanjib, National Institute of Technology, Nagaland, India

Clark Andrew, Paris School of Economics, France

Dittmann Paweł, Wroclaw University of Economics, Poland

Domański Czesław, University of Lodz, Poland

Domański Henryk, Polish Academy of Sciences, Warsaw, Poland

Dwivedi Alok, Texas Tech University Health Sciences Center, El Paso, USA

Fabrizi Enrico, Università Cattolica del Sacro Cuore, Piacenza, Italy

Gabler Siegfried, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

Gamrot Wojciech, University of Economics in Katowice, Poland

Garg Neha, Statistics School of Sciences, IGNOU, New Delhi, India

Gedam Vinayak K., University of Pune, India

Getka-Wilczyńska Elżbieta, Warsaw School of Economics, Poland

Gołaś Zbigniew, Poznan University of Life Sciences, Poland

Gołata Elżbieta, University of Economics, Poznan, Poland

Gurgul Henryk, AGH University of Science and Technology, Cracow, Poland

Jajuga Krzysztof, Wroclaw University of Economics, Poland

Jędrzejczak Alina, University of Lodz, Poland

Joe Dominique, University of Lausanne, Switzerland

Kadilar Cem, Hacettepe University, Ankara, Turkey

Kapadia Asha Seth, University of Texas School of Public Health, Houston, USA

Kordos Jan, Warsaw Management Academy, and Central Statistical Office of Poland

Körner Thomas, Federal Statistical Office, Wiesbaden, Germany

Korzeniewski Jerzy, University of Lodz, Poland

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 357

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., followed by a hard copy addressed to
Prof. Wlodzimierz Okrasa,
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://www.stat.gov.pl/pts/15_ENG_HTML.htm

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 359–378

ESTIMATION OF INCOME INEQUALITY AND THE POVERTY RATE IN POLAND, BY REGION AND FAMILY TYPE

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

ABSTRACT

High income inequality can be a source of serious socio-economic problems, such as increasing poverty, social stratification and polarization. Periods of pronounced economic growth or recession may impact different groups of earners differently. Growth may not be shared equally and economic crises may further widen gaps between the wealthiest and poorest sectors. Poverty affects all ages but children are disproportionately affected by it. The reliable inequality and poverty analysis of both total population of households and subpopulations by various family types can be a helpful piece of information for economists and social policy makers. The main objective of the paper was to present some income inequality and poverty estimates with the application to the Polish data coming from the Household Budget Survey. Besides direct estimation methods, the model based approach was taken into regard. Standard errors of estimates were also considered in the paper.

Key words: income inequality, poverty, variance estimation, small area statistics.

1. Introduction

The range of survey data analysis has expanded enormously over time in response to growing demands of policy makers. Recently, the demand for estimates at a small level of aggregation has increased, in contrast to national estimates that were commonly used in the past. Since income inequality in Poland increased significantly in the period of transformation from the centrally planned to the market economy, reliable inequality and poverty analysis of the total population of households and subpopulations by various family types can provide helpful information for economists and social policy makers.

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

In the paper, some direct and indirect model-based estimation methods for income inequality and poverty parameters are presented and applied. Income inequality is measured by the Gini and Zenga indices. Among the indicators of poverty we consider: at risk of poverty rate, poverty gap and poverty severity with special attention paid to the estimation of their standard errors. The estimates of inequality measures are produced only for large domains (regions or family types considered separately) using direct estimation, while poverty related measures are also calculated for small domains that require small area estimation. For subsets of the Polish population cross-classified by region and family type small area estimators based on linear mixed models are used.

Measures of income inequality and poverty are presented in Section 2. Section 3 provides a brief survey of direct variance estimation methods, while in Section 4 the outline of EBLUP theory is presented. Some empirical applications based on Polish HBS data are included in Section 5.

2. Measures of income inequality and poverty

Income inequality refers to the degree of difference in earnings among various individuals or segments of a population. Measures of inequality, also called concentration coefficients, are widely used to study income, welfare and poverty issues. They can also be helpful in analyzing the efficiency of a tax policy or in measuring the level of social stratification and polarization. They are most frequently applied to dynamic comparisons, i.e. comparing inequality across time. Among numerous inequality measures, the Gini and Zenga coefficients are of the greatest importance. The Gini concentration coefficient is the most widely used measure of income inequality, mainly because of its clear economic interpretation. The Zenga „point concentration" measure based on the Zenga curve has recently received some attention in the literature.

The Gini inequality index, based on the Lorenz curve, can be expressed as follows:

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

where: $p = F(y)$ is a cumulative distribution function of income, $L(p)$- the Lorenz function given by the following formula:

$$L(p) = \mu^{-1}\int_0^p F^{-1}(t)dt, \tag{2}$$

where μ denotes the expected value of a random variable Y and $F^{-1}(p)$ is the p^{th} quantile.

One can estimate the value of the Gini index from the survey data using the following formula:

$$\hat{G} = \frac{2\sum_{i=1}^{n}(w_i y_{(i)} \sum_{j=1}^{i} w_j) - \sum_{i=1}^{n} w_i y_{(i)}}{(\sum_{i=1}^{n} w_i) \sum_{i=1}^{n} w_i y_{(i)}} - 1 \ , \tag{3}$$

where: $y_{(i)}$ – household incomes in a non-descending order, w_i- survey weight for *i*-th economic unit, $\sum_{j=1}^{i} w_j$ - rank of *i*-th economic unit in *n*-element sample.

An alternative to the Lorenz curve (2) is the concentration curve proposed by Zenga (1984, 1990), defined in terms of quantiles of the size distribution and the corresponding quantiles of the first-moment distribution. It is called "*point concentration measure*", as it is sensitive to changes of inequality in each part (point) of a population.

The Zenga point measure of inequality is based on the relation between income and population quantiles:

$$Z_p = [y_p^* - y_p] / y_p^*, \tag{4}$$

where y_p denotes the population p^{th} quantile and y_p^* is the corresponding income quantile defined as follows:

$$y_p^* = Q^{-1}(p). \tag{5}$$

The function $Q(p)$, called *first-moment distribution function*, can be interpreted as cumulative income share related to the mean income. Thus, the Zenga approach consists in comparing the abscissas at which $F(p)$ and $Q(p)$ take the same value p.

Zenga synthetic inequality index can be expressed as the area below the Zenga curve (4), and is defined as simple arithmetic mean of point concentration measures Z_p, $p \in \langle 0.1 \rangle$:

$$Z = \int_0^1 Z_p dp \ . \tag{6}$$

The commonly used nonparametric estimator of the Zenga index (6) was introduced by Aly and Hervas (1999) and can be expressed by the following equation:

$$\hat{Z} = 1 - \frac{1}{n\bar{y}} \left\{ y_{1:n} + \sum_{j=1}^{n-1} y_{\left\langle \frac{\sum_{i=1}^{j} y_{i:n}}{\bar{y}} \right\rangle :n} \right\} , \tag{7}$$

where: $y_{i:n}$ – *i*-th order statistics in *n*-element sample based on weighted data,
$\bar{y}$ – sample arithmetic mean.

The poverty measures are statistical functions which translate the comparison of the indicator of well-being and the poverty line, made for each household, into one aggregate number for the population as a whole or a population sub-group. Since the publication of Sen (1976) article on the axiomatic approach to the measurement of poverty, several indices of poverty have been developed that make use of the three basic poverty indicators (Panek, 2008). The most popular poverty measure is *headcount ratio* also called *at-risk-of-poverty rate ARPR*. It represents the share of the population whose equivalent income or consumption is below the poverty line:

$$H = \frac{n_p}{n} 100, \tag{8}$$

where: n_p- number of the poor, n- total number of households.

Poverty gap index provides information regarding the distance of households from the poverty line. This measure captures the mean aggregate income or consumption shortfall relative to the poverty line across the whole population. It is obtained by adding up all the shortfalls of the poor and dividing the total by the population size:

$$PG = \frac{1}{n} \sum_{i=1}^{n_p} \left(\frac{y^* - y_i}{y^*} \right), \tag{9}$$

where: y^* denotes the poverty line (poverty threshold). The poverty gap (9) can be used as a measure of the minimum amount that one would have to transfer to the poor under perfect targeting, i.e. each poor person getting exactly the amount he/she needs to be lifted out of poverty so that they are all brought out of poverty. By replacing the number of households n by the number of the poor n_p in the formula (9), we obtain the alternative poverty gap index:

$$PG_p = \frac{1}{n_p} \sum_{i=1}^{n_p} \left(\frac{y^* - y_i}{y^*} \right). \tag{10}$$

Poverty severity index (squared poverty gap) takes into account not only the distance separating the poor from the poverty line (the poverty gap), but also the inequality among the poor. That is, higher weights are placed on those households which are further away from the poverty line.

$$PS_p = \frac{1}{n_p} \sum_{i=1}^{n_p} \left(\frac{y^* - y_i}{y^*} \right)^2. \tag{11}$$

According to their definitions, the headcount index (ARPR), the poverty gap index and the poverty severity index (Panek, 2008) can be expressed as ratio

estimators. Thus, the precision estimation algorithms can be similar to the algorithm for a ratio estimate. The headcount index estimator can be expressed as follows:

$$\hat{H} = \frac{\sum_{i \in U} I_i w_i}{\sum_{i \in U} w_i}, \tag{12}$$

while the estimator of poverty gap index given by (10) takes the form:

$$\hat{PG}_p = \frac{\sum_{i \in U_p} ((y^* - y_i)/y^*) w_i}{\sum_{i \in U_p} w_i}, \tag{13}$$

where: I_i – indicator function taking value 1 when *i*-th household equivalent income is below a poverty line, and taking value 0 in the opposite situation, w_i – survey weight for *i*-th economic unit, U_p denotes the poor families population (or subpopulation).

3. Methods of variance estimation

The precision of an estimator is usually discussed in terms of its variance or standard error. When the standard errors of inequality and poverty measures are large, many conclusions about the comparisons over time and between groups may not be warranted. For most income concentration measures, the Gini and Zenga indices included, explicit variance estimators are theoretically complicated, i.e. it is hard to derive general mathematical formulas for nonlinear statistics, especially when the sampling design is complex. Also, most widely used poverty statistics are nonlinear functions of sampling observations so their standard errors are rather difficult to obtain and have been rarely reported in practice. To solve this problem, some special approximate techniques for variance estimation can be used. They include: Taylor linearization technique, random groups method, jackknife, bootstrap, balanced half samples, also called balanced repeated replication BRR. (Wolter, 2003; Särndal et al., 1997).

In the context of inequality measures, the Taylor linearization, the bootstrap and the parametric approach based on a theoretical income distribution model are the methods of variance estimation most often used (Jędrzejczak, 2011); while standard errors of poverty statistics are usually estimated by means of the bootstrap and balance repeated replication. An interesting outline of the variance estimations methods for the Gini index was offered by Langel and Tillé (2013).

The parametric approach uses a model-based variance with respect to hypothesized data generating process, provided that an empirical income distribution can be approximated by a theoretical model described by a probability density function $f(y,\boldsymbol{\theta})$. Applying the maximum likelihood (ML)

theory, the estimators obtained are asymptotically unbiased and normally distributed with variances given by the Cramer-Rao bound. Let us assume that an inequality measure of interest can be expressed as a function $g(\boldsymbol{\theta})$ of the model parameters $\boldsymbol{\theta}$. The variance of the ML estimator of an inequality measure $g(\boldsymbol{\theta})$ takes the form:

$$\mathrm{D}^2[\mathrm{g}(\hat{\boldsymbol{\theta}})] = \left[\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \mathbf{I}_\theta^{-1}\left[\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right], \tag{14}$$

where: $\mathbf{I}_{\boldsymbol{\theta}}$ denotes the Fisher information matrix.

The estimator of the variance (14) can be obtained by replacing the unknown parameter values $\boldsymbol{\theta}$ by their large sample ML estimates $\hat{\boldsymbol{\theta}}$. It preserves the asymptotic properties of maximum likelihood estimators (Zehna, 1966). The parametric approach can be very effective for large samples, assuming that the income distribution model as well as the parametric formula for an inequality statistic are both known.

Another method of variance estimation that can be used for poverty and inequality measures is balanced repeated replication. It proves especially useful when data come from a complex survey design with a large number of strata.

Let a stratified sampling frame with H strata be considered, with two subsets of primary sampling units (PSU) obtained for each stratum. They should be constructed in such a way that every stratum consists of two subsamples, each of them having similar number of units. A half-sample is a set consisting of one of the two subsets for each stratum. The number of all possible half-samples is 2^H, what may cause complication when the number of strata is large. To avoid such difficulties, we can choose the balanced set of R half-samples, so that the number of variants is significantly smaller than 2^H. The subset of balanced half-samples can be defined as a matrix of dimension $R \times H$ with the elements (r,h) equal δ_{rh} = +1 or –1, indicating whether the PSU from the h-th stratum selected for the r-th half sample is the first or the second PSU. The set of R half-samples is considered as balanced, if

$$\sum_{r=1}^{H} \delta_{rh}\delta_{rh'} = 0 \qquad \forall h = h'. \tag{15}$$

Balanced matrix RH can be obtained from Hadamard matrix, that has dimensions $R \times R$. The rows of Hadamard matrix denote half-samples while columns denote the strata, and the following condition is satisfied $H+1 \le R \le H+4$. Because the lines and columns in such a matrix are mutually orthogonal the half-samples selected are mutually independent (examples of Hadamard matrices to be found in: Bruch, Münnich and Zins, 2011).

The weights for selected elements may be equalized and are usually multiplied by 2 (Shao et al., 1998). Next, the estimated values for parameter of

interest $\hat{\theta}_r$ are determined by balanced repeated replication for each half-sample. The standard variance estimator can be expressed by the following formula:

$$\hat{V}_{BRR}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \hat{\theta}_{StrRS})^2 \ , \tag{16}$$

where $\hat{\theta}_{StrRS}$ is parameter estimate for the whole sample in the case of stratified random sampling (StrRS).

4. Model-based approach and EBLUP estimation

Sample survey data can be used to derive reliable direct estimates for large domains (discussed in Section 3), but sample sizes in small domains are seldom large enough for direct estimators to provide adequate precision for these domains. Thus, it is necessary to employ indirect estimation methods that borrow strength from related areas. Many subpopulation parameters, including means and totals, can be expressed as linear combinations of fixed and random effects of small area models. Best linear unbiased prediction (BLUP) estimators of such parameters can be obtained in a classical way using BLUP estimation procedure. BLUP estimators minimize Mean Square Error (MSE) within the class of linear unbiased estimators and do not depend on the normality of random effects. Maximum likelihood (ML) or restricted maximum likelihood (REML) methods can be used to estimate the variance and covariance components, assuming normality.

The EBLUP procedure has been applied in many important statistical surveys conducted all over the world. The pioneer work in this area was that of Fay and Herriot (Fay, Herriot, 1979), where the EBLUP technique was used for evaluating per capita income and some other statistics obtained for counties. The model-based approach to small area estimation of inequality indices for regions in Poland was applied in the paper of Jędrzejczak, Kubacki (2010). The authors provided empirical Bayes (EB) and empirical best linear unbiased prediction (EBLUP) estimators under area level models. Molina and Rao (2010) estimated poverty indicators as examples of nonlinear small area population parameters by using the empirical Bayes (best) method, based on a nested error model. Hierarchical Bayes multivariate estimation of poverty rates for small domains was lately discussed by Fabrizi et al. (2011).

Many applications of EBLUP in the context of small area estimation are based on a special kind of the general linear mixed model, widely known as basic area level model (Rao, 2003):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} + \mathbf{e} \tag{17}$$

where: **y** is $n \times 1$ vector of sample observations, **X** – known matrix of explanatory variables, **β** is a vector of linear regression coefficients, **v** denotes area-specific random effect vector, **e** is sampling error vector.

It is usually assumed that **v** and **e** are independently distributed with mean **0** and covariance matrices **G** and **R,** respectively. EBLUP estimator for the small area model given by (17) has the following form:

$$\boldsymbol{\theta}_{EBLUP} = \mathbf{X}\boldsymbol{\beta} + \mathbf{MGV}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{18}$$

where: $\boldsymbol{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{y}$, **M** is the identity matrix, **G** is the matrix with non-zero diagonal and its values are equal to σ_v^2, which is the model variance. It is usually computed using special iterative procedure that applies Fisher algorithm.

Mean square error estimate (MSE) of EBLUP can be obtained from the following formula:

$$MSE(\theta_{EBLUP}) = g_1(\hat{\delta}) - b_\delta^T(\hat{\delta})\nabla g_1(\hat{\delta}) + g_2(\hat{\delta}) + 2g_3(\hat{\delta}) \tag{19}$$

where δ is a variance dependent parameter. Using this formula we usually assume that the mean square error of EBLUP is the sum of three main elements g_1, g_2 and g_3 which are described by the following equations (Rao, 2003):

$$g_1(\hat{\delta}) = diag(\mathbf{G} - \mathbf{GV}^{-1}\mathbf{G}) \tag{20}$$

$$g_{2i}(\hat{\delta}) = (\mathbf{X_i} - \mathbf{m}_i^T\mathbf{GV}^{-1}\mathbf{X})(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X_i} - \mathbf{m}_i^T\mathbf{GV}^{-1}\mathbf{X}) \tag{21}$$

$$g_{3i}(\hat{\delta}) = (\mathbf{m}_i^T(\mathbf{V}^{-1} - \mathbf{G}(\mathbf{V}^{-1}\mathbf{V}^{-1}))\mathbf{V}(\mathbf{m}_i^T(\mathbf{V}^{-1} - \mathbf{G}(\mathbf{V}^{-1}\mathbf{V}^{-1}))^{\mathbf{T}})\mathbf{I} \tag{22}$$

where $\mathbf{m_i}$ is a vector with zeros for all elements with exception for the element having an index *i* while **I** is the inversed Fisher information matrix.

5. Application

The methods given above were applied to the estimation of inequality and poverty measures in Poland by region and family type. The basis for the calculations was micro data coming from the Polish Household Budget Survey (HBS) conducted in 2009. The data obtained from the household budget survey allow for the analysis of the living conditions of the population, being the basic source of information on the revenues and expenditure of the population. In 2009 the randomly selected sample covered 37,302 households, i.e. approximately 0.3% of the total number of households. The sample was selected by two-stage stratified sampling with unequal inclusion probabilities for primary sampling units. In order to maintain the relation between the structure of the surveyed population and the socio-demographic structure of the total population, the data obtained from the HBS were weighted with the structure of households by

number of persons and class of locality coming from the Population and Housing Census 2002.

The analysis has been conducted after dividing the overall sample by region NUTS1, constructed according to the EUROSTAT classification, and by 6 family types, classified according to the number of children. The variable of interest was household's available income that can be considered the basic characteristic of its economic condition. It is defined as a sum of households' current incomes from various sources reduced by prepayments on personal income tax made on behalf of a tax payer by tax-remitter (this is the case of income derived from hired work and social security benefits and other social benefits); by tax on income from property; taxes paid by self-employed persons, including professionals and individual farmers, and by social security and health insurance premiums.

To obtain the estimates of income inequality coefficients for selected subpopulations, the formulas (3) and (7) were applied, while poverty indicators were estimated by means of (12) and (13). Standard errors of the Gini and Zenga inequality measures were estimated using the parametric approach based on the three-parameter Burr type-III distribution, also called the Dagum model. The model is known to be well fitted to empirical income and wage distributions in different divisions. First, the maximum likelihood estimates of the Dagum model were calculated and then the formula (14) was applied to obtain the variances of the Gini and Zenga indices. The precision of poverty indicators was estimated by means of the balanced repeated replication technique BRR (See: formula (16)). It is worth mentioning that BRR estimators are typically estimators of design-based variances while methods based on ML, and specifically the one based on the Dagum model, are model-based. They generalise sample observations using predefined income distribution model instead of survey weights.

The results of the calculations are presented in Tables 1-3 and in Figures 1-6. To allow comparing the conditions of households of different sizes and different demographic structures, various income equivalence scales are used. The square root scale, popular in recent OECD publications, was applied in the paper. It is based on division of individual household income by the square root of the household size. As a poverty threshold we used 60% of median equivalised income value.

Table 1. Estimates of inequality and poverty indices with their standard errors by family type

No.	Number of children	Inequality coefficient		Poverty index		
		Gini	Zenga	Headcount ratio	Poverty gap	Poverty severity
1	0	0.36 (0.004)	0.37 (0.007)	15.4 (0.230)	26.2 (0.442)	12.2 (0.402)
2	1	0.32 (0.009)	0.30 (0.016)	13.2 (0.405)	29.7 (0.952)	16.0 (0.930)
3	2	0.33 (0.014)	0.31 (0.021)	16.9 (0.597)	28.4 (0.912)	14.2 (0.911)

Table 1. Estimates of inequality and poverty indices with their standard errors by family type (cont.)

No.	Number of children	Inequality coefficient		Poverty index		
		Gini	Zenga	Headcount ratio	Poverty gap	Poverty severity
4	3	0.32 (0.031)	0.30 (0.059)	27.1 (1.525)	27.8 (1.278)	13.3 (1.234)
5	4	0.30 (0.057)	0.28 (0.125)	33.5 (2.763)	31.1 (2.515)	15.7 (2.521)
6	5 …	0.29 (0.072)	0.26 (0.137)	39.6 (4.188)	29.0 (2.593)	15.8 (2.435)
7	Total	0.35 (0.003)	0.36 (0.005)	15.9 (0.253)	27.2 (0.396)	13.2 (0.381)

Source: Authors' calculations on the basis on HBS 2009.

Table 2. Estimates of inequality and poverty indices with their standard errors by region

No.	Region	Inequality coefficient		Poverty index		
		Gini	Zenga	Headcount ratio	Poverty gap	Poverty severity
1	Central	0.39 (0.006)	0.43 (0.011)	13.9 (0.419)	30.0 (1.137)	16.3 (1.228)
2	Southern	0.32 (0.008)	0.30 (0.015)	13.2 (0.405)	24.5 (0.755)	10.4 (0.673)
3	Eastern	0.35 (0.008)	0.36 (0.014)	23.5 (0.630)	28.6 (0.737)	14.3 (0.676)
4	North-western	0.33 (0.009)	0.32 (0.016)	14.5 (0.859)	23.7 (1.046)	10.1 (0.852)
5	South-western	0.35 (0.010)	0.36 (0.018)	15.3 (0.880)	28.5 (0.974)	14.3 (0.913)
6	Northern	0.34 (0.009)	0.35 (0.016)	15.7 (0.689)	26.7 (1.166)	13.0 (0.913)
7	Total	0.35 (0.003)	0.36 (0.005)	15.9 (0.253)	27.2 (0.396)	13.2 (0.381)

Source: Authors' calculations on the basis on HBS 2009.

Table 3. Estimates of poverty indices and their standard errors by region and family type

	Region	Number of children	Sample size	Poverty index					
				Headcount ratio		Poverty gap		Poverty severity	
				Estimate	Standard Error	Estimate	Standard Terror	Estimate	Standard error
1	Central	0	5469	18.047	0.629	27.482	0.676	13.173	0.655
		1	1393	13.516	0.892	31.206	2.502	17.854	2.784
		2	962	19.465	1.238	31.002	2.518	17.927	2.564
		3	216	33.325	4.478	34.347	3.226	19.130	3.276
		4	39	48.989	9.800	33.177	5.712	17.840	6.112
		5…	22	38.354	12.793	30.144	12.016	17.045	13.231
2	Southern	0	4871	13.510	0.633	24.624	0.958	10.033	0.794
		1	1374	13.407	1.095	26.455	1.358	12.261	1.267
		2	924	16.706	1.292	22.026	1.011	8.614	1.063
		3	235	29.368	3.447	21.598	2.850	7.384	1.968
		4	55	29.153	4.683	26.079	4.139	12.678	4.284
		5…	26	51.757	8.413	22.207	7.121	10.841	6.507
3	Eastern	0	4009	16.270	0.731	27.201	1.286	14.540	1.169
		1	1276	12.619	0.889	35.823	2.077	21.860	1.826
		2	914	14.765	1.050	33.173	2.165	18.960	2.130
		3	293	22.069	1.554	28.704	3.654	15.467	3.201
		4	76	23.275	5.846	20.817	4.501	6.446	2.439
		5…	35	37.941	8.772	32.748	6.853	17.063	7.159
4	North-western	0	3618	14.759	1.136	23.029	1.301	9.202	1.039
		1	1155	13.004	1.063	24.543	2.004	11.538	1.834
		2	719	17.003	1.934	23.129	1.654	9.522	1.390
		3	197	24.825	4.309	23.835	3.164	11.100	3.352
		4	47	32.767	6.989	32.015	9.682	20.310	9.774
		5…	23	38.979	7.540	13.452	3.689	3.154	2.089
5	South-western	0	2640	16.196	0.926	28.338	1.225	13.440	1.174
		1	752	12.962	1.062	33.131	3.589	20.641	3.696
		2	418	19.401	1.518	25.956	2.719	12.101	2.470
		3	109	30.216	6.194	24.314	2.995	9.256	2.364
		4	32	44.162	7.695	29.473	4.901	13.911	5.686
		5…	5	29.683	–	27.222	–	12.639	–
6	Northern	0	3378	13.037	0.648	25.336	1.286	12.460	1.424
		1	996	12.824	0.925	28.000	2.734	13.766	2.185
		2	720	18.372	2.048	27.459	2.103	13.194	1.841
		3	223	22.826	2.457	32.090	2.636	16.275	3.107
		4	48	42.105	6.384	35.491	5.465	18.427	5.683
		5…	33	35.697	12.660	36.954	8.481	25.483	9.988

Source: Authors' calculations on the basis on HBS 2009.

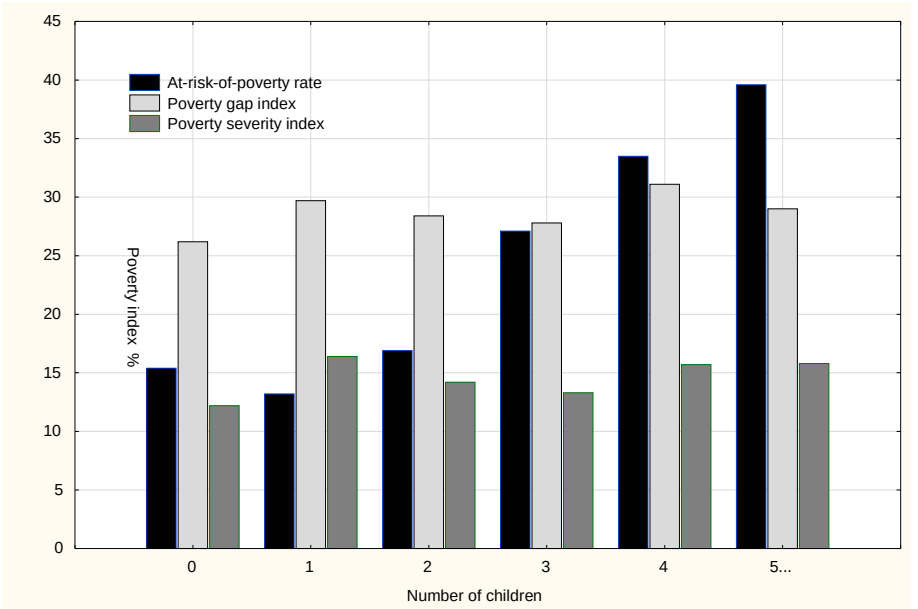


Figure 1. Poverty characteristics of families with different number of children

Source: Authors' calculations.

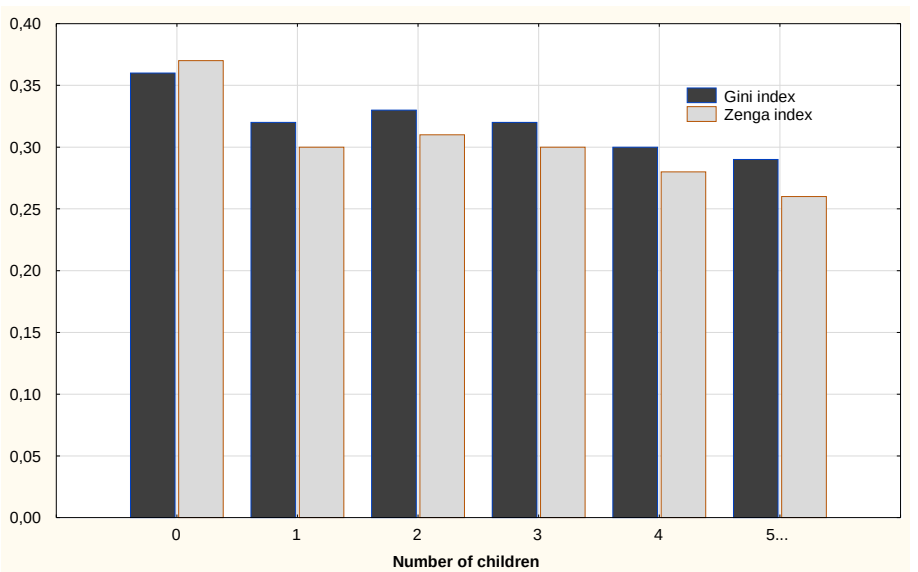


Figure 2. Inequality characteristics of families with different number of children

Source: Authors' calculations.

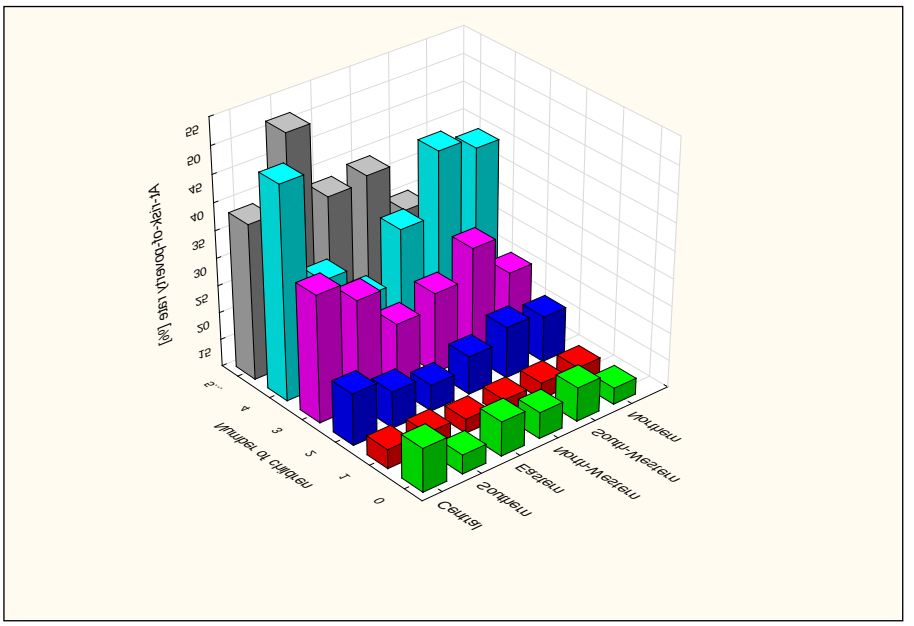


Figure 3. At- Risk-of-Poverty Rates (ARPR) by family type and region
Source: Authors' calculations.

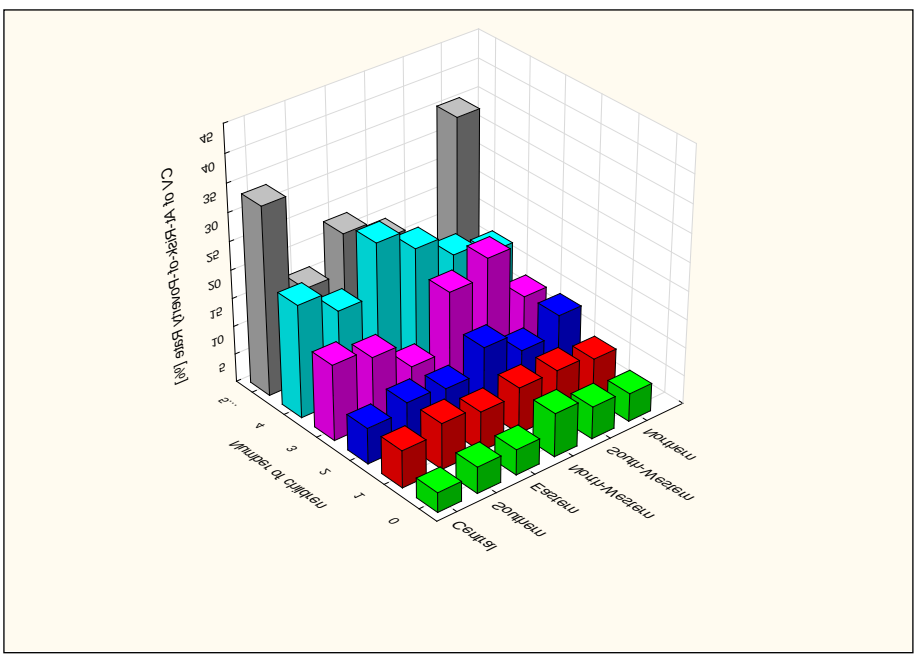


Figure 4. Coefficients of variation for ARPRs by family type and region
Source: Authors' calculations.

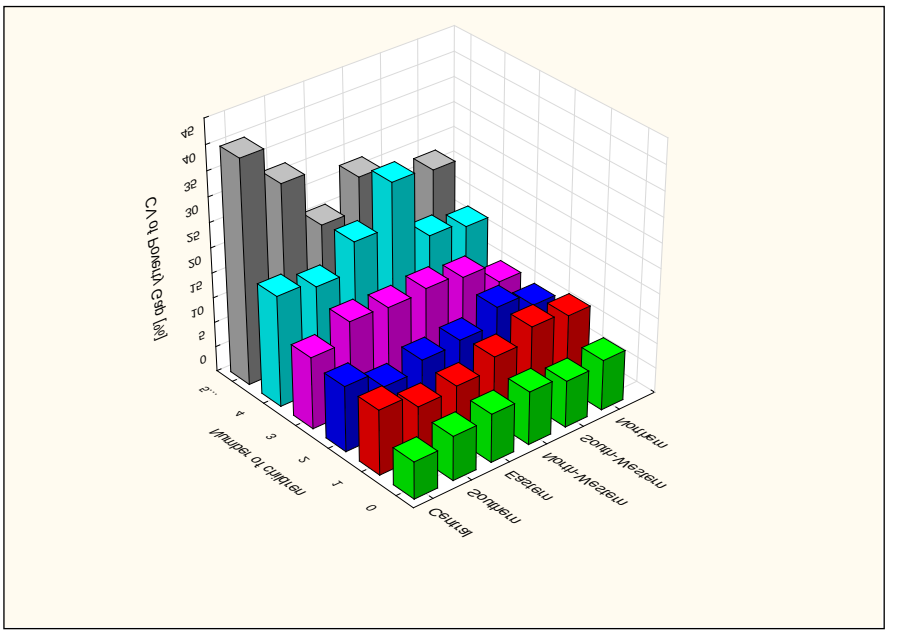


Figure 5. Coefficients of variation for Poverty Gaps by family type and region
Source: Authors' calculations.

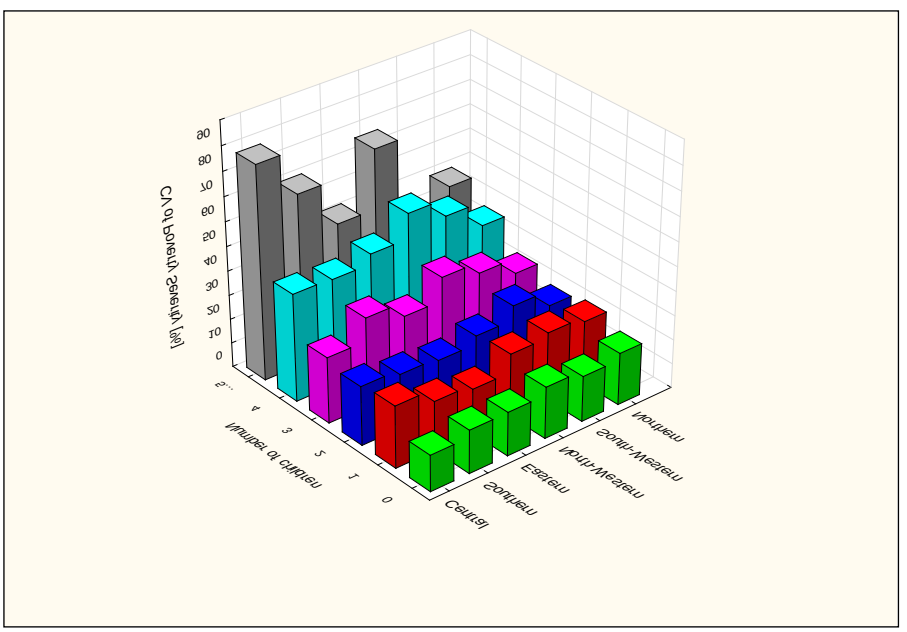


Figure 6. Coefficients of variation for Poverty Severity indices by family type and region

Source: Authors' calculations.

Tables 1 and 2 comprise estimates of inequality and poverty indices by family type and by region, separately. Diversification of household income, expressed by the Gini and Zenga coefficients (relatively stable since 2003), can be considered high in comparison with other European countries. The value of the Gini index for the total household was 0.35 while the Zenga index estimate was even higher (0.36). It is interesting to observe that the number of children is a factor clearly differentiating the level of income inequality and poverty of households. The higher the number of children, the higher the level of poverty indices (especially ARPR), and generally the lower is the level of income inequality. The latter is expressed, among others, by the fact that the poorest families with many children live mainly on social benefits, so their equivalised incomes are relatively similar. Families with five or more children present at-risk-of- poverty rate close to 40%, with simultaneously low level of the Gini index (0.29). On the other hand, discrepancies between regions are much smaller, with the *Eastern* region still having the worst position in terms of income (*ARPR*=23.5%, *G*=0.35, *Z*=0.36).

The detailed estimation results for the division of the entire population by family type (number of children), and at the same time by region, are presented in Table 3. The estimated values of three basic poverty measures for subpopulations, headcount ratio (ARPR), poverty gap index and poverty severity index, are accompanied by their standard errors estimated by means of balanced repeated estimation technique BRR. The calculations were done using WesVar package.

All the results presented in Tables 1-3 are supported by their estimated standard errors. Moreover, Figures 4-6 show the coefficients of variation (CV) for poverty coefficients outlined in Table 3. Analyzing the results of the calculations given above, one can easily notice that the precision of poverty indicators is unsatisfactory, especially when the division of households by family type and simultaneously by region is considered (See: Table 3 and Figures 4-6). The standard errors are relatively small only for the first household type, i.e. families without children, and usually account for about 5% of the corresponding estimates. For the remaining family types the efficiency of estimation is poor as coefficients of variation exceed 30% in many cases.

Thus, we have come to the conclusion that sample sizes of the subpopulations are apparently too small to provide reliable direct estimates. When a subpopulation is too small to yield direct estimates with adequate precision, it is regarded as a small area. Some indirect estimation methods based on borrowing strength in time and in space have been developed to overcome small area estimation problems (Rao, 2003). It is now generally accepted that when indirect

estimation is to be used, it should be based on explicit small areas models (See: §4). To improve the precision of head-count-ratio estimation for family types in regions, a model-based approach using the basic area level model (17) was applied. First, a standard linear regression model was constructed using some auxiliary sources of data that come from Polish Public Statistics and administrative registers. Regional per capita income GDP and the number of children NC played the role of explanatory variables in the model:

Model parameter	Coefficient value	Standard error	t-statistic	*p*-value
Intercept (α_0)	0.000865	0.050108	0.017259	0.986334
GDP (α_1)	0.001067	0.000477	2.239633	0.031971
NC (α_2)	0.056487	0.005636	10.02267	1.53E-11

For the specified model, the value of determination coefficient ($R^{2)}$ is 0.762, the value of corrected R^2 is equal to 0.747, the F statistics is $F(2.33)=52.735$ $p<0.00000$, and standard estimation error is equal to 0.05775.

The results of EBLUP estimation of ARPRs are summarized in Table 4. As it can be easily noticed, the model-based approach induced significant refinement for almost all subpopulations. Mean squared errors of EBLUP estimates are smaller than the corresponding standard deviations of direct estimates. In the last column we show the reduction of CV which exceeds 60% in some cases, yet on average it is equal to 21.54% .

Because the correlations between poverty gap (or poverty severity) and the same auxiliary variables as for the model of headcount ratio are relatively weak, we do not include the models for these cases. However, a preliminary correlation analysis for other explanatory variables, including the Gini and Zenga coefficients, has been carried out. It reveals a relatively strong correlation between the poverty and inequality measures that was found to be 0.46 for Gini and poverty gap and 0.43 for Gini and poverty severity (for the Zenga measure the correlation coefficients were 0.52 for poverty gap and 0.48 for poverty severity). It can be assumed that also other poverty related variables, e.g. unemployment rate, may be included in such computations. A more detailed analysis of such cases goes beyond the scope of the paper yet it may be performed in the future.

Table. 4. ARPR estimation results using direct and model-based approach

	Region	Number of children	Sample size	Direct estimation		EBLUP estimation		CV Red. [%]
				Estimate	Standard error	Estimate	Root mean squared error	
1	Central	0	5469	18.047	0.629	17.926	0.626	**0.1**
		1	1393	13.516	0.892	13.834	0.877	**3.9**
		2	962	19.465	1.238	19.901	1.197	**5.4**
		3	216	33.325	4.478	30.071	3.085	**23.7**
		4	39	48.989	9.800	34.564	3.895	**43.7**
		5…	22	38.354	12.793	37.416	4.210	**66.3**
2	Southern	0	4871	13.510	0.633	13.447	0.627	**0.6**
		1	1374	13.407	1.095	13.649	1.059	**5.0**
		2	924	16.706	1.292	17.190	1.233	**7.3**
		3	235	29.368	3.447	27.438	2.598	**19.3**
		4	55	29.153	4.683	29.739	3.063	**35.9**
		5…	26	51.757	8.413	37.250	3.740	**38.2**
3	Eastern	0	4009	16.270	0.731	15.996	0.722	**-0.5**
		1	1276	12.619	0.889	12.719	0.872	**2.7**
		2	914	14.765	1.050	15.105	1.021	**4.9**
		3	293	22.069	1.554	22.313	1.466	**6.7**
		4	76	23.275	5.846	26.895	3.362	**50.2**
		5…	35	37.941	8.772	33.555	3.837	**50.5**
4	North-western	0	3618	14.759	1.136	14.447	1.098	**1.2**
		1	1155	13.004	1.063	13.263	1.030	**5.0**
		2	719	17.003	1.934	17.870	1.743	**14.2**
		3	197	24.825	4.309	25.158	2.886	**33.9**
		4	47	32.767	6.989	30.577	3.441	**47.2**
		5…	23	38.979	7.540	35.436	3.666	**46.5**
5	South-western	0	2640	16.196	0.926	15.906	0.906	**0.4**
		1	752	12.962	1.062	13.246	1.029	**5.2**
		2	418	19.401	1.518	19.654	1.423	**7.5**
		3	109	30.216	6.194	26.696	3.245	**40.9**
		4	32	44.162	7.695	32.696	3.510	**38.4**
		5…	5	29.683	–	34.364	3.841	**-**
6	Northern	0	3378	13.037	0.648	12.962	0.641	**0.5**
		1	996	12.824	0.925	12.988	0.903	**3.6**
		2	720	18.372	2.048	18.786	1.826	**12.8**
		3	223	22.826	2.457	23.425	2.105	**16.5**
		4	48	42.105	6.384	32.281	3.379	**31.0**
		5…	33	35.697	12.660	34.064	3.927	**67.5**

Source: Authors' calculations on the basis of HBS 2009 and CSO Local Data Bank.

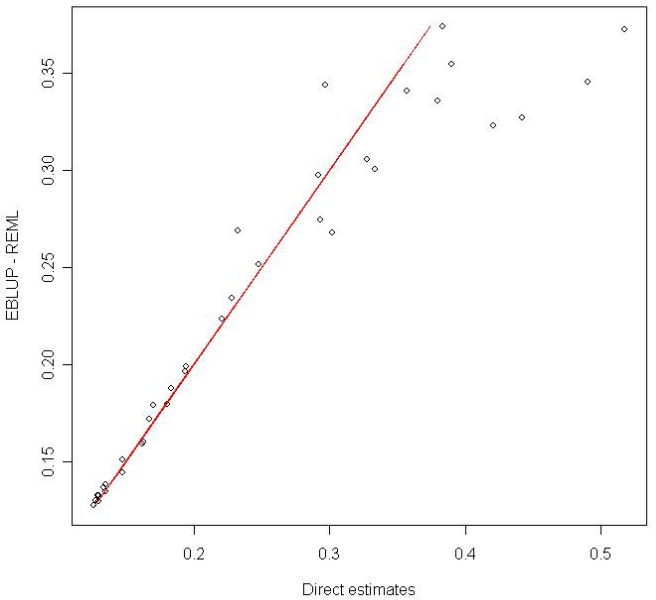


Figure 7. EBLUP (REML) estimates versus direct estimates

6. Conclusions

Efficient estimation of income distribution parameters, especially inequality and poverty characteristics and their standard errors, can be a serious problem for small domains and should be analyzed in detail. The EBLUP estimators for headcount ratios applied in the paper have proved to be more efficient than the corresponding direct estimators, as a result of "borrowing strength" from other subpopulations. The EBLUP estimation procedure, based on a general linear mixed model, has an additional advantage of taking into account the between-area variation beyond that explained by the auxiliary variables included in a classical regression model. The estimates of inequality and poverty measures by subpopulations presented in the paper can provide economists and social policy makers with valuable information that can help them to improve the decision-making process and bring them to adequate economic allocations. Understanding the domain-specific profiles may prove crucial in developing appropriate polices on most efficient reduction of the global poverty. In the future, it would also be interesting to consider more advanced cases of small area models, including poverty gap and poverty severity ratio models, that can account for differences between domains of interest more precisely.

REFERENCES

ALY, E., HERVAS, M., (1999). Nonparametric Inference for Zenga's Measure of Income Inequality, Metron, LVII (1-2), pp. 69–84.

BRUCH, CH., MÜNNICH, R., ZINS, S., (2011). Variance Estimation for Complex Surveys, AMELI project, Deliverable 3.1.

FABRIZI, E., FERRANTE, M. R., PACEI, S., TRIVISANO, C., (2011). Hierarchical Bayes Multivariate Estimation of Poverty Rates Based on Increasing Thresholds for Small Domains, Computational Statistics & Data Analysis, 55 (4), pp. 1736–1747.

FAY, R. E., HERRIOT, R. A., (1979). Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data, Journal of the American Statistical Association, 74, pp. 269–277.

JĘDRZEJCZAK, A., (2011). Metody analizy rozkładów dochodów i ich koncentracji, Łódź University Press, Łódź.

JĘDRZEJCZAK, A., KUBACKI, J., (2010). Estimation of Gini Coefficient for Regions from Polish Household Budget Survey using Small Area Estimation Methods, [in:] *Survey Sampling Methods in: Economic and Social Research*, Edited by Wywiał J., Gamrot W., Akademia Ekonomiczna w Katowicach, Katowice, pp. 109–124.

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

LANGEL, A., TILLÉ, Y., (2013). Variance Estimation of the Gini Index: Revisiting a Result Several Times Published, Journal of the Royal Statistical Society: Series A, 176 (2), pp. 521–540.

PANEK, T., (2008). Ubóstwo i nierówności: dylematy pomiaru, in: Statystyka społeczna – dokonania, szanse, perspektywy, BWS, t. 57, GUS, Warszawa, 2008, pp. 96–108.

RAO, J. N. K., (2003). Small Area Estimation, Wiley, London.

SÄRNDAL, C. E., SWENSSON, B., WRETMAN, J., (1997). Model Assisted Survey Sampling, Springer, New York.

SEN, A., (1976). Poverty - an Ordinal Approach to Measurement, Econometrica 44, pp. 219–231.

SHAO, J., CHEN, Y., CHEN, Y., (1998). Balanced Repeated Replication for Stratified Multistage Survey Data under Imputation. Journal of the American Statistical Association, 93 (442), pp. 819–831.

WOLTER, K., (2003). Introduction to Variance Estimation, Springer-Verlag, New York.

ZEHNA, P. W., (1966). Invariance of Maximum Likelihood Estimation, Annals of Mathematical Statistics 37, pp. 744–745.

ZENGA, M., (1984). Proposta per un Indice di Concentrazione Basato sui Rapporti tra Quantili di Popolazione e Quantili di Reddito. Giornale degli Economisti e Annali di Economia, 48, pp. 301–326.

ZENGA, M., (1990). Concentration Curves and Concentration Indices Derived from Them, in: Income and Wealth Distribution, Inequality and Poverty, Springer-Verlag, Berlin, 94–110.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 379–390

IMPROVED ESTIMATORS FOR SIMPLE RANDOM SAMPLING AND STRATIFIED RANDOM SAMPLING UNDER SECOND ORDER OF APPROXIMATION

Prayas Sharma[1], **Rajesh Singh**[2]

ABSTRACT

Singh and Solanki (2012) and Koyuncu (2012) proposed estimators for estimating population mean $\overline{Y}$. Up to the first order of approximation and under optimum conditions, the minimum mean squared error of both the estimators is equal to the MSE of the regression estimator. In this paper, we have tried to find out the second order biases and mean square errors of these estimators using information on auxiliary variable based on simple random sampling. Finally, we have compared the performance of these estimators with some numerical illustration.

Key words: simple random sampling, stratified random sampling, population mean, study variable, exponential ratio type estimators, bias and MSE.

1. Introduction

Suppose n pairs (x_i, y_i) (i=1,2,…,n) observations are taken on n units sampled from N population units using simple random sampling without replacement scheme. For estimating the population mean $\overline{Y}$ of a study variable Y, let us consider X be the auxiliary variable that is correlated with the study variable Y, taking the corresponding values of the units. In sampling theory the use of suitable auxiliary information results in considerable reduction in MSE of the ratio estimators. Many authors suggested estimators using some known population parameters of an auxiliary variable. Many authors including Upadhyaya and Singh (1999), Kadilar and Cingi (2006), Khoshnevisan et al. (2007), Singh et al. (2007), Singh and Kumar (2011) suggested estimators in simple random sampling using auxiliary variables. Most of the authors discussed the properties of estimators along with their first order bias and MSE. Hossain et al. (2006) studied some estimators in second order of approximation. In this study

[1] Department of Statistics, Banaras Hindu University. Varanasi-221005, India.
E-mail: prayassharma02@gmail.com.
[2] Department of Statistics, Banaras Hindu University. Varanasi-221005, India.
E-mail: rsinghstat@gmail.com.

we have studied the properties of some estimators under second order of approximation.

2. Some estimators in simple random sampling

For estimating the population mean $\overline{Y}$ of Y, Singh and Solanki (2012) proposed a ratio type estimator t_1 as

$$t_1 = \overline{y}\left(\frac{a\overline{X}+bc}{a\overline{x}+bc}\right)^{\alpha} \tag{2.1}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, a and c are either real numbers or function of known parameters of the auxiliary variable, and b is an integer which takes values +1 and -1 for designing the estimators such that $a\overline{X}+bc$ and $a\overline{x}+bc$ are non-negative. The scalar α takes values -1, (for product-type estimator) and +1 (for ratio-type estimator).

Koyuncu (2012) proposed an estimator t_2 as

$$t_2 = \overline{y}\exp\left[\frac{d\left(\overline{X}-\overline{x}\right)}{\overline{X}\left(\overline{X}+\overline{x}\right)+2e}\right] \tag{2.2}$$

where d and e is either real number or a function of the known parameter associated with auxiliary information.

3. Notations used

Let us define, $e_0 = \frac{\overline{y}-\overline{Y}}{\overline{Y}}$ and $e_1 = \frac{\overline{x}-\overline{X}}{\overline{X}}$, such that $E(e_0) = E(e_1) = 0$.

For obtaining the bias and MSE, the following lemmas will be used:

Lemma 3.1

(i) $$V(e_0) = E\{(e_0)^2\} = \frac{N-n}{N-1}\frac{1}{n}C_{02} = L_1C_{02}$$

(ii) $$V(e_1) = E\{(e_1)^2\} = \frac{N-n}{N-1}\frac{1}{n}C_{20} = L_1C_{20}$$

(iii) $$COV(e_0,e_1) = E\{(e_0e_1)\} = \frac{N-n}{N-1}\frac{1}{n}C_{11} = L_1C_{11}$$

Lemma 3.2

(iv) $$E\{(e_1^{\,2}e_0)\} = \frac{(N-n)}{(N-1)}\frac{(N-2n)}{(N-2)}\frac{1}{n^2}C_{21} = L_2C_{21}$$

(v) $$E\{(e_1^{\,3})\} = \frac{(N-n)}{(N-1)}\frac{(N-2n)}{(N-2)}\frac{1}{n^2}C_{30} = L_2C_{30}$$

Lemma 3.3

(vi) $$E(e_0e_1^3) = L_3C_{31} + 3L_4C_{20}C_{11}$$

(vii) $$E\{(e_1^4)\} = \frac{(N-n))(N^2+N-6nN+6n^2)}{(N-1)(N-2)(N-3)}\frac{1}{n^3}C_{30} = L_3C_{40} + 3L_4C_{20}^2$$

(viii) $$E(e_0^2e_1^2) = L_3C_{40} + 3L_4C_{20}$$

where

$$L_3 = \frac{(N-n))(N^2+N-6nN+6n^2)}{(N-1)(N-2)(N-3)}\frac{1}{n^3},$$

$$L_4 = \frac{N(N-n))(N-n-1)(n-1)}{(N-1)(N-2)(N-3)}\frac{1}{n^3}$$

And $$Cpq = \sum_{i=1}^{N}\frac{(Xi - \overline{X})^p}{\overline{X}^p}\frac{(Y_i - \overline{Y})^q}{\overline{Y}^q}$$

The proofs of these lemmas are straightforward by using SRSWOR (see Sukhatme and Sukhatme (1970)).

4. First order biases and mean squared errors

The bias expressions of the estimators t_1 and t_2 are respectively written as

$$Bias(t_1) = \overline{Y}\left[\frac{1}{2}\alpha(\alpha+1)A^2L_1C_{20} - \alpha AL_1C_{11}\right] \tag{4.1}$$

$$Bias(t_2) = \overline{Y}\left[\frac{3}{2}B^2L_1C_{20} - BL_1C_{11}\right] \tag{4.2}$$

where $A = \frac{a\overline{X}}{a\overline{X} + bc}$ and $B = \frac{d\overline{X}}{2e + 2d\overline{X}}$.

The MSE expressions of the estimators t_1 and t_2 are respectively given by

$$MSE(t_1) = \overline{Y}^2\left[L_1C_{02} + A^2\alpha^2 L_1C_{20} - 2A\alpha L_1C_{11}\right] \tag{4.3}$$

$$MSE(t_2) = \overline{Y}^2\left[L_1C_{02} + B^2 L_1C_{20} - 2BL_1C_{11}\right] \tag{4.4}$$

Under optimum conditions $MSE(t_1) = MSE(t_2)$, and is the same as that of MSE of usual regression estimator. In search of the best estimator, we have extended our study to the second order of approximation.

5. Second order biases and mean squared errors

Expressing estimator t_1 in terms of e's (i=0,1), we get

$$t_1 = \overline{Y}(1 + e_0)(1 + Ae_1)^{-\alpha}$$

Or

$$t_1 - \overline{Y} = \overline{Y}\{e_0 - A\alpha e_1 + MA^2e_1^2 - A\alpha e_0e_1 + MA^2e_0e_1^2 - NA^3e_1^3 - NA^3e_0e_1^3 + OA^4e^4\} \tag{5.1}$$

Taking expectations and using lemmas, we get the bias of the estimator t_1 to the second order of approximation, given by

$$Bias_2(t_1) = \overline{Y}\left[MA^2L_1C_{20} - A\alpha L_1C_{11} + MA^2L_2C_{21} - NA^3L_2C_{30} - NA^3(L_3C_{31} + 3L_4C_{20}C_{11}) + OA^2(L_3C_{40} + 3L_4C_{20}^2)\right] \tag{5.2}$$

Similarly, we get the bias of the estimator $t_2$to the second order of approximation as

$$Bias_2(t_2) = \overline{Y}\left[\frac{3}{2}A^2L_1C_{20} - AL_1C_{11} + \frac{3}{2}A^2L_2C_{21} - \frac{7}{6}A^3L_2C_{30} - \frac{7}{6}A^3(L_3C_{31} + 3L_4C_{20}C_{11}) + \frac{25}{24}A^4(L_3C_{40} + 3L_4C_{20}^2)\right] \tag{5.3}$$

where, $M = \frac{\alpha(\alpha+1)}{2}$ $N = \frac{\alpha(\alpha+1)(\alpha+2)}{6}$ $O = \frac{\alpha(\alpha+1)(\alpha+2)(\alpha+3)}{24}$.

Squaring equation (5.1) and then taking expectations and using lemmas we get MSE of t_1 up to the second order of approximation as

$$\begin{aligned} MSE_2(t_1) = \overline{Y}^2 \Big[& L_1C_{02} + A^2\alpha^2 L_1C_{20} - 2A\alpha L_1C_{11} - 2M\alpha A^3 L_2C_{30} \\ & - 2A\alpha L_2C_{12} + A^2(\alpha^2 + 2M)\left(L_3C_{22} + 3L_4(C_{20}C_{02} + C_{11}^2)\right) \\ & + 2(M + \alpha^2)A^2 L_2C_{21} + \left(M^2 + 2\alpha N\right)(L_3C_{40} + 3L_4C_{20}^2) \\ & + (2N - 4M\alpha)A^3\left(L_3C_{31} + 3L_4C_{20}C_{11}\right)\Big] \end{aligned} \tag{5.4}$$

Similarly, the MSE of the estimator t_2 up to the second order of approximation is given by

$$\begin{aligned} MSE_2(t_2) = \overline{Y}^2 \Big[& L_1C_{02} + B^2 L_1C_{20} - 2BL_1C_{11} - 3B^3 L_2C_{30} + 5B^2 L_2C_{21} - 2BL_2C_{12} \\ & - \frac{25}{3}B^3\left(L_3C_{31} + 3L_4C_{20}C_{11}\right) + 4B^2\left(L_3C_{22} + 3L_4(C_{20}C_{02} + C_{11}^2)\right) \\ & + \frac{55}{12}B^4(L_3C_{40} + 3L_4C_{20}^2\Big] \end{aligned} \tag{5.5}$$

6. Numerical illustration

For natural population data, we calculate the bias and the mean square error of the estimators and compares biases and MSE of the estimators under the first and second order of approximation.

Data Set

The data for the empirical analysis are taken from 1981, Utter Pradesh District Census Handbook, Aligar. The population consists of 340 villages under koil police station, with Y=Number of agricultural workers in 1981 and X=Area of the villages (in acre) in 1981. The following values are obtained:

$\overline{Y} = 73.76765,\ \overline{X} = 2419.04,\ N = 340,$ n=70, C_{02}=0.7614, C_{11}=0.2667, C_{12}=0.0747,

C_{03}=2.6942, C_{12}=0.1589, C_{30}=0.7877,C_{13}=0.1321, C_{31}=0.8851, C_{04}=17.4275, C_{40}=1.3051

C_{22}=0.8424,

Table 6.1. Bias and MSE of estimators

Estimators	**Bias**		**MSE**	
	First order	**Second order**	**First order**	**Second order**
$\mathbf{t_1}$	0.214866	0.214695	**47.142303**	**48.69108**
$\mathbf{t_2}$	0.4297319	0.428945	**47.142303**	**47.86158**

In the Table 6.1 the bias and MSE of the estimators t_1 and t_2 are written under the first order and the second order of approximations. From the Table 6.1 it is observed that the biases of the estimators t_1 and t_2 decrease and the mean squared errors increase for the second order of approximation. MSEs up to the first order of approximation on their optimum values are equal for both the estimators, which prompt us to the study of the estimators up to the second order of approximation, and on the basis of the study up to the second order of approximation we conclude that the estimator t_2 is better than t_1 for the given data set.

7. Estimators under stratified random sampling

While planning surveys, stratified random sampling has often proved useful in improving the precision of unstratified sampling strategies to estimate the finite population mean of the study variable $\overline{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hi}$. Assume that the population U consists of L strata as $U=U_1, U_2,\ldots,U_L$. Here, the size of the stratum U_h is N_h, and the size of the simple random sample in stratum U_h is n_h, where h=1, 2,---,L. In this study, we consider our proposed estimators from section (2) to estimate $\overline{Y}$ under stratified random sampling without replacement scheme, given respectively by

$$t_1' = \bar{y}_{st}\left(\frac{a\overline{X} + bc}{a\bar{x}_{st} + bc}\right)^{\alpha} \tag{7.1}$$

$$t_2' = \bar{y}_{st} \exp\left[\frac{d\left(\bar{X} - \bar{x}_{st}\right)}{X\left(\bar{X} + \bar{x}_{st}\right) + 2e}\right] \tag{7.2}$$

where $\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h$, $\bar{x}_{st} = \sum_{h=1}^{L} w_h \bar{x}_h$,

and $\bar{X} = \sum_{h=1}^{L} w_h \bar{X}_h$.

Here, $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$, and $\bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$.

Notations used under stratified random sampling

Let us define $e_0 = \frac{\bar{y}_{st} - \bar{Y}}{\bar{Y}}$ and $e_1 = \frac{\bar{x}_{st} - \bar{X}}{\bar{X}}$, then $E(e_0) = E(e_1) = 0$.

To obtain the bias and MSE of the proposed estimators, we use the following notations in the rest of the article:

$$V_{rs} = \sum_{h=1}^{L} W_h^{r+s} \frac{1}{\bar{Y}^r \bar{X}^s} E\left[\left(\bar{y}_h - \bar{Y}_h\right)^r \left(\bar{x}_h - \bar{X}_h\right)^s\right]$$

where $\bar{y}_h$ and $\bar{Y}_h$ are the sample and population means of the study variable in the stratum h, respectively. Similar expressions for X and Z can also be defined.

We also have,

$$E(e_0^2) = \frac{\sum_{h=1}^{L} w_h^2 \gamma_h S_{yh}^2}{\bar{Y}^2} = V_{20}$$

$$E(e_1^2) = \frac{\sum_{h=1}^{L} w_h^2 \gamma_h S_{xyh}^2}{\bar{X}^2} = V_{02}$$

$$E(e_0 e_1) = \frac{\sum_{h=1}^{L} w_h^2 \gamma_h S_{xyh}^2}{\bar{X}\bar{Y}} = V_{11}$$

where $S_{yh}^2 = \dfrac{\sum_{i=1}^{N_h} (\bar{y}_h - \bar{Y}_h)^2}{N_h - 1}$, $S_{xh}^2 = \dfrac{\sum_{i=1}^{N_h} (\bar{x}_h - \bar{X}_h)^2}{N_h - 1}$,

$$S_{xyh} = \frac{\sum_{i=1}^{N_h} (\bar{x}_{h-} \bar{X}_h)(\bar{y}_h - \bar{Y}_h)}{N_h - 1},$$

$$\gamma_h = \frac{1 - f_h}{n_h},\ f_h = \frac{n_h}{N_h},\ w_h = \frac{N_h}{N}.$$

Some additional notations used for the second order of approximation are

$$C_{rs(h)} = \frac{1}{N_h} \sum_{i=1}^{N_h} \left[\left(\bar{y}_h - \bar{Y}_h \right)^s \left(\bar{x}_h - \bar{X}_h \right)^r \right],$$

$$V_{12} = \sum_{h=1}^{L} W_h^3 \frac{k_{1(h)} C_{12(h)}}{\bar{Y} \bar{X}^2},$$

$$V_{21} = \sum_{h=1}^{L} W_h^3 \frac{k_{1(h)} C_{21(h)}}{\bar{Y}^2 \bar{X}},$$

$$V_{30} = \sum_{h=1}^{L} W_h^3 \frac{k_{1(h)} C_{30(h)}}{\bar{Y}^3},$$

$$V_{03} = \sum_{h=1}^{L} W_h^3 \frac{k_{1(h)} C_{03(h)}}{\bar{X}^3},$$

$$V_{13} = \sum_{h=1}^{L} W_h^4 \frac{k_{2(h)} C_{13(h)} + 3k_{3(h)} C_{01(h)} C_{02(h)}}{\bar{Y} \bar{X}^3},$$

$$V_{04} = \sum_{h=1}^{L} W_h^4 \frac{k_{2(h)} C_{04(h)} + 3k_{3(h)} C_{02(h)}^2}{\bar{X}^4},$$

$$V_{22} = \sum_{h=1}^{L} W_h^4 \frac{k_{2(h)} C_{22(h)} + k_{3(h)} \left(C_{01(h)} C_{02(h)} + 2C_{11(h)}^2 \right)}{\bar{Y}^2 \bar{X}^2},$$

where $k_{1(h)} = \dfrac{(N_h - n_h)(N_h - 2n_h)}{n^2(N_h - 1)(N_h - 2)}$,

$$k_{2(h)} = \frac{(N_h - n_h)(N_h + 1)N_h - 6n_h(N_h - n_h)}{n^3(N_h - 1)(N_h - 2)(N_h - 3)},$$

$$k_{3(h)} = \frac{(N_h - n_h)N_h(N_h - n_h - 1)(n_h - 1)}{n^3(N_h - 1)(N_h - 2)(N_h - 3)}.$$

8. First order biases and mean squared errors under stratified random sampling

The bias of the estimators t_1' and t_2' under stratified random sampling is respectively written as

$$\text{Bias}(t_1') = \overline{Y}\left[\frac{1}{2}\alpha(\alpha+1)A^2V_{02} - \alpha AV_{11}\right] \tag{8.1}$$

$$\text{Bias}(t_2') = \overline{Y}\left[\frac{3}{2}B^2V_{02} - BV_{11}\right] \tag{8.2}$$

The MSE of the estimators t_1' and t_2' under stratified random sampling is given by

$$\text{MSE}(t_1') = \overline{Y}^2\left[V_{20} + A^2\alpha^2V_{02} - 2A\alpha V_{11}\right] \tag{8.3}$$

$$\text{MSE}(t_2') = \overline{Y}^2\left[V_{20} + B^2V_{02} - 2BV_{11}\right] \tag{8.4}$$

Note that the optimum value of A (for $\alpha = 1$) and B is obtained as $A_{opt} = B_{opt} = \dfrac{V_{11}}{V_{02}}$. Using these optimal values, the minimum MSEs of the estimators t_1' and t_2' are given by

$$\text{MSE}_{\min}(t_1') = \left(V_{20} - \frac{V_{11}^2}{V_{02}^2}\right) = \sum_{h=1}^{L} W_h^2\gamma_h S_{yh}^2\left(1-\rho^2\right) = \text{MSE}(RE_s) \tag{8.5}$$

assuming $\alpha = 1$.

Similarly, for the estimator t_2', the minimum MSE is given by

$$\mathrm{MSE}_{\min}(t_2') = \left(V_{20} - \frac{V_{11}^2}{V_{02}^2} \right) = \sum_{h=1}^{L} W_h^2 \gamma_h S_{yh}^2 \left(1-\rho^2\right) = \mathrm{MSE}(\mathrm{RE}_s) \qquad (8.6)$$

The minimum MSE of the estimators t_1' and t_2' is equal to the MSEof the regression estimator under stratified random sampling. To find the most efficient estimator among $t_1'$and t_2', it is useful to find their MSE equations up to the second order of approximation.

9. Second order biases and mean squared errors under stratified random sampling

Expressing the estimator t_1 in terms of e's (i=0,1), we get

$$t_1' = \overline{Y}(1+e_0)(1+Ae_1)^{-\alpha}$$

or

$$t_1' - \overline{Y} = \overline{Y}\{e_0 - A\alpha e_1 + MA^2 e_1^2 - A\alpha e_0 e_1 + MA^2 e_0 e_1^2 - NA^3 e_1^{\,3} - NA^3 e_0 e_1^3 + OA^4 e^4\} \qquad (9.1)$$

Taking expectations and using lemmas we get the bias of the estimator t_1'up to the second order of approximation given by

$$\mathrm{Bias}_2(t_1') = \overline{Y}\left[MA^2 V_{02} - A\alpha V_{11} + MA^2 V_{12} - NA^3 V_{03} - NA^3 V_{13} + OA^4 V_{04})\right] \qquad (9.2)$$

Similarly, we get the bias of the estimator t_2' up to the second order of approximation as

$$\mathrm{Bias}_2(t_2') = \overline{Y}\left[\frac{3}{2}A^2 V_{02} - AV_{11} + \frac{3}{2}A^2 V_{12} - \frac{7}{6}A^3 V_{03} - \frac{7}{6}A^3 V_{13} + \frac{25}{24}A^4 V_{04})\right] \qquad (9.3)$$

Squaring equation (9.1) and then taking expectations and using lemmas we get the MSE of t_1' up to the second order of approximation as

$$\mathrm{MSE}_2(t_1') = \overline{Y}^2\left[V_{20} + A^2\alpha^2 V_{02} - 2A\alpha V_{11} - 2M\alpha A^3 V_{03} + 2(M+\alpha^2)A^2 V_{12} - 2A\alpha V_{21} + A^2(\alpha^2 + 2M)V_{22} + (2N - 4M\alpha)A^3 V_{13} + \left(M^2 + 2\alpha N\right)V_{04}\right] \qquad (9.4)$$

Similarly, the MSE of the estimator t_2' is given by

$$MSE_2(t_2') = \overline{Y}^2\left[V_{20} + B^2V_{02} - 2BV_{11} - 3B^3V_{03} + 5B^2V_{12} - 2BV_{21} - \frac{25}{3}B^3V_{13} + 4B^2V_{22} + \frac{55}{12}B^4V_{04}\right] \quad (9.5)$$

10. Numerical illustration

For the natural population data, we calculate the bias and the mean square error of the estimators for the first and second order of approximations.

Data Set

To illustrate the performance of the above estimators, we have considered the natural data given in Singh and Chaudhary (1986, p.162). The data were collected in a pilot survey for estimating the extent of cultivation and production of fresh fruits in three districts of Uttar-Pradesh in the years 1976-1977.

$\overline{Y} = 443.53$, $\overline{X} = 8.8$, V_{02}=0.062801, V_{20}=0.050034, V_{11}=0.054013,

$V_{13(2)}$=000236

$V_{03(2)}$=0.000741, $V_{11(2)}$=0.000602, $V_{02(2)}$=2.7784E-07, $V_{20(2)}$=0.000554,

$V_{04(2)}$=0.000277, $V_{12(2)}$=0.000624, $V_{21(2)}$=0.000524, $V_{22(2)}$=0.000204.

Table 10.1. Biases and MSEs of the estimators

Estimators	**Bias**		**MSE**	
	First order	**Second order**	**First Order**	**Second order**
t_1'	-3.07761E-15	-0.00391163	**704.0452321**	**897.926434**
t_2'	10.30211607	10.35740036	**704.0452321**	**711.5662075**

From Table 10.1 we observe that the MSEs of the estimators t_1' and t_2' are the same up to the first order of approximation but the biases are different. The MSE of the estimator t_2' is less than t_1' under the second order of approximation. Thus, on the basis of the second order of approximation we conclude that the estimator t_2' is better than the estimator t_1' for this data set.

11. Conclusion

In this study we have considered two estimators motivated by Singh and Solanki (2012) and Koyuncu (2012). The MSEs of these estimators are the same up to the first order of approximation. We have extended the study to the second order of approximation to search for best estimator in the case of the minimum variance. The properties of the estimators are studied under simple random sampling without replacement and stratified random sampling. We have observed from Table 6.1 and Table 10.1 that the behavior of the estimators changes dramatically when we consider the terms up to the second order of approximation.

REFERENCES

HOSSAIN, M. I., RAHMAN, M. I., TAREQ, M., (2006). Second order biases and mean squared errors of some estimators using auxiliary variable. SSRN.

KADILAR, C., CINGI, H., (2006). Ratio estimators in stratified random sampling. Biom. Jour., 45, 2, 218–225.

KHOSHNEVISAN, M., SINGH, R., CHAUHAN, P., SAWAN, N., SMARANDACHE, F., (2007). A general family of estimators for estimating population mean using known value of some population parameter(s), Far East Journal of Theoretical Statistics 22, 181–191.

KOYUNCU, N., (2012). Efficient estimators of population mean using auxiliary attributes. Appl. Math. Comput.

SINGH, D., CHUDHARY, F. S., (1986). Theory and analysis of sample survey designs. Wiley Eastern Limited, New Delhi.

SINGH, H. P., SOLANKI, R. S., (2012). Improved estimation of population mean in simple random sampling using information on auxiliary attribute. Appl. Math. Comput., 218, 7798–7812.

SINGH, R., CAUHAN, P., SAWAN, N., SMARANDACHE, F., (2007). Auxiliary Information and A Priori Values in Construction of Improved Estimators. Renaissance High Press.

SINGH, R., KUMAR, M., (2011). A note on transformations on auxiliary variable in survey sampling. MASA, 6:1, 17–19.

SUKHATME, P. V., SUKHATME, B. V., (1970). Sampling theory of surveys with applications. Iowa State University Press, Ames, U.S.A.

UPADHYAYA, L. N., SINGH, H. P., (1999). Use of transformed auxiliary variable in estimating the finite population mean. Biom. Jour., 41, 627–636.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 391–398

A RATIO-CUM-PRODUCT ESTIMATOR OF FINITE POPULATION MEAN IN SYSTEMATIC SAMPLING

Rajesh Tailor, Narendra K. Jatwa, Housila P. Singh[1]

ABSTRACT

In this paper we consider the problem of estimation of population mean using information on two auxiliary variables in systematic sampling. We have extended Singh (1967) estimator for estimation of population mean in systematic sampling. We have derived the expressions for the bias and mean squared error of the suggested estimator up to the first degree of approximation. We have compared the suggested estimator with existing estimators and obtained the conditions under which the suggested estimator is more efficient. An empirical study has been carried out to demonstrate the performance of the suggested estimator.

Key words: systematic sampling, ratio-cum-product estimator, bias, mean squared error.

1. Introduction

The use of auxiliary information for the estimation of population parameters in simple random sampling and stratified random sampling has been widely used. But in systematic sampling which is frequently used when up-to-date sampling frame is available the use of auxiliary information in estimation of population parameters has not been discussed much.

Cochran (1940) used auxiliary information at estimation stage and suggested a ratio estimator. The ratio estimator is more efficient when study and auxiliary variates are positively correlated and regression line passes through the origin. In case of negative correlation, Robson (1957) developed product method of estimation that provides a product estimator which is more efficient than the simple mean estimator. Hansen et al. (1946) developed a combined ratio estimator in stratified sampling. In systematic sampling, Swain (1964) defined a ratio estimator whereas Shukla (1971) suggested a product estimator. Many authors including Kushwaha and Singh (1989), Singh and Singh (1998) and Singh and Solanki (2012) discussed various estimators of population mean. Singh et al. (2011) suggested a general family of estimators for estimating population mean in

[1] S. S. in Statistics, Vikram University, Ujjain-456010, M. P., India.

systematic sampling using auxiliary information in the presence of missing observations. Singh and Jatwa (2012) suggested a class of exponential-type estimators in systematic sampling. Singh et al. (2011) studied some modified ratio and product estimators for population mean in systematic sampling.

Singh (1967) used information on population means of two auxiliary variates and developed a ratio-cum-product estimator in simple random sampling. Motivated by Singh (1967), we have suggested a ratio-cum-product estimator in systematic sampling.

Suppose N units in the population are numbered from 1 to N in some order. Select a sample of n units, if a unit at random is taken from the first k units and every k^{th} subsequent unit, then $N = nk$. This sampling method is similar to that of selecting a cluster at random out of k clusters (each cluster containing n units), made such that i^{th} Cluster contains serially numbered units $i, i+k, i+2k, ..., i+(n-1)k$. After sampling of n units, observe both the study variate y and the auxiliary variate x. Let y_{ij} and x_{ij} denote the observations regarding the variate y and the variate x respectively on the unit bearing the serial number $i+(j-1)k$ in the population $(i = 1, 2, ..., k; j = 1, 2, ..., n)$. If the i^{th} sampling unit is taken at random from the first k units, then $\bar{y}_{sy}$ and $\bar{x}_{sy}$ are defined as

$$\bar{y}_{sy} = \bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}, \quad \bar{x}_{sy} = \bar{x}_{i.} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}.$$

The usual ratio estimator for estimating the population mean $\bar{Y}$ in systematic sampling given by Swain (1964) is defined as

$$\hat{\bar{Y}}_R^{sys} = \bar{y}_{sys}\left(\frac{\bar{X}}{\bar{x}_{sys}}\right), \tag{1.1}$$

where $\bar{x}_{sys} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$ is an unbiased estimator of population mean $\bar{X} = \frac{1}{N}\sum_{j=1}^{N} x_{ij}$, the population mean of the auxiliary variate x. Here, $\bar{X}$ is assumed to be known.

Singh et al. (2011) suggested a ratio type exponential estimator for estimating the population mean $\bar{Y}$ in systematic sampling as

$$\bar{y}_{Resy} = \bar{y}_{sy}\exp\left(\frac{\bar{X} - \bar{x}_{sy}}{\bar{X} + \bar{x}_{sy}}\right). \tag{1.2}$$

When the study variate and auxiliary variate are negatively correlated, Shukla (1971) suggested a product estimator for population mean $\overline{Y}$ as

$$\hat{\overline{Y}}_P^{sys} = \overline{y}_{sys}\left(\frac{\overline{z}_{sys}}{\overline{Z}}\right). \tag{1.3}$$

Singh et al. (2011) suggested a product type exponential estimator for estimating the population mean $\overline{Y}$ in systematic sampling as

$$\overline{y}_{Pesy} = \overline{y}_{sy}\exp\left(\frac{\overline{x}_{sy}-\overline{X}}{\overline{x}_{sy}+\overline{X}}\right). \tag{1.4}$$

Variances of the ratio estimators $\hat{\overline{Y}}_R^{sys}$ and $\overline{y}_{Resy}$ up to the first degree of approximation are respectively given by

$$V(\hat{\overline{Y}}_R^{sys}) = \left(\frac{N-1}{nN}\right)\overline{Y}^2\left[\rho_y^* c_y^2 + \rho_x^* c_x^2\left(1-2k\sqrt{\rho^{**}}\right)\right] \tag{1.5}$$

and

$$V\left(\overline{y}_{Resy}\right) = \left(\frac{N-1}{nN}\right)\overline{Y}^2\left[\rho_y^* c_y^2 + \rho_x^*\left(c_x^2/4\right)\left(1-4k\sqrt{\rho^{**}}\right)\right], \tag{1.6}$$

where

$k = \rho_{yx}c_y/c_x$, $\rho_y^* = \{1+\rho_y(n-1)\}$, $\rho_x^* = \{1+\rho_x(n-1)\}$, $\rho^{**} = \{\rho_y^*/\rho_x^*\}$,

$$S_{yx} = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{X})(y_{ij}-\overline{Y}),\quad \rho_{yx} = \frac{S_{yx}}{S_y S_x},$$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{X})^2, \qquad S_y^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\overline{Y})^2,$$

$$S_z^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(z_{ij}-\overline{Z})^2,$$

and (ρ_y, ρ_x, ρ_z) being the intra-class correlation between the units of a cluster corresponding the (y,x,z) variates.

Further, the variances of the product estimator $\hat{\overline{Y}}_P^{sys}$ and $\overline{y}_{Resy}$ up to the first degree of approximation are respectively given by

$$V(\hat{\overline{Y}}_P^{sys}) = \left(\frac{N-1}{nN}\right)\overline{Y}^2\left[\rho_y^* c_y^2 + \rho_z^* c_z^2\left(1+2k^*\sqrt{\rho_2^{**}}\right)\right] \tag{1.7}$$

and

$$Var\left(\overline{y}_{Pesy}\right) = \left(\frac{N-1}{nN}\right)\overline{Y}^2\left[\rho_y^* c_y^2 + \rho_z^*\left(c_z^2/4\right)\left(1+4k^*\sqrt{\rho_2^{**}}\right)\right], \tag{1.8}$$

where

$$k^* = \rho_{yz}\{c_y / c_z\} \text{ and } \quad \rho_2^{**} = \{\rho_y^* / \rho_z^*\}$$

In the situation when the study variate y is positively correlated with the auxiliary variable x and negatively correlated with another auxiliary variable z we have suggested a ratio-cum-product estimator of the population mean $\overline{Y}$ in line with Singh (1967) in systematic sampling. We have derived the bias and mean squared error of the suggested estimator up to the first degree of approximation. Conditions are obtained under which the suggested estimator is better than the usual unbiased estimator, Swain (1964) ratio estimator, Shukla's (1971) product estimator and Singh et al. (2011) estimators. The result of this paper has been supported through an empirical study.

2. Suggested estimator

Motivated by Singh (1967) we have suggested a ratio-cum-product estimator in systematic sampling for population mean $\overline{Y}$ as

$$\hat{\overline{Y}}_{RP}^{sys} = \overline{y}_{sys}\left(\frac{\overline{X}}{\overline{x}_{sys}}\right)\left(\frac{\overline{z}_{sys}}{\overline{Z}}\right) . \tag{2.1}$$

To obtain the bias and the mean squared error of the suggested estimator $\hat{\overline{Y}}_{RP}^{sys}$, we write

$\overline{y}_{sys} = \overline{Y}(1+e_0)$, $\overline{x}_{sys} = \overline{X}(1+e_1)$, $\overline{z}_{sys} = \overline{Z}(1+e_2)$ such that

$E(e_0) = E(e_1) = E(e_2) = 0$

and $\quad E(e_0^{\,2}) = \theta c_y^2 \rho_y^*$, $E(e_1^{\,2}) = \theta c_x^2 \rho_x^*$, $E(e_2^{\,2}) = \theta c_z^2 \rho_z^*$,

$$E(e_0 e_1) = \theta k c_x^2 \sqrt{\rho_y^* \rho_x^*} \ , \ E(e_0 e_2) = \theta k^* c_z^2 \sqrt{\rho_y^* \rho_z^*} \ ,$$

$$E(e_1 e_2) = \theta k^{**} c_z^2 \sqrt{\rho_x^* \rho_z^*} .$$

where

$$c_z^2 = S_z^2 / \overline{Z}^2, \ \rho_z^* = \{1 + \rho_z (n-1)\}, \ k^{**} = \rho_{xz}\left(c_x / c_z\right),$$

$$S_z^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(z_{ij} - \overline{Z})^2, \ \rho_{yz} = S_{yz}/(S_y S_z), \ \rho_{xz} = S_{xz}/(S_x S_z),$$

$$S_{xz} = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \overline{X})(z_{ij} - \overline{Z}), \ S_{yz} = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y})(z_{ij} - \overline{Z}) \text{ and}$$

$$\theta = \left(\frac{N-1}{nN}\right).$$

Expressing the suggested estimator $\hat{\bar{Y}}_{RP}^{sys}$ in terms of e's we have

$$\hat{\bar{Y}}_{RP}^{sys} = \bar{Y}(1+e_0)(1+e_1)^{-1}(1+e_2). \tag{2.2}$$

We assume that $|e_1| < 1$ so that $(1+ e_1)^{-1}$ is expandable. Expanding that right-hand side, multiplying out and neglecting terms of e's having power greater than two we have

$$\hat{\bar{Y}}_{RP}^{sys} = \bar{Y}\left[1+e_0-e_1+e_2-e_oe_1+e_oe_2-e_1e_2+e_1^2\right]$$

or

$$\left(\hat{\bar{Y}}_{RP}^{sys}-\bar{Y}\right)=\bar{Y}\left[e_0+e_1+e_2-e_oe_1+e_oe_2-e_1e_2+e_1^2\right]. \tag{2.3}$$

Taking expectation of both sides of (2.3) we get the bias of the estimator $\hat{\bar{Y}}_{RP}^{sys}$ to the first degree of approximation as

$$B(\hat{\bar{Y}}_{RP}^{sys})=\left(\frac{N-1}{nN}\right)\bar{Y}\left[\rho_x^*c_x^2\left(1-k\sqrt{\rho^{**}}\right)+\sqrt{\rho_z^*}c_z^2\left(k^*\sqrt{\rho_y^*}-k^{**}\sqrt{\rho_x^*}\right)\right]. \tag{2.4}$$

Squaring both sides of (2.3) and neglecting terms of e's having power greater than two we have

$$\left(\hat{\bar{Y}}_{RP}^{sys}-\bar{Y}\right)^2=\left[e_0^2-e_1^2+e_2^2-2e_oe_1+2e_oe_2-2e_1e_2\right]$$

Taking expectation of both sides of the above equation, we get the mean square error of $\hat{\bar{Y}}_{RP}^{sys}$ to the first degree of approximation as

$$MSE(\hat{\bar{Y}}_{RP}^{sys})=\left(\frac{N-1}{nN}\right)\bar{Y}^2\left[\rho_y^*c_y^2+\rho_x^*c_x^2\left(1-2k\sqrt{\rho^{**}}\right)+\sqrt{\rho_z^*}c_z^2\left(1-2k^{**}\sqrt{\rho_1^{**}}\right)+2k^*c_z^2\sqrt{\rho_y^*\rho_z^*}\right], \tag{2.5}$$

where $\rho_1^{**}=\left(\rho_x^*/\rho_z^*\right)$.

3. Efficiency comparisons

In systematic sampling, the variance of the usual unbiased estimator $\bar{y}_{sys}$ is given by

$$V(\bar{y}_{sys})=\left(\frac{N-1}{nN}\right)\bar{Y}^2\rho_y^*c_y^2 \tag{3.1}$$

From (2.5), (1.5), (1.6), (1.7) and (1.8) it is observed that the suggested estimator $\hat{\bar{Y}}_{RP}^{sys}$ would be more efficient than

(i) $V(\bar{y}_{sys})$ if

$$\left[\rho_x^* c_x^2 + \left(1 - 2k\sqrt{\rho^{**}}\right) + \sqrt{\rho_z^*} c_z^2 \left(1 - 2k^{**}\sqrt{\rho_1^{**}}\right) + 2k^*\sqrt{\rho^{**}}\right] < 0 \tag{3.2}$$

(ii) $\hat{\bar{Y}}_R^{sys}$ if

$$d > 1/2 \tag{3.3}$$

(iii) $\hat{\bar{Y}}_P^{sys}$ if

$$d_1 < 1/2 \tag{3.4}$$

(iv) $\bar{y}_{Resy}$ if

$$\left[\rho_x^* c_x^2 \left(\frac{3}{4} - k\sqrt{\rho^{**}}\right) + \rho_z^* c_z^2 \left(1 - 2k^{**}\sqrt{\rho_1^{**}}\right) + 2k^*\sqrt{\rho_2^{**}}\right] < 0 \tag{3.5}$$

(v) $\bar{y}_{Pesy}$ if

$$\left[\rho_z^* c_z^2 \left(\frac{3}{4} + k^*\sqrt{\rho_2^{**}}\right) + \rho_x^* c_x^2 \left(1 - 2k\sqrt{\rho^{**}}\right) + 2k_1^*\sqrt{\rho_3^*}\right] < 0 \tag{3.6}$$

where $d = \left(k^{**}\sqrt{\rho_1^{**}} - k^*\sqrt{\rho_2^{**}}\right), d_1 = \left(k\sqrt{\rho^{**}} + k_1^*\sqrt{\rho_3^*}\right), \rho_1^{**} = (\rho_x^*/\rho_z^*)$, $\rho_2^{**} = (\rho_y^*/\rho_z^*), \rho_3^{**} = (\rho_z^*/\rho_x^*)$ and $k_1^* = \rho_{xz}(c_z/c_x)$.

Expressions (3.2), (3.3), (3.4), (3.5) and (3.6) provide the conditions under which the suggested estimator $\hat{\bar{Y}}_{RP}^{sys}$ would be more efficient than usual unbiased estimator $\bar{y}_{sys}$, ratio estimator $\hat{\bar{Y}}_R^{sys}$, product estimator $\hat{\bar{Y}}_P^{sys}$, ratio type exponential estimator $\bar{y}_{Resy}$ and product type exponential estimator $\bar{y}_{Pesy}$.

4. Empirical study

To illustrate the performance of the proposed ratio-cum-product estimator $\hat{\bar{Y}}_{RP}^{sys}$ relative to the usual unbiased estimator, we assumed (artificially) the following values of parameters:

$\bar{X} = 44.47$, $S_x^2 = 149.55$, $c_x = 0.28$, $S_{xy} = 538.57$, $\bar{Y} = 80$, $S_y^2 = 2000$, $c_y = 0.56$, $S_{yz} = -902.86$, $\bar{Z} = 48.40$, $S_z^2 = 427.83$, $c_z = 0.43$, $S_{xz} = -241.06$, $\rho_{xy} = 0.9848$ $\rho_{yz} = -0.9760$, $\rho_{zx} = -0.9530$, $\rho_x = 0.707$, $\rho_y = 0.6652$, $\rho_z = 0.5487$, N=15, n=3.

Table 4.1. Percent Relative Efficiency of $\bar{y}_{sys}$, $\hat{\bar{Y}}_{R}^{sys}$, $\hat{\bar{Y}}_{P}^{sys}$ and $\hat{\bar{Y}}_{RP}^{sys}$ with respect to $\bar{y}_{sys}$.

Estimator	PRE $(., \bar{y}_{sys})$
$\bar{y}_{sys}$	100.00
$\hat{\bar{Y}}_{R}^{sys}$	389.620
$\hat{\bar{Y}}_{p}^{sys}$	189.452
$\bar{y}_{Resy}$	177.434
$\bar{y}_{Pesy}$	139.318
$\hat{\bar{Y}}_{RP}^{sys}$	777.790

5. Conclusions

Section 3 provides the conditions under which the suggested estimator has less mean squared errors in comparison to the simple mean estimator, the ratio estimator and the product estimator in systematic sampling. Table 4.1 exhibits that the suggested estimator $\hat{\bar{Y}}_{RP}^{sys}$ has the largest percent relative efficiency as compared to the simple mean estimator $\bar{y}_{sys}$, the ratio estimator $\hat{\bar{Y}}_{R}^{sys}$ and the product estimator $\hat{\bar{Y}}_{p}^{sys}$, the ratio type exponential estimator $\bar{y}_{Resy}$ and the product type exponential estimator $\bar{y}_{Pesy}$. Thus, the suggested estimator $\hat{\bar{Y}}_{RP}^{sys}$ is recommended for its use in practice.

Acknowledgements

The authors are thankful to the Editor and to the learned referee for his valuable suggestions regarding improvement of the paper.

REFERENCES

COCHRAN, W. G., (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. J. Agri. Sci., 30, 262–275.

HANSEN, M. H., HURWITZ, W. N., GURNEY, M., (1946). Problems and methods of the sample survey of business. J. Amer. Statist. Assoc., 41, 173–189.

KUSHWAHA, K. S., SINGH H. P., (1989). Class of almost unbiased ratio and product estimators in systematic sampling. J. Ind. Soc. Agri. Statist, 41, 2, 193–205.

ROBSON, D. S., (1957). Application of multivariable polykays to the theory of unbiased ratio type estimation. J. Amer. Statist. Assoc., 50, 1225–1226.

SHUKLA, N. D., (1971). Systematic sampling and product method of estimation. In *Proceedings of all India Seminar on Demography and Statistics*. B.H.U. Varanasi: India.

SINGH, H. P., SINGH R., (1998). Almost unbiased ratio and product type estimators in systematic sampling. Questiio, 22, 3, 403–416.

SINGH, H. P., SOLANKI, R. S., (2012). An efficient class of estimators for the population mean using auxiliary information in systematic sampling. Journal of Statistical Theory and Practice, 6, (2), 274–285.

SINGH, R., MALIK, S., CHAUDHARY, M. K., VERMA, H., ADEWARA, A. A., (2012). A general family of ratio type estimators in systematic sampling. J. Reli. & Stat. Studs., 5, 1, 73–82.

SINGH, H. P., JATWA, N. K., (2012). A class of exponential type estimators in systematic sampling. Eco. Qulty. Control., 27, 195–208.

SINGH, H. P., TAILOR, R., JATWA, N. K., (2011). Modified ratio and product estimators for population mean in systematic sampling. J. Modern. Appl. Statist. Meth. 10 , 2, 424–435.

SWAIN, A. K. P. C., (1964). The use of systematic sampling ratio estimate. J. Ind. Statist. Assoc., 2, 160–164.

SINGH, M. P., (1967). Ratio-cum-product method of estimation. Metrika, 12, 34–42.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 399–428

PROCESSES IN TRANSBORDER AREAS – SIGNIFICANT IMPACT ON THE ECONOMIC GROWTH

Marek Cierpiał-Wolan[1]

ABSTRACT

Dynamics and interdependence of socio-economic phenomena in the contemporary world require the appropriate instruments to be developed to monitor them. This becomes especially evident in the case of transborder areas, which are differentiated in terms of the socio-economic potentials, on the one hand, along with a growing gap in the data system for transborder areas, on the other. Therefore, it seems to be essential (especially in countries in transition) to create a coherent research system that collects, processes and disseminates information for such areas. Recognizing these needs, the Polish official statistics has initiated works towards establishing a suitable infrastructure. The results of this research (based on household surveys, border traffic surveys, entrepreneurship surveys, etc.) have shown, rather unexpectedly, that estimates of some items in Balance of Payment (BoP) are being changed. Consequently, they should be taken into account in the calculation of Gross Domestic Product (GDP), as discussed in this article.

Key words: research system for transborder areas, improvement in GDP estimates.

1.Introduction

Analyses of the processes of socio-economic development in the modern world clearly indicate the need to study phenomena occurring in transborder areas. Integration and disintegration processes between countries and regions result in higher interaction within socio-economic issues (through the difference of potential) including an increase in the scale of unregistered and illegal transactions. It is worth stressing that unregistered trade usually means all goods which are not registered on customs documents but can be legally transferred across the border. It does not cover grey or black zones. All these circumstances cause the specific behaviour of both households and businesses.

[1] Statistical Office, Rzeszow, Poland. E-mail: m.wolan@stat.gov.pl.

Therefore, the unique character of transborder areas requires a great number of various surveys of socio-economic matters to be carried out. Establishing a consistent research system should include a wide spectrum of methodological structure, which will be useful both in the countries covered and not covered by liberalization of the rules on requirements as to crossing the border (it will be particularly helpful in the countries with both kinds of border crossings, e.g. internal and, at the same time, external borders of the European Union). Effective functioning of such a system requires support from standardized sources of information (official registers, administrative sources of data, bank registers, automatic measurement of traffic, other Big data sources), as well as creation of projects which will not only include surveys on borders, but will primarily concentrate on processes ongoing around the border.

The economic crisis that emerged in the world economy in 2008 showed how important it is to have detailed information on economic activity. It is worth noting that an important part of the structure of GDP is the trade balance and the balance of spending part of foreigners and residents of a country.

Despite the fact that humankind in the last year produced more information than ever, paradoxically, we can observe growing information gaps and failures in the functioning of the information systems concerning transborder areas. Thus, it is important to work on continuous improvement methodology of research system for transborder areas.

In order to meet the demand for information, both research community and Polish official statistics undertake various actions concerning the use of different sources of information, monitoring socio-economic phenomena in transborder areas, and above all, improving and designing new surveys for these areas.

Within the framework of these projects, the integrated surveys for the needs of tourism statistics, national accounts and balance of payments has been launched by Polish official statistics. The results of the pilot survey turned out to be unexpected, and influenced the development of a coherent research system on transborder areas.

The objective of this paper is to develop a survey methodology for transborder areas in order to use the results of this survey on the micro-meso-macroeconomic level. When it comes to macroeconomic level, incomplete information causes a disturbance of GDP structure from expenditure approach and failures in economic growth rate.

2.Towards coherent research system for transborder areas

The process of creation of a coherent research system for transborder areas consists in three main interlocking parts: delimitation of transborder areas, monitoring of socio-economic phenomena and data sources and comprehensive survey. Obviously, they require harmonised methodology, applicable for

countries covered and not covered by liberalization of the rules of crossing the border. Taxonomic methods and spatial models are especially helpful in this kind of research system. Firstly, we ought to work out of a uniform set of variables for transborder areas in order to apply both kinds of methods. We know that the choice of variables is always a little subjective. Therefore, we shall engage to this project a team of experts from different backgrounds (business, self-government, government institution, scientists, etc.) in order to choose appropriate set of measures. A synthetic indicator can be given as a result of the first method, for example the transborder index. AHP method which is a part of spatial models is used for evaluating competitiveness of spatial units and tracking demand and supply shocks both in time and space. It also allows for detecting whether there is diffusion, exchange, interaction - whether other regions become infected or whether local changes are a response to exogenous shocks (Cierpiał-Wolan, 2011), (Cierpiał-Wolan, Wierzbiński, 2013).

Delimitation of the transborder area is the first step that should be taken, namely with preliminary delimitation based on, for example, regulations (according to which the border zone covers an area of 30 to 50 km from the border), etc. We must not forget about dynamic delimitation which means systematic analysis of socio-economic phenomena in this area and environment.

All sorts of information which we can find in statistical databases and administrative registers adjusted to transborder areas fall within the ambit of the monitoring. In this case inventory of information resources is often fruitful. It may happen that we have unknowingly a lot of information concerning transborder areas. Sometimes only a deep insight into statistical databases or a little modification in statistical forms is required so as to adapt a survey to our needs. The next interesting path to follow is combining information from the registers and sample surveys especially in terms of budget constraints. Using non-statistical sources of information, Big Data in particular, should not be omitted. Therefore, monitoring involves continuous reporting on socio-economic phenomena in transborder areas using various kinds of sources so as to obtain profiles of a borderland.

Comprehensive surveys are the third pillar of the coherent research system for transborder areas. Based on previous and current achievements, four kinds of surveys should be conducted: questionnaire survey at the border and in the vicinity of the border, household survey, survey of unregistered economy, survey of travelling foreigners in tourist accommodation establishments. According to this plan of actions Polish official statistics launched a pilot survey which consists of two modules. The first one comprises surveys at the EU's external and internal borders (on the territory of Poland) including Questionnaire survey at the borders and in the vicinity of the border as well as Traffic intensity survey only at the EU's internal border. Households survey constitutes the second module**.**

3. Border traffic and movement of goods and services at the European Union's external border

After Poland's accession to the Schengen zone, the part of eastern Polish border became at the same time a part of the external border of the European Union. Moreover, the growing gap in potentials between Polish and Belarusian as well as Ukrainian economies caused an increase in the scale of non-registered and illegal trade. Therefore, due to integration process and higher intensity of economic phenomena a greater demand for information concerning transborder areas appeared. In response to these, Polish official statistics resumed in 2008 the survey of goods and services turnover in border traffic. Initially, it was carried out on the Polish-Ukrainian border, and since 2010 at the whole European Union's external border on the territory of Poland.

The survey of goods and services turnover in border traffic is conducted using the representative method, which allows generalization, with specified error, of the obtained results for a total number of persons who cross the surveyed border, by foreigners and the Poles. It is conducted on a sample of about 1%.

A two-stage scheme for drawing elements for a sample with determining the strata is used. First, the days (time intervals) undergoing survey are drawn, then persons are drawn out of those who cross the border. The strata are determined according to the days of the week as well as the border crossings and kind of traffic. For each of the strata one, selected at random, 12-hour interval in a quarter coinciding with a day shift of the border guard is surveyed. Drawing of a sample is the same for the Poles and foreigners. For each of the selected shift (a unit of the first stage of drawing which participate in the survey) a sample of persons undergoing survey is selected by means of systematic sampling. In case a selected person rejects to participate in the survey, a successive person is surveyed.

The questionnaire survey is carried out in quarter periods, in selected days of a week chosen from the total number of days in a given period (7 times in the quarter). Non-representative days are not included in drawing, e.g. national and religious holidays. The questionnaire survey is conducted simultaneously at all border crossings covered by the survey.

Estimation of survey results is based on data gathered from questionnaires and information of the Border Guard on border traffic which concern respective crossings, including the way of crossing the border. These data cover the number of the Poles and foreigners who cross the border according to the crossing, direction and kind of traffic (the way of crossing the border) in a surveyed quarter and in 12-hour shifts during which the questionnaire surveys were carried out.

Data are generalized separately for the Poles and foreigners in each stratum. Results for the border sections and regions (NUTS 2 level) are calculated on the basis of the results from all strata.

The results of the this survey analysis allow us to say that the phenomena associated with the border traffic are of great importance for social-economic development of transborder areas. The greatest intensity of these phenomena occurred in areas in the strip of up to 50 kilometres along the border. It is reflected by high percentage of people crossing the border who incurred expenses in this strip, as well as the fact that the inhabitants of villages located in this area were among the vast majority of people crossing the border, and the majority of expenses was incurred in this strip. Analysis of the results of the survey carried out at the European Union's external border on the territory of Poland also shows some variation occurring on its particular sections, especially in the case of the Polish-Russian border.

The value of foreigners' expenses in Poland and Poles abroad was significant in comparison with the Polish foreign trade turnover. Expenses incurred for the purchase of goods in Poland by foreigners declaring Ukraine as a country of residence amounted to 4.4 billion zl in 2013, whereas expenses of Poles returning from Ukraine – 224.7 million zl. Data on exchange of goods presented by statistics of foreign trade for Poland showed that exports to Ukraine amounted to 18.0 billion zl in 2013, whereas imports from Ukraine – as a country of dispatch – amounted to 7.0 billion zl.

The value of expenses incurred by the Belarusians in Poland is also significant. In 2013, the expenses on the purchase of goods incurred in Poland by foreigners declaring Belarus as the country of residence amounted to 2.4 billion zl and expenses of Poles in Belarus – 76.3 million zl. Exports of goods from Polish to Belarus amounted to 7.7 billion zl, whereas imports from Belarus (as a country of dispatch) –2.5 billion zl.

Expenses on the purchase of goods incurred in Poland by foreigners declaring Russia as a country of permanent residence amounted to 855.7 million zl, whereas expenses of Poles declaring Russia as a country of stay – 370.2 million zl. Exports of goods from Poland to Russia amounted to 34.1 billion zl, whereas imports from Russia – as a country of dispatch – totalled to 78.5 billion zl. Expenses on goods incurred by the Poles in Russia and Russians in Poland were much lower compared to expenses on purchase of goods incurred by Belarusians and Ukrainians in Poland. The Polish exports to Russia and imports from Russia was, however, much higher than in the case of Ukraine and Belarus. Therefore, the expenses on the purchase of goods incurred in Poland by foreigners declaring Russia as a country of permanent residence were small compared with exports from Poland to Russia.

Expenses incurred in Poland by foreigners crossing the border in territorial units on NUTS 2 level accounted for around 18% compared with the value of export sales of entities of these units in 2012. As to expenses incurred abroad by

Poles crossing the border in corresponding units, they accounted for 2.4% compared with the value of import purchases of entities from these units in 2012.

It is also worth noting that the value of purchases of goods made in Poland by foreigners crossing the European Union's external border in 2013 at current prices was 11.5% compared with the value of retail sales of the regions located by the European Union's external border.

The results of the survey of goods and services turnover in border traffic at the European Union's external border on the territory of Poland showed that the vast majority of people go abroad and return in one day, mainly to make purchases. The expenses structure was dominated by the resources allocated to the purchase of goods, and only a small part was spent on services.

There were differences between studied phenomena on individual sections of the border. In particular, the Polish-Russian border could be seen as different one, both with regard to the purposes for which foreigners visited Poland, as well as to distance to travel, structure of expenses and the frequency of crossing the border.

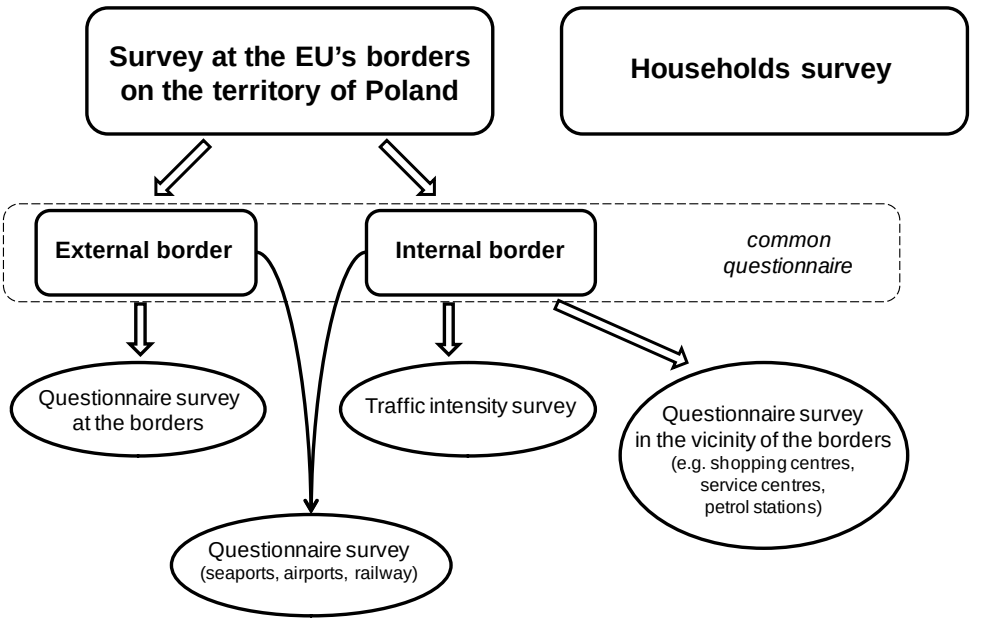
Changes in legislation relating to crossing the border, introduced by Poland or neighbouring country, were almost immediately reflected in changes in the level of border traffic intensity and border trade. The functioning of local border traffic, which has been present at the Polish-Ukrainian border since July 2009, has had a positive impact on increasing traffic and the amount of expenses made by foreigners in the Polish border area. Thus, it has positively influenced the revival of the area, as evidenced by the higher rate of increase in the number of economic entities in this area than in areas by other sections of the border. This can, in turn, be beneficial for the development of units on the LAU level 2 that are in the border area.

The volume of the border traffic and the presence of border trade are affected by different endo- and exogenous factors, especially relationships between prices of goods on both sides of the border, supply-side deficiencies in the internal market of a given country and the possibility of crossing the border. The way they evolve will determinate the development of border trade, which is one of the forms of border cooperation, conducive to its development as well as strengthening and improving the living conditions of the inhabitants of border regions.

4. Comprehensive survey

In order to obtain information from both the survey of foreigners' trips to Poland and Pole's participation in trips, a pilot survey "The integration of statistical surveys of travel tourism for the needs of tourism statistics, national accounts and balance of payments" has been launched.

The survey consists of two parts - the survey at the border and in its vicinity and the household survey of participation of Poles in travel.



4.1. Typology of border crossings

In order to select border crossings for observation their typology was created. In the first stage the crossings were divided by type of the border, with the following three types of crossings: land border – road, rail and river crossings; sea border – ports; air border – airports.

The land crossings were divided according to the nature (permeability) of the borders and the neighbouring country, that is the borders located at the EU's external border on the territory of Poland (which includes Polish-Russian, Polish-Belarusian and Polish-Ukrainian borders), and internal border of the EU and Poland (which includes Polish-Lithuanian, Polish-Slovak, Polish-Czech and Polish-German borders). Then, on the basis of the Border Guard data on border traffic, the profiles of border crossings were made.

After a thorough analysis, the border crossings which were selected were the ones whose share in the border traffic of persons at the Polish border with the neighbouring country was higher than 1%, and in the case of ports and airports – the share in border traffic at the sea and air border, respectively. The number of border crossings meeting the accepted criteria was 88, which accounted for 33.0% of all crossings. However, what needs to be stressed, those crossings were handling as much as 96.0% of all border traffic.

At the internal border these border crossings were grouped, with the use of Ward's method, according to particular sections of the border into subsets characterized with high internal similarity due to selected features. The following

features of each crossing were considered: total volume of border traffic of persons, the proportion of foreigners among persons crossing the border, the proportion of passenger cars in border traffic of vehicles, the percentage of trucks in the border traffic of vehicles.

The adoption of the first two characteristics is, as it seems, obvious, taking into account the purpose of the survey. The share of passenger cars was also analyzed because of different nature of crossings for cars and trucks. The movement of cars dominated on vast majority of border crossings. The share of trucks in border traffic of vehicles is highly correlated with the share of passenger cars (correlation coefficient close to -1), so it was omitted in the method of grouping of border crossings.

In grouping the border crossings according to the features adopted in the framework of individual sections of the border 5 subsets were obtained (with the exception of the border with Lithuania). The results of this grouping are used to select the crossings for the survey of border traffic.

In the next step, within the groups, crossings are drawn to the survey. This helps to ensure the observation of diverse traffic of people and vehicles at the internal border. When choosing crossings for the survey geographical distances between the selected crossings, which are another element of stratification, are taken into account. It is important to drawn crossings which are not centred around one location, but are spread across the border with of a given country. This is due to the varying traffic on account of the geographical location of the crossing.

4.2. The survey at the border and in its vicinity

a) The border traffic survey covers people and vehicles crossing the Polish border with the countries of the European Union at selected border crossings. It records the number of vehicles crossing the border from and to Poland and the number of people travelling in these vehicles, as well as people travelling on foot (including bicycles, wheelchairs, etc.). The survey is carried out every quarter.

The survey of border traffic at the internal border introduced the rotation of border crossings. It means that the selected crossings for a given section of the border included crossings which were surveyed in the previous year. Such an approach will ensure continuity of information on changing conditions on the crossings.

The survey of border traffic at the airports and seaports was developed separately due to its specificity and with the use of appropriate data sources (in particular reporting on air and sea transport). When selecting the crossings the possibility of obtaining relevant data on the structure of people travelling is taken into account. Due to the small share of rail crossings in border traffic it is not expected to maintain a continuous survey on these crossings. In order to discern

the specificity of trips made by foreigners crossing the border by rail, temporary surveys of this kind of travellers are conducted at selected railway stations (including border localities through which trains pass).

Taking into consideration the fact that Poland borders with four the EU countries (i.e. Germany, Czech Republic, Slovakia and Lithuania) and the traffic at the sections of the border with these countries is very diverse, the total road border traffic at the EU's internal border EU on the territory of Poland is the sum of the values for individual sections of the border. Data on border traffic of Poles are estimated on the basis of information from a survey conducted in households and from counting of vehicles and people at selected border crossings. The border traffic of foreigners leaving Poland, however, is estimated using regression analysis for each of the section of the border and on the basis of data from counting of vehicles and people, as well as other available data sources.

Using the regression analysis, equations describing border traffic on particular sections were created as follows:

$$Y_{Co} = \begin{cases} Y_{CN} = \alpha_{N_0} + \alpha_{N_1} X_{N_1} + ... + \alpha_{N_m} X_{N_m} + e_N \\ Y_{CC} = \alpha_{C_0} + \alpha_{C_1} X_{C_1} + ... + \alpha_{C_m} X_{C_m} + e_C \\ Y_{CS} = \alpha_{S_0} + \alpha_{S_1} X_{S_1} + ... + \alpha_{S_m} X_{S_m} + e_S \\ Y_{CL} = \alpha_{L_0} + \alpha_{L_1} X_{L_1} + ... + \alpha_{L_m} X_{L_m} + e_L \end{cases} \quad (1)$$

where:

Y_{Co} – (for $o = \{N, C, S, L\}$) takes the form depending on the border: with Germany Y_{CN} – the number of foreigners crossing the Polish-German border, similarly to the border with the Czech Republic (Y_{CC}), Slovakia (Y_{CS}) and Lithuania (Y_{CL}),

X_{o_m} – variables describing traffic at the selected section of the border,

α_{o_m} – coefficients of variables.

The following variables were selected to be used in the estimates:

- Y – the number of foreigners crossing the EU's internal border in Poland (data for 2005-2007 available from the Border Guard),
- X_1 – data of the Border Guard concerning traffic intensity in 2005-2007 and data from pilot surveys on traffic of people and vehicles at the EU's internal border in Poland carried out by the CSO),

- X_2 – the number of foreign tourists in collective accommodation establishments (the report on the use of collective tourist accommodation establishment - as a part of the survey conducted by the CSO).

Using the first variable X_1 seems obvious taking into account the purpose of the survey. The number of foreign tourists accommodated in collective accommodation establishments was used in the analysis due to the high degree of correlation between total number of crossings, which in turn allows for using the current data to characterize the changes in the structure of crossings. It should be emphasised that in the case of selecting new crossings for the survey in subsequent years, the model will be re-estimated again.

All regression models were analyzed in detail in terms of form, nature of relation between variables and properties of residuals. The models meet the established criteria and therefore guarantee high precision and accuracy of the estimated results.

The estimates of the number of foreigners by country are based on the analysis of registered countries according to on registration numbers of vehicles and pedestrians' declarations of country of origin counted on selected border crossings. In addition, information about country of origin are combined with the database on the use of collective accommodation establishments and the structure of people who responded to questionnaires conducted on trips made by foreigners to Poland. A synthetic summary of the information will be the basis to determine traffic intensity of people in particular countries.

b) **The questionnaire survey of tourist and same-day trips of foreigners on selected crossings on the EU's internal and external border on the territory of Poland** includes the questionnaire survey of foreigners (non-residents) leaving Poland in order to obtain travel information (including the expenses incurred in connection with travel to Poland, the purpose of visit, the length of stay in Poland) conducted in the vicinity of selected road crossings at the EU's internal and external borders on the territory of Poland and at seaports and airports.

Due to lack of survey frame, the survey of trips made by foreigners to Poland uses elusive population[1], but owing to the fact the place of the survey is known proper representativeness of the sample is provided. This is a questionnaire survey carried out in the vicinity of selected border crossings on the EU's internal and external border on the territory of Poland, including seaports and airports, in the form of direct interviews made by interviewers. Participation in the survey is voluntary. People are surveyed using systematic sampling. In case a selected person refuses to participate in the study, the next person is surveyed. For

[1] L. Kish (1991), A taxonomy of elusive populations, Journal of Official Statistics, 7, 339–347.

individual border crossings sampling intervals are determined taking into account the projected volume of travellers traffic on individual crossings and the possibility of interviewer to interview at a certain time. The survey is carried out every quarter. The survey of trips made by foreigners to Poland is a sample survey.

It is important, in developing appropriate methods for estimating the border traffic, to select a proper representative sample and obtain good quality data. The survey of this phenomenon must therefore be limited to the necessary group of crossings and certain time periods (days and hours) per year, which at the same time ensure the quality of the results.

Information on the size of the total border traffic in Poland is a total compilation of statistics which includes:

- the results of estimating the border traffic at the EU's internal border on the territory of Poland on road crossings using regression analysis,
- data of the Border Guard,
- data of the Civil Aviation Authority on passenger traffic (information on domestic and international traffic, flights directions of target cities),
- structure of passengers obtained from the questionnaire survey conducted at selected airports.

Because data on traffic at the internal border are generalized to relevant border crossings and sections of the border, one should calculate the values for the number of people crossing the border which correspond to different strata that were separated on account of border crossings. It is therefore necessary to define the weights for particular strata.

Let us assume that Y_{oi} is the number of people crossing a given section of the border o, where $o = \{N, C, S, L, M, Lot\}$ (N – the Polish-German border, S – the Polish-Slovakian border, C – the Polish-Czech border, L – the Polish-Lithuanian border, M – sea border, Lot – air border) in the i-th quarter ($i = \{1,2,3,4\}$). Let us use $N_{oi(s)}$ as the number of people crossing the border at a given section of the border at a given border crossing s ($s = 1,2,3,...$).

Let us assume, additionally, that $Q_{oi(s)}$ is the number of people crossing the border on account of the stratum s. The value $Q_{oi(s)}$ is calculated by the formula:

$$Q_{oi(s)} = Y_{oi} \cdot \frac{N_{oi(s)}}{\sum_{s} N_{oi(s)}}. \tag{2}$$

The above formula implies that $Y_{oi} = \sum_{s} Q_{oi(s)}$.

Due to specificity of individual border sections the weights used for generalization of the results are calculated on the basis of different data for

internal and external border. The basis for estimating the results on external land border is data obtained from the questionnaires and information from the Border Guard on border traffic. On the internal land border, the estimated results are based on data obtained from questionnaires and data on traffic based on counting vehicles and people which are then generalized to respective border crossings and sections of the borders. Data are generalized separately for each stratum. The results for individual sections of the borders are calculated on the basis of the results of relevant strata.

In the border survey a two-stage sampling design with separated strata was used.

Let us assume that $N_{ij(s)}$ is a real number of people crossing borders on *j*-th day of the week of *i*-th quarter in stratum s, and $n_{ij(s)}$ is the number of respondents crossing borders on *j*-th day of the week and *i*-th quarter in stratum s, $i = 1,2,3,4,\ j = 1,2,...,7$.

A generalizing weight is assigned to respondents belonging to the stratum s in *i*-th quarter and *j*-th day of the week by the following formula:

$$w_{ij(s)} = \frac{N_{ij(s)}}{n_{ij(s)}}. \tag{3}$$

Let Z denote a set of all categories of generalization.

If $G_{ij(s)}(z)$ is an unknown number of people crossing the border with the feature of the category $z \in Z$, $g_{ij(s)}(z)$ is the number of respondents with the feature of the z category, then we can note:

$$\frac{G_{ij(s)}(z)}{N_{ij(s)}} = \frac{g_{ij(s)}(z)}{n_{ij(s)}}, \tag{4}$$

and thus:

$$G_{ij(s)}(z) = \frac{N_{ij(s)}}{n_{ij(s)}} \cdot g_{ij(s)}(z) = w_{ij(s)} \cdot g_{ij(s)}(z). \tag{5}$$

The number of people crossing the border, belonging to the surveyed z category is generalized by the formula:

$$L_s(z) = \sum_i \sum_j G_{ij(s)}(z). \tag{6}$$

The total number of people $L(z)$ crossing the border, belonging to the surveyed z category, is the sum of estimates $L_s(z)$ throughout the year, i.e.:

$$L(z) = \sum_{s} L_s(z). \tag{7}$$

Generalization of travel expenses incurred in Poland for the purchase of goods and services (including accommodation services, catering and transport) is carried out in the same categories of crossing the border as the number of people crossing the border.

If $x_{ij(s)}(k,z)$ is the value of expenses incurred by *k*-th respondent, $k = 1,2,...,n_{ij(s)}$, belonging to z category, then expenses for each category of people crossing the border are generalized according to the formula:

$$T_s(z) = \sum_{i}\sum_{j} G_{ij(s)}(z) \cdot \overline{x_{ij(s)}(z)}, \tag{8}$$

where:

$\overline{x_{ij(s)}(z)}$ – denotes average value of expenses incurred in *i*-th day of *j*-th quarter by a respondent belonging to the *z* category, *w*hich is described by the relation:

$$\overline{x_{ij(s)}(z)} = \frac{\sum_{k=1}^{g_{ij(s)}(z)} x_{ij(s)}(k,z)}{g_{ij(s)}(z)} \ . \tag{9}$$

Hence, and from condition (19) it follows that:

$$T_s(z) = \sum_{i}\sum_{j} w_{ij(s)} \cdot g_{ij(s)}(z) \cdot \frac{\sum_{k=1}^{g_{ij(s)}(z)} x_{ij(s)}(k,z)}{g_{ij(s)}(z)}, \tag{10}$$

therefore:

$$T_s(z) = \sum_{i}\sum_{j}\sum_{k=1}^{g_{ij(s)}(z)} w_{ij(s)} \cdot x_{ij(s)}(k,z). \tag{11}$$

The total value of expenses $T(z)$ of people crossing the border, belonging to the z category throughout a year, is the sum of estimates $T_s(z)$, i.e.:

$$T(z) = \sum_{s} T_s(z). \tag{12}$$

Let $X_{ij(s)}(k)$ denote an amount of total expenses incurred by *k*-th respondent on *j*-th day of the week and *i*-th quarter in stratum s, $k = 1,2,...,n_{ij(s)}$.

Let:

$$W_{ij(s)} = \sum_{k=1}^{n_{ij(s)}} X_{ij(s)}(k) \tag{13}$$

be a sum of total expenses of respondents on *j*-th day of the week and *i*-th quarter in stratum s, then:

$$\frac{W_{ij(s)}}{n_{ij(s)}} \tag{14}$$

is an estimation of average amount of total expenses per respondent on *j*-th day of the week and *i*-th quarter in stratum s, whereas:

$$T_{ij(s)} = \frac{N_{ij(s)}}{n_{ij(s)}} \cdot W_{ij(s)} \tag{15}$$

is an estimation of total expenses incurred by people crossing the border on *j*-th day of the week and *i*-th quarter in stratum s.

The estimation of total amount of expenses T_s in stratum s is the sum of estimates $T_{ij(s)}$, i.e.:

$$T_s = \sum_i \sum_j T_{ij(s)} = \sum_i \sum_j \frac{N_{ij(s)}}{n_{ij(s)}} \cdot \sum_{k=1}^{n_{ij}(s)} X_{ij(s)}(k). \tag{16}$$

The estimation of total amount of expenses T throughout a year in the survey is the sum of estimates T_s, i.e.:

$$T = \sum_s T_s. \tag{17}$$

Let us assume that $N_s = \sum_i \sum_j N_{ij(s)}$ is the total real number of people crossing the border in a given year in stratum s and $n_s = \sum_i \sum_j n_{ij(s)}$ is the total number of respondents in a given year in stratum s.

If the sample is selected in a way that the weights from formula (3) do not depend on *i* and *j*, i.e. when $w_{ij(s)} = \frac{N_s}{n_s}$, then T_s is reduced to

$$T_0 = \frac{N_s}{n_s} \cdot \sum_i \sum_j \sum_{k=1}^{n_{ij(s)}} X_{ij(s)}(k), \tag{18}$$

i. e., to an average amount of total expenses $\frac{1}{n_s}\sum_i\sum_j\sum_{k=1}^{n_{ij(s)}} X_{ij(s)}(k)$, multiplied by the real number of people crossing the border N_s for a given stratum s.

The results of the pilot survey indicate that the number of arrivals of foreigners to Poland amounted to 15 779 thousand. Compared with 2007, the number of crossings made increased by more than 7%. The largest increase in traffic of foreigners leaving Poland was reported at airports (over 36%), followed by seaports (an increase of over 59%). At the eastern border the number of arrivals of foreigners, compared with the data of 2007, increased by over 15%, on the southern border - an increase of more than 5.6%, and on the western border - by almost 3%. The value of the total expenditure which were incurred by foreigners on account of the trip to Poland - after the generalization of data obtained in the pilot survey - was 6 335 825 zl. Foreigners from the EU countries spent a total of 4 038 067 zl, whereas from the non-EU countries - 2 287 758 zl.

4.3. Household survey of participation of Poles in travel

The survey of participation of Poles in travel is a sample questionnaire survey, carried out in face-to-face interview made by interviewers. Participation in the survey is voluntary. A quarter is reference period while the survey is carried out in the month following the quarter.

The sample is drawn from a frame built on the basis of census enumeration areas base (from which census enumeration areas with zero flats are excluded) with applying a two-stage sampling by means of stratification on the first stage. Census enumeration areas or a set of census enumeration areas with the minimum of 5 dwellings are the first-stage sampling units (primary sampling units – PSU). Census enumeration areas which do not fulfil this condition are combined into a unit within the same statistical division. The second-stage sampling units are dwellings. Five dwellings are drawn from each first-stage sampling unit.

Census enumeration areas are sorted by strata, which were defined using the following criteria: 1. subregion 2. variable p – as the size of a locality. Additionally, the strata were modified for large cities.

The strata containing border areas are divided into two parts: border part and central part. A border zone consists of gminas located not further than 30 kilometres from the border or the coastline. Areas in the coastal zone without access to marine connection with foreign countries are treated as the central area. If a part of gmina is situated in a distance between 30 and 50 kilometres from the border line, it is included in the border area as well. This zone has been set along Polish border based on definition of the border area and the results of the survey

of goods and services turnover in border traffic at the EU's external border in Poland.

In this way 254 strata were formed out of 191. However, due to the fact that 5 strata were so small (single gminas) that a single unit of the first degree could not be allocated to them, they were attached to strata of adjacent subregion with the same class of localities, and located in the same zone (border or non-border one). Finally, 249 strata were obtained.

The sample size (for Poland) was determined on the basis of data from National Census of Population and Housing 2011, while the basis for calculating the sample in voivodships was the number of households in gminas. The number of dwellings in a gmina is taken from the TERYT[1] database which is periodically updated.

In order to obtain sufficient number of questionnaires for same-day abroad trips, the sample is doubled in border zones as these areas see the highest number of same-day trips. Therefore, half of the sample is allocated to the border strata and the other half to the central strata. Within each of these strata the sample allocation is proportional to the number of dwellings in a stratum. The adopted method allows for generalization of the results at the voivodship level with division of a voivodship into border and central areas.

The following notation is assumed in the sampling:

w – symbol of voivodship,

h – stratum number in voivodship,

k – PSU number in stratum,

N_{wh} – number of PSUs in *h*-th stratum of *w*-th voivodship,

n_{wh} – number of PSUs in the sample in *h*-th stratum of *w*-th voivodship,

M_{wh} – number of dwellings in *h*-th stratum of *w*-th voivodship,

M_{whk} – number of dwellings in *k*-th PSU of *h*-th stratum of *w*-th voivodship.

The first-stage sampling units (PSU) are drawn separately in each stratum. In a given *h*-th stratum of *w*-th voivodship (denoted by N_{wh}, n_{wh}, M_{wh}, M_{whk} respectively as N, n, M and M_k) PSUs are sorted randomly in such a way that first each PSU is given a random number, then PSUs are sorted by increasing order of the random numbers.

In the next step a sequence of accumulated values is constructed:

$$\{S\} = \begin{cases} S_0 = 0 \\ S_k = S_{k-1} + M_k \end{cases}, \tag{19}$$

[1] National Official Register of Territorial Division of the Country.

for $k = 1,2,...,N$, hence:

$$S_N = \sum_k M_k = M, \tag{20}$$

where M – a number of dwellings in stratum.

After constructing the sequence $\{S\}$ an interval of sampling is calculated:

$$IN = \frac{M}{n}. \tag{21}$$

Moreover, a random start number P_0 is drawn from the range $(0; IN)$. The values IN and P_0 are the real numbers.

Then a numerical sequence $\{X_i\}$ is constructed:

$$X_i = P_0 + IN \cdot (i-1) \quad \text{for } i \in \{1,2,...,n\}. \tag{22}$$

If for some $i \in \{1,2,...,n\}$ the following inequality is satisfied:

$$S_{k-1} < X_i < S_k, \tag{23}$$

then k-th PSU is added to the sample.

Samplings in each strata are carried out in the same way.

Sampling of dwellings is carried out in each PSU which has been drawn to the sample. From each PSU 5 dwellings are drawn.

The following information is available for a given PSU - the address, the PSU number, the number of dwellings in PSU, i.e. M_{whk}.

Dwellings are drawn using simple random sampling without replacement, i.e. 5 integers are selected without repetition from the set $[1; M_{whk}]$.

The procedure of sampling dwellings is the same in all PSUs selected for the sample.

Selected dwellings are sorted sequentially according to: PSU, census enumeration areas (if PSU consists of 2 or more census enumeration areas), the dwelling number in a census enumeration area.

The generalization of the results of the survey include:

- probabilities of selecting households,
- level of the completeness of the survey by class of locality,
- structure of household's population by current demographic data.

Accordingly the following weights for each household are calculated:

- $w1g$ – sampling weight,
- $w2g$ – adjusted weight calculated by taking into account the level of the completeness of the survey by class of locality,
- wg – weight calculated by taking into account structure of household's population.

The weight $w1g$ for households results from the established sampling scheme. The weight is reciprocal of probability of selection of a household which surveyed household lives in. For the household which belongs to *h*-th stratum and *k*-th PSU:

$$w1g_{hk} = \frac{1}{\pi_{hk}}, \tag{24}$$

whereas

$$\pi_{hk} = \frac{n_h \cdot M_{hk}}{M_h} \cdot \frac{m}{M_{hk}} = \frac{n_h \cdot m}{M_h} = \frac{m_h}{M_h}, \tag{25}$$

where:

n_h – number of PSU to be selected from stratum *h*,

m – number of dwellings selected from one PSU,

m_h – number of dwellings to be selected from stratum *h*

M_h – number of dwellings in *h*-th stratum,

M_{hk} – number of dwellings in *k*-th PSU of *h*-th stratum.

Therefore, for all dwellings which belong do stratum *h* the following weight is assigned:

$$w1g_h = \frac{M_h}{m_h}. \tag{26}$$

The weight *wlg* is then adjusted if the interviewer was unable to contact the selected dwellings, if there is a lack of information because the respondents refused to participate in the survey, if there was temporary absence of persons of household and the like. This weight is adjusted in six classes of locality *p* separately because there is a relation between a class of locality and the level of the completeness of the survey. Therefore, the rate R_p of the completeness of the survey is calculated:

$$R_p = R1_p \cdot R2_p \tag{27}$$

where:

$R1_p$ – rate of making contact with dwellings in class p,

$R2_p$ – rate of responses in class p.

The first rate relates to dwellings and it is a quotient of the number of dwellings in which households reside and with whom contact was made by the

interviewer to the number of actually existing dwellings. Dwellings closed down, transformed into non-residential facilities or wrong addresses are not taken into account in calculating this rate.

The latter rate concerns households and indicates a fraction of households which were interviewed. Weights *wlg* are used for estimating rates $R1_p$ and $R2_p$.

Because each stratum belongs to one class of locality, the weight *wlg* for households from stratum *h* of class *p* is adjusted as follows:

$$w2g_h = \frac{w1g_h}{R_p}. \tag{28}$$

In this way adjusted weighs *w2g* on account of non-responses are obtained.

The next step is to calculate weights *wg* taking weight *w2g* as a basis. The weights are calculated using demographic data from other sources. As additional variables information about number of households is used according to six size classes, i.e. 1-person, 2-person, 3-person, 4-person, 5-person and 6 or more person in the division of urban and rural areas. The values of these variables come from NSP 2011.

The following calibration method was used. For each of 12 categories of households (1-person households, 2-person households,…, 6 or more persons households in urban and rural areas) the values are calculated:

$$M_j = \frac{G_j}{\hat{g}_j}, \tag{29}$$

where:

G_j – number of households of *j*-th category in population (i.e. according to NSP 2011),

$\hat{g}_j$ – number of households of *j*-th category estimated on the basis of the sample.

Finally, for a household which belongs to *h*-th stratum and *j*-th category:

$$wg_{hj} = w2g_h \cdot M_j. \tag{30}$$

As the number of the surveyed households is a small subset of the population of all households in Poland, the data on the number of abroad trips made by Poles were also obtained on the basis of data derived from: the results of the survey carried out at road crossings at the EU's internal border on the territory of Poland; data of the Civil Aviation Authority on the number of passengers checked-in to foreign airports, data of the Central Statistical Office reports on the volume of traffic of passengers on ships, data of the Border Guard.

The traffic on road crossings at the internal border was estimated using regression analysis. The model included the variables:

• Y - the number of Poles returning home through the EU's internal border on the territory of Poland (available data of the Border Guard for the years 2003-2007),
• X - data on the volume of traffic on selected crossings.

The outcome of the analyzes were four equations describing the movement of Poles at selected sections of the border. As in the case of estimating the movement of foreigners, the estimated structural parameters of the models are statistically significant. Moreover, these models were analyzed in terms of the relationships between variables and properties of residuals. The value of the assessment of the F test significance for all equations was much lower than the assumed level of significance of 0.05. The models meet the basic assumptions of the method of least squares, which provide the basis for their practical use.

According to the results of the pilot survey, in conjunction with the trips completed in the first quarter of 2013, Poles spent 4232.5 million zl, of which 59% were spent during foreign trips. On average, the cost of one short-term domestic trip was 228 zl, a long-term domestic trip - 695 zl, and a foreign trip - 2170 zl.

5. The specificity of transborder research

Questionnaire surveys conducted at the border are unique. This is a very important issue, which has a large impact on the organization of the survey, as well as to the limit in the selection of survey methods.

It is not easy to acquire a respondent in border surveys as the travellers who are surveyed are usually in a hurry. At the EU's internal border it is even more difficult to acquire respondents due to free movement of vehicles and persons after abolition of border controls. For that reason many vehicles do not stop in the vicinity of the border crossing but goes to points distant from the border. In the case of some crossings at the EU's internal border, particularly those characterized by a high share of local border traffic, only foreigners residing in the border area stop in the vicinity of the crossing. They stop to make purchases and it is difficult to capture persons arriving for other purposes (with at least one overnight stay). Hence, it is important to choose and constantly verify places for the survey.

Another issue concerning the conduct of the border surveys is the risk associated with the occurrence of a dangerous situation for interviewers, especially in the late hours of the day and night. Therefore, it is important to cooperate with the Police, the Border Guard and to reduce surveys at these times to a minimum.

An important problem arising in carrying out the surveys of border traffic of vehicles and people and trips made by foreigners to Poland is all kinds of difficulties in obtaining information on the number of persons in vehicles and the country from which the vehicle comes, as well as obtaining responses from foreigners leaving Poland.

5.1. "Random route" technique for collecting questionnaires

The analysis of the data obtained from border surveys conducted so far by the official statistics shows that border traffic is generated by a small group of vehicles, which implies that the probability of selecting a household whose members travel abroad is small. In other words, the population of people who travel abroad is a small subset of the population of all households in the border area. The survey of trips made by Poles introduced a modified technique of collecting questionnaires, maintaining at the same time the principles of a representative selection of households for the survey. In case the interviewer fails to make an interview during the first visit in the selected flat, he/she is obliged to retry the contact. If, despite retried attempts, a household cannot be contacted or if the selected household was taking part in travelling, the "random route" technique is applied to collect questionnaires. According to this technique, when it is impossible to conduct the interview the interviewer goes to the next apartment to make an interview in accordance with the appropriate algorithm for the selection of subsequent flats, visiting a maximum of 8 apartments. The number of maximum searches has been determined on the basis of the number of vehicles crossing the border and the number of households in border powiats. In the case of households not taking part in travelling, the interviewer writes down the relevant characteristics of the visited household and moves to another flat. If the interviewer finds the flat which is taking part in travelling, he determines the number of households taking part in travelling, but if it was one household, he conduct an interview with it. In case of more than one household he shall draws just one of them according to the following principles - the household whose householder was the last to celebrate his birthday is selected for the survey. Then, he moves to the next starting point. Conversely, when a household which is taking part in travelling is found, the interviewer conduct a survey in the last (eighth) step and proceeds to the next starting point.

5.2. Calibration of weights due to generalized results of estimations of border traffic of Poles

In households surveys of trips usually modifications of parameters are required. In this case the calibration of weights was made using estimated data on trips of Poles and individual data obtained from the household survey.

Based on the data obtained from the pilot survey, three categories of households were separated on account of the length of stay of Poles abroad:

- households with only same-day trips,
- households with only one or more overnight stays (multi-day trips),

- households with same-day and multi-day trips.

For each of them the number of completed trips was assigned. In the next stage, the estimation of the number of trips taking into account the foreign trips of Poles was made. The calculations consisted of:

- for households that participated in the trips with just one or more overnight stays the total number of trips was estimated depending on the number of trips from a given household along with the use of the structure of weights assigned to these households,

- for households that had both foreign same-day and multi-day trips the estimation method was similar, i.e. the total number of trips was estimated depending on the number of trips from a given household, along with the use of the structure of weights assigned to these households,

- for households that had only same-day trips a different procedure was assigned. In this case, the households were divided according to the country visited. These households were assigned the number of trips that should be executed in them to a particular country, then the total number of trips was estimated depending on the number of trips from a given household, along with the use of the structure of weights assigned to these households.

In this way new weights for households that participated in foreign trips were calculated. For the remaining households the weights were reduced proportionately.

5.3. Method for estimating for all countries of the world the trips made by foreigners to Poland and their expenses

In order to estimate the results of the trips made by foreigners (non-residents) to Poland and their expenses the following data are used additionally: data on the use of tourist accommodation establishments, data of the Border Guard on border crossings made by foreigners broken down by country of origin, and data on crossings of the EU's internal border on the territory of Poland based on information obtained in the survey of border traffic and from airports and seaports. The sources of data listed above contain information about the trips made by foreigners to Poland from approximately 193 countries around the world.

In the first step all countries of the world were divided into 19 categories due to different specificities of average expenditure, the type and length of stay, purpose of visit, the distance from Poland, etc. Among these categories the countries bordering Poland were separated individually. Other European countries were divided into 4 groups (Eastern Europe, Southern Europe, Western Europe, and Northern Europe), Africa - into two groups (North Africa and South Africa), Asia - into two groups (Middle East Asia and Far East Asia), America divided

into 3 groups (North America, Central America and South America) and Australia and Oceania. In some cases the calculations are performed on the combined categories due to the specific topic (e.g. the calculation of the average expenditure for same-day visitors).

In the next step a comparison of the number of trips for each country on the basis of the border guard is made. These values are then adjusted on the basis of the report which contains information on the number of foreigners using collective accommodation establishments.

The next step was to calculate the number of overnight stays of tourists from different countries. The principle was to calculate the average length of stay (number of nights per single trip) for each group of countries. The analysis of the data showed that in several categories the number of nights per trip significantly differs from those in other categories. This was due to the small number of registered questionnaires in a given category, or extreme cases registered (few tourists staying in Poland for a long time).

Total expenditure for the country is the product of the average expenditure for the category in which a given country is located and the estimated number of trips for this country.

5.4. Additional sources of data

The method of generalizing the results uses additional data sources apart from data obtained directly from the count at the border and conducted questionnaire survey. These sources are used as follows:

• Border Guard data are used when estimating the overall traffic at the EU's internal border on the territory of Poland. The relevant time series of the volume of total crossings, as well as the number of vehicles crossing the border are used in the econometric model which is prepared. In addition, these data are used in the process of estimating the number of trips on account of the country of residence of foreigners visiting Poland,

• data on the use of tourist accommodation establishments - the data used in the estimation of the overall traffic in relevant sections of the border. The corresponding time series are included in the econometric models as additional explanatory variables. Furthermore, the information from accommodation establishments are taken into account in estimating the number of trips made by foreigners to Poland as well as in estimating the number of overnight stays,

• data on air transport - data of the Civil Aviation Authority and the Central Statistical Office reports. These data are used to estimate the number of trips made by foreigners and Poles travelling by air,

• data on sea transport - data of the CSO reporting; they are used to estimate the number of trips made by foreigners and Poles travelling by sea,

• data on railway transport - data used to estimate the number of trips made by foreigners and Poles travelling by rail,

• data of travel agencies - data collected on the basis of a specially developed questionnaire. The information contained in the questionnaire are used to estimate expenses incurred for the purchase of travel packages with tour operators,

• EUROSTAT data for non-typical countries (countries of low tourist traffic with Poland).

6. Results

a) Impact on the economic growth

The developed research system proves that there is a greater economic activity and entrepreneurship in border areas. Obviously, this has an impact on the condition of companies and macroeconomic indicators.

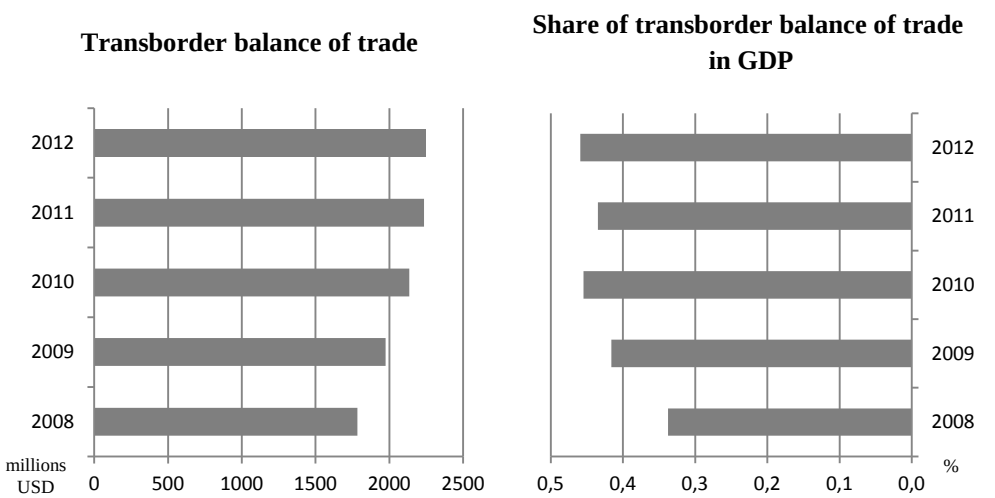


Figure 1. Transborder balance of trade and share of transborder balance of trade in GDP

The bar chart on the right shows the share of the balance of trade in tourism for transborder areas in GDP - transborder balance of trade (expenses of foreigners in Poland minus expenses of Polish citizens during foreign travel). What is important is the fact that an increasing tendency can be seen in the analyzed period with the exception of 2011. It is also worth stressing that this share does not exceed 0.5 %.

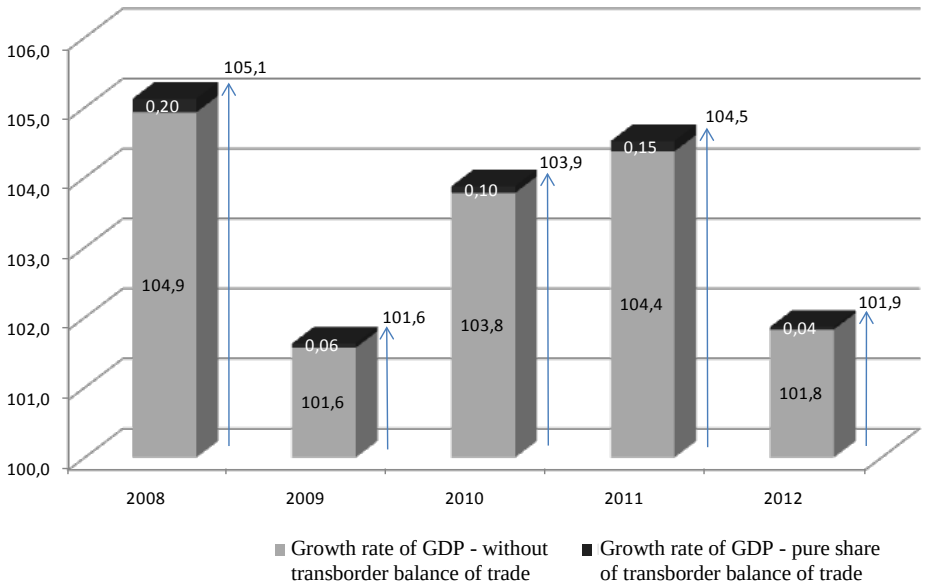


Figure 2. The rate of economic growth

In order to calculate how the balance of foreigners and Poles' expenditure influenced GDP growth, appropriate calculations of expenditure values into constant prices of the previous year and the fixed prices of the base year were made, using appropriate price indices for this purpose. In this way, the GDP growth rate was calculated without the participation of tourism expenditure. The resulting dynamics was compared with the officially published data on economic growth.

The grey part of the bars shows the rate of economic growth if we would not take into consideration the transborder balance of trade in tourism. The black parts illustrate pure share of this balance in economic growth.

We noticed systematic increase in transborder balance of trade. However, the nature o transborder processes is the reason why this increase cannot last indefinitely. So, as you can see, this red shares can be slightly higher or lower when comparing one period to another, but they still sustain the process of the economic growth in the analyzed period.

a) Entrepreneurship in border areas – selected aspects

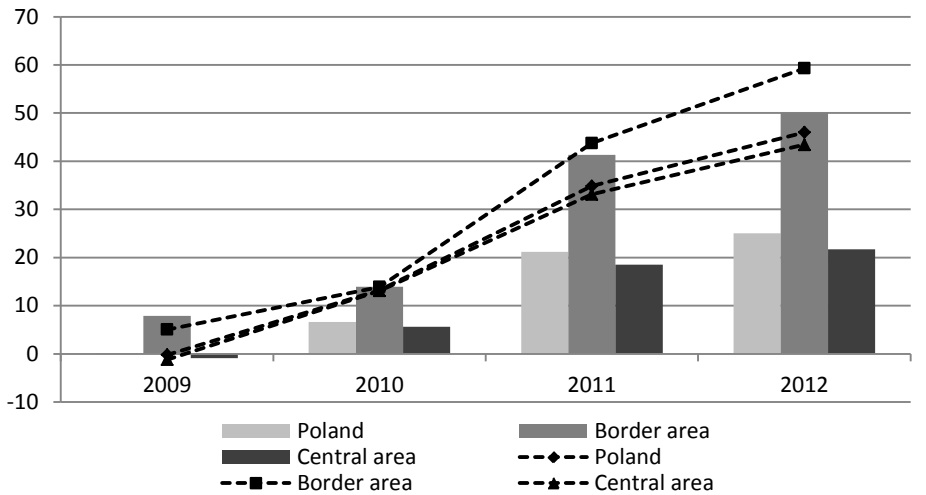


Figure 3. Dynamics of revenues from sale of products, goods and materials as well as dynamics of export

The dynamics of both revenues and export show a clear upward trend. For both these indicators the largest increase occurred in 2011. Border area throughout the whole period is characterized by higher dynamics than the national average.

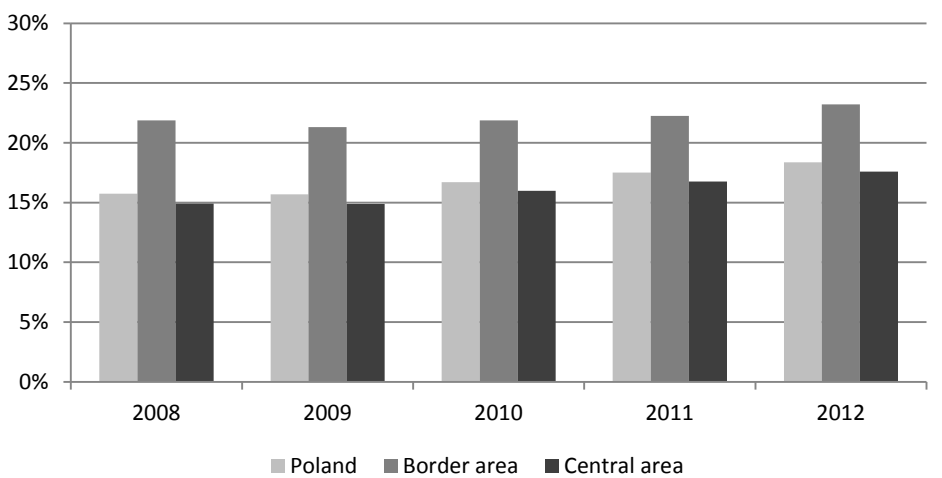


Figure 4. The share of revenues from exports of products, goods and materials in revenues from sales of products

The share of revenues from exports of products, goods and materials in revenues from sales of products, goods and materials is the evidence for competitiveness of companies in the market.

The level of the rate for the border area was growing steadily from 2010 up to 23% in 2012. It is worth noting that since 2010 the difference in this ratio between border area and Poland remained at a similar level (about 5 percentage points).

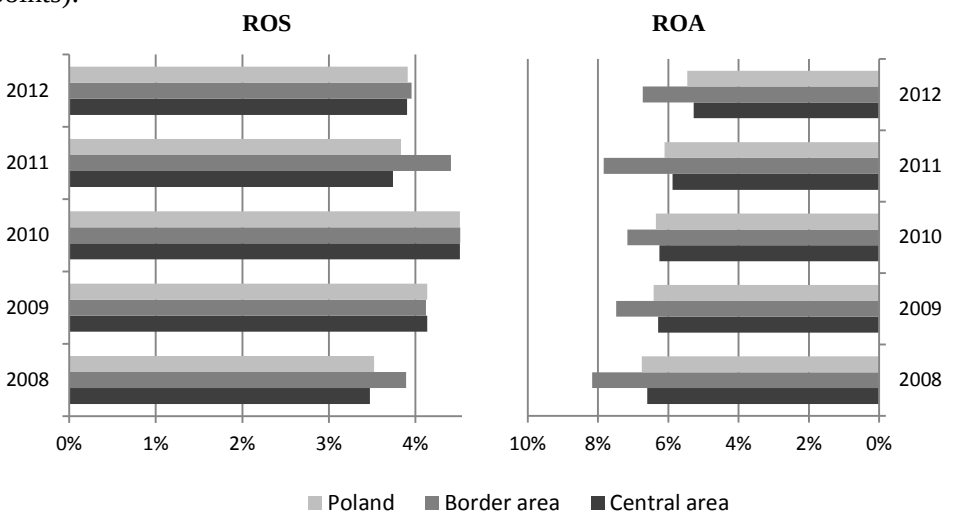


Figure 5. Return on sales rate (ROS) and return on assets rate (ROA)

The rate of return on sales, calculated as a share of net profit in revenues from total activity, illustrates the level of profit margins used by the company. High margins contribute to increase profitability but can also adversely affect the company's competitiveness in the market. By 2010, in the border area a steady increase in the rate of return on sales was recorded. In 2011 and 2012, the rate was lower than in the two previous years, but still remained above the index calculated for the country.

Return on assets, calculated as the ratio of operating profit (EBIT) to total assets, reflects the financial result subject to loads due to financial costs (interest and taxes).

Both in the country and in central area the rate decreases in the reporting period, while in the border area there was an increase in 2011 compared to the years 2009-2010. Throughout the entire period under analysis one can clearly observe a higher value of this index for the border zone. In 2012, in all areas the index recorded a decrease to the lowest level in the last five years.

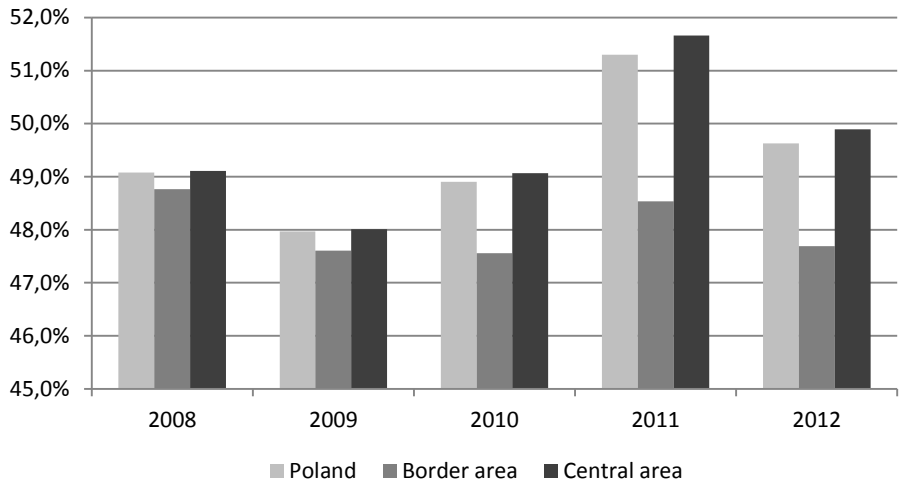


Figure 6. The debt ratio

The debt ratio is calculated as the ratio of liabilities to assets, informs about the amount of debt per unit of assets. Even high value of the index at a satisfactory profitability and liquidity preserved need not be the reason for a negative assessment. Thus, this indicator should be analyzed in the context of other financial indicators.

In 2008-2010, the debt ratio for the border area was declining from year to year. In 2011 it recorded an increase (insignificant one compared with the index for Poland), and in 2012 returned to the level of 2010. It is worth emphasizing that throughout the analyzed period this rate is lower both compared with the rate for Poland and the central area.

On the basis of the total debt ratio and the return on assets rate one can determine the rate of return on equity. The combination of these three indicators allows one to evaluate the effect of leverage. It is worth noting that the lower level of overall debt in the border area did not influence significantly the deterioration of financial leverage, because the disparities in the values of ROE and ROA between the border zone and the whole Poland are similar.

7. Conclusions

In terms of the economic slowdown that we see in many parts of the world, precise estimates of individual components of GDP are becoming particularly important. The experience of several European countries shows that the phenomena occurring in the border areas have a significant impact on the balance of payments. Nowadays, economic growth is determined by details, and export is one of the most sensitive component of GDP. Therefore, precise information

concerning BoP obtained from a coherent research system for transborder areas is crucial.

The analysis of the results from monitoring and surveys show clearly higher activity of enterprises and households operating in the border areas. This also means that the socio-economic phenomena in these areas have a significant impact on the processes of economic growth. This is evidenced by the relatively high proportion of transborder balance in creating economic growth in Poland. It is particularly important during economic downturn when higher activity of enterprises and households in these areas acts as a stabilizer of socio-economic situation.

What is worth stressing is the supranational and multidimensional nature and the scale of transborder processes. In consequence, the functioning of a coherent research system for transborder areas has been providing opportunity to use of the results of analyses on micro-meso-macroeconomic level. The results of the survey on the scale of foreigner's expenditures allow entrepreneurs to set up firms or branches in transborder areas. Simultaneously, local authorities, having this kind of information, can create additional incentives for development of entrepreneurship. On the regional level, functioning of such system makes it possible for self-government and government institutions to lead politics to increase competitiveness of each region. As regards the question of the national level, thanks to this system we can more precisely estimate GDP, BoP to be exact. By means of coherent research system we can take common or compatible decisions on the both side of the border (e.g. common road, migration policy, new border crossings, legislation on local border traffic).

Therefore, it is necessary to develop a system of transborder surveys, which includes both monitoring socio-economic phenomena based on the statistical and non-statistical sources of information as well as introduces and modifies surveys dedicated to transborder areas.

REFERENCES

CIERPIAŁ-WOLAN, M., (2008). Cross-Border Surveys – Some Methodological Aspects, Statistics in Transition new series, Warsaw, Vol. 9, No. 2, 361–366.

CIERPIAŁ-WOLAN, M., (2009). The measures of adaptability in Poland's system of official statistics under crisis, Statistics in Transition new series, Warsaw, Vol. 10, No. 1, 163–170.

CIERPIAŁ-WOLAN, M., (2011). Directions for development of transborder areas – state and prospects, Statistics in Transition new series, Warsaw, Vol.12, No. 3, 537–545.

CIERPIAŁ-WOLAN, M., LIBERDA, Z. B., ŁAGODZIŃSKI, W. W., (2010). Combining of Survey and Administrative Data for Cross-Border Areas, IARIW 31st General Conference, Switzerland.

CIERPIAŁ-WOLAN, M., WOJNAR, E., (2013). Statistical surveying of border traffic and movement of goods and services, [in:] Trans-Border Economies - New Challenges of Regional Development in Democratic World, Edited by Cierpiał-Wolan M., Oleński J., Wierzbieniec W., Jarosław, 99–108.

KISH, L., (1991). A taxonomy of elusive populations, Journal of Official Statistics, 7.

Metodologia badania budżetów gospodarstw domowych. Zeszyt metodologiczny zaopiniowany przez Komisję Metodologiczną GUS, (2010). Central Statistical Office – Living Conditions Department, Warsaw.

OKRASA, W., CIERPIAŁ-WOLAN, M., MARKOCKI, P., (2013). Statistical Issues in Analyzing Trans-Border Non-Institutional Activities. Effect on Spatial Inequality of Wellbeing, Paper delivered at the 59th ISI World Statistics Congress, Hong Kong, 25–30 September 2013.

Surveys of trips made by Poles and arrivals of foreigners to Poland - Methodological book, (2013). Edited by Cierpiał-Wolan, M., Jeznach, M., CSO, Warsaw.

THOMPSON, S. K., (2012). Sampling, Third Edition, Wiley.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 429–442

MULTI-STATE MARKOV MODEL: AN APPLICATION TO LIVER CIRRHOSIS

Gurprit Grover[1], V. Sreenivas[2], Sudeep Khanna[3], Divya Seth[4]

ABSTRACT

The control and treatment of chronic diseases is a major public health challenge, particularly for patients suffering from liver disease. In this paper, we propose a frame to estimate survival and death probabilities of the patients suffering from liver cirrhosis and HCC in the presence of competing risks. Database of the admitted patients in a hospital in Delhi has been used for the study. A stochastic illness-death model has been developed describing two liver illness states (Cirrhosis and HCC) and two death states (death due to liver disease and death due to competing risk). Individuals in the study were observed for one year of life at any age x_i. The survival and death probabilities of the individuals suffering from liver cirrhosis and HCC have been estimated using the method of maximum likelihood. The probability of staying in the cirrhotic state is estimated to be threefold higher than that of developing HCC (0.64/0.21) in one year of life. The probability of cirrhotic patient moving to HCC state is twice (0.21/0.11) the probability of dying due to liver disease. HCC being the severe stage, the probability of patient dying due to HCC is three times that of cirrhosis. Markov model proves to be a useful tool for analysis of chronic degenerative disease like liver cirrhosis. It can provide in-depth insight for both the researchers and policy makers to resolve complex problems related to liver cirrhosis with irreversible transitions.

Key words: illness-death model, maximum likelihood, Cirrhosis, Hepatocellular Carcinoma (HCC).

[1] Department of Statistics, University of Delhi, New Delhi. E-mail: gurpritgrover@yahoo.com.
[2] Department of Biostatistics, All India Institute of Medical Sciences, New Delhi. E-mail: sreenivas_vishnu@yahoo.com.
[3] Department of Gastroentrology, Pushpawati Singhania Research Institute, New Delhi. E-mail: khannasudeep@hotmail.com.
[4] Department of Statistics, University of Delhi, New Delhi. E-mail: divyaseth1234@gmail.com.

1. Introduction

Chronic disease is a long-lasting condition that can be controlled but not cured. Although chronic diseases are among the most common and costly health problems, they are also among the most preventable diseases and can be effectively controlled. Our health care system is heavily weighted in dealing with the problems that are best managed with the patient in a passive role. Chronic conditions and diseases such as heart disease, stroke, cancer, diabetes, etc. are leading causes of mortality and morbidity, accounting for 60% of total premature deaths all over the world out of which 53% is in India [WHO]. Research suggests that complex conditions such as diabetes & depression will impose an even greater health burden in the near future. Study after study, regardless of the underlying disease, has shown generally poor performance in caring for patients with chronic disease.

In studies of chronic disease progression, interest focuses on the rate at which individual progress through a defined set of disease states. The evolution of chronic degenerative disease is characterized by progression of the disease through intermediate states to advanced disease and death. Analysis of such studies can be successfully performed by multi-state models [4]. In the multi-state framework, issues of interest include the study of the relationship between covariates and disease evolution, estimation of transition probabilities and survival rates. A multi state Markov model is developed for survival data analysis for patients suffering from chronic liver disease. Stochastic multistate or competing models, like Markov chains are those best suited to the analysis of such phenomena [10-13]. In this research paper two illness states, viz. cirrhosis and HCC, and two death states, viz. death due to liver disease and death due to competing risk respectively are considered. Accordingly, transition probabilities from one state to another are estimated in the absence of covariates.

In chronic liver disease, the deterioration of the liver functions occurs slowly, over a period of time. Patients who suffer from chronic liver disease may develop cirrhosis after years of disease. Cirrhosis is a form of chronic liver injury that represents an end stage of virtually any progressive liver disease. In 1977, the World Health Organization defined cirrhosis as a diffuse liver process characterized by fibrosis and the conversion of normal liver architecture into structurally abnormal nodules. Liver failure/cancer occurs when the liver loses its ability to function properly. It is a progressive condition that causes severe damage to the liver. It may take months or even years for liver cancer to develop.

In human population, all individuals are not equally healthy and their chance of dying varies from individual to individual. It is known that illness and death are two different types of events. Illness may be transient, repetitive and reversible, whereas death is an irreversible or absorbing state [11]. Further complexity is introduced when the probability of an individual dying from one cause is influenced by the presence of competition of other causes. Death is usually attributed to a single cause, however various risks compete for the life of an

individual. Competing risks can be defined as an event whose occurrence precludes the occurrence of another event under examination [8]. The expressions for survival and death probabilities of liver patients have been obtained by using the concept of crude probability of death under competing risks. The likelihood estimates of the survival and death probabilities have also been obtained. Several stochastic illness-death models have been proposed for studying progression of human diseases. Fix and Neyman proposed a model with two illness and two death states to study human cancer [7]. Grover et al. proposed an illness-death model for estimating survival and death probabilities under cardio vascular shocks in the presence of competing risks [9]. Various shock models categorizing shocks as major and minor have been developed for predicting the survival time function of CHD patients [2].

Chronic infection with hepatitis C virus (HCV) has been estimated to affect 3.2 million persons in the United States and 130 million worldwide and is a leading cause of liver failure and the need for liver transplant [1,14]. Of patients exposed to the hepatitis C virus (HCV), approximately 80% develop chronic hepatitis C and of those, about 20–30% will develop cirrhosis over 20–30 years. Many of these patients have had concomitant alcohol use, and the true incidence of cirrhosis due to hepatitis C alone is unknown. HCV is a noncytopathic virus, and liver damage is probably immune-mediated. Progression of liver disease due to chronic hepatitis C is characterized by portal-based fibrosis with bridging fibrosis and nodularity developing, ultimately culminating in the development of cirrhosis. In cirrhosis due to chronic hepatitis C, the liver is small and shrunken with characteristic features of a mixed micro and macronodular cirrhosis seen on liver biopsy. In addition to the increased fibrosis that is seen in cirrhosis due to hepatitis C, an inflammatory infiltrate is found in portal areas with interface hepatitis and occasionally some lobular hepatocellular injury and inflammation.

Similar findings are seen in patients with cirrhosis due to chronic hepatitis B. Of patients exposed to hepatitis B, about 5% develop chronic hepatitis B, and about 20% of those patients will go on to develop cirrhosis. Special stains for HBc (hepatitis B core) and HBs (hepatitis B surface) antigen will be positive and ground glass hepatocytes signifying HBsAg (hepatitis B surface antigen) may be present. The worldwide incidence of HCC has increased, mostly due to persistent HBV or HCV infection; presently it constitutes the fifth most common cancer, representing around 5% of all cancers. Approximately 25% of these individuals may ultimately develop cirrhosis. Hepatocellular Carcinoma occurs at a rate of 1% to 4% per year after cirrhosis is established [5] and cirrhosis underlies HCC in approximately 80% to 90% of cases worldwide [3].

The purpose of this paper is to develop a stochastic illness-death model using the concept of competing risks theory. The survival and death probabilities of the individuals in different states of chronic liver disease have been estimated using the method of maximum likelihood.

2. Materials and methods

In our model, we have considered two illness states viz. cirrhosis (C) and liver cancer (HCC) and two death states viz. death due to liver disease (DLD) and death due to competing risks (DCR).

A schematic representation of the illness-death model is provided in Fig.1. It is assumed that initially the patient is in cirrhosis stage from which he can transfer to liver cancer state, i.e. HCC, and then to DLD state, representing death due to liver disease or DCR state, representing death due to competing risks. Patient in HCC state can either die due to liver disease so he can transfer to DLD or he can die of competing risks so can move to DCR. It is assumed that liver cancer or hepatocellular carcinoma (HCC) is an irreversible disease therefore model does not allow a transfer from HCC back to cirrhosis.

Patients enrolled in the study have been categorized into two groups namely, cirrhosis and HCC respectively depending upon their initial condition. Suppose that there are n_1 patients in cirrhosis and n_2 patients in HCC state, with the total sample size $n = n_1 + n_2$. Individuals in the study are observed for one year of life i.e. (0,1) at any age x_i so that one year of life means age interval (x_i, x_{i+1}) which corresponds to (0,1).

Let $\lambda_1 t$ and $\lambda_2 t$ be the unknown morbidity intensity functions and $\lambda_3 t$, $\lambda_4 t$, $\lambda_5 t$ and $\lambda_6 t$ be the unknown mortality intensity functions. Each intensity function represents instantaneous transition between respective states at time t ($x_i < t \leq x_{i+1}$).

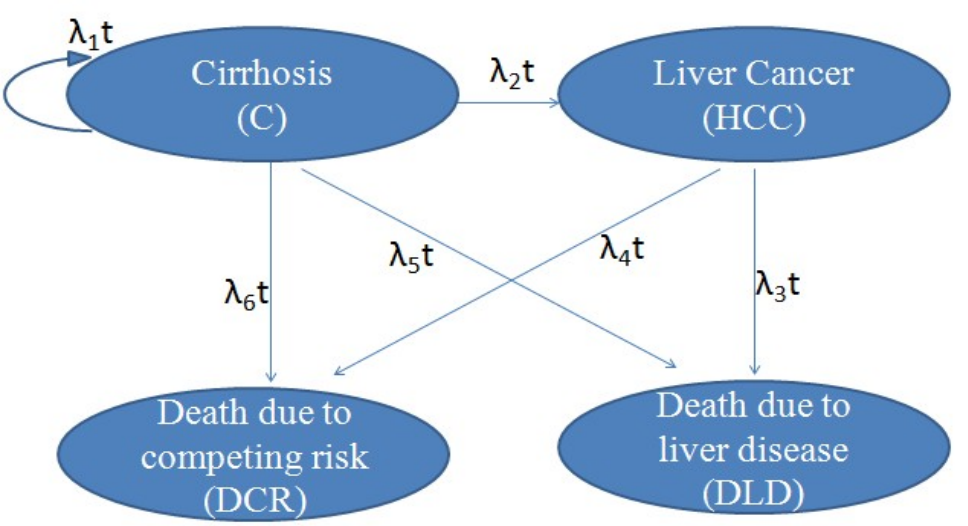


Figure 1. Illustration of illness and death states along with possible transitions

2.1. Calculation of transition probabilities

Let

$\lambda_1 t = -(\lambda_2 t + \lambda_5 t + \lambda_6 t)$

and

$\lambda_2 t = -(\lambda_3 t + \lambda_4 t)$

The relationships between the intensity functions and transition probabilities for the general illness-death process are given by Chiang (1968). In the present case, the survival and death probabilities of an individual in different states of the model may be defined as follows:

$P_{11}(0,1)$ = P (of surviving one year of life while being in the cirrhotic state with hazard rate $\lambda_1 t$)

$$\Rightarrow P_{11}(0,1) = \exp\{\int_0^1 \lambda_1 t dt\} \quad (1)$$

$P_{12}(0,1)$ = P (of surviving one year of life while being in HCC state with hazard rate $\lambda_2 t$)

$$\Rightarrow P_{12}(0,1) = \int_0^t \exp(\int_0^u \lambda_1 \tau d\tau)\lambda_2 u \exp(\int_u^t \lambda_2 \tau d\tau) dt \quad (2)$$

$Q_{13}(0,1)$ = P (of dying due to liver disease after experiencing HCC with hazard rate $\lambda_3 t$ during period (0,1))

$$\Rightarrow Q_{13}(0,1) = \int_0^t \int_0^t e \exp(\int_0^u \lambda_1 \tau d\tau)\lambda_2 u du \exp(\int_u^t \lambda_2 \tau d\tau)\lambda_3 t dt \quad (3)$$

$Q_{14}(0,1)$ = P (of dying from competing risks after experiencing HCC with hazard rate $\lambda_4 t$ during period (0,1))

$$\Rightarrow Q_{14}(0,1) = \int_0^t \int_0^t e \exp(\int_0^u \lambda_1 \tau d\tau)\lambda_2 u du \exp(\int_u^t \lambda_2 \tau d\tau)\lambda_4 t dt \quad (4)$$

$Q_{15}(0,1)$ = P (of dying due to liver disease without experiencing HCC with hazard rate $\lambda_5 t$ during the period (0,1))

$$\Rightarrow Q_{15}(0,1) = \int_0^t \exp(\int_0^t \lambda_1 \tau d\tau)\lambda_5 t dt \quad (5)$$

$Q_{16}(0,1)$ = P (of dying from competing risks without experiencing HCC with hazard rate $\lambda_6 t$ during the period (0,1))

$$\Rightarrow Q_{16}(0,1) = \int_0^t \exp\left(\int_0^t \lambda_1 \tau d\tau\right)\lambda_6 t dt \quad (6)$$

Using equations (1) to (6), it can be easily verified that

$$P_{11}(0,1) + P_{12}(0,1) + Q_{13}(0,1) + Q_{14}(0,1) + Q_{15}(0,1) + Q_{16}(0,1) = 1$$

Now, the survival and death probabilities of an individual in different states of the model, while being initially in HCC state, may be defined as follows:

$P_{22}(0,1)$ = P (of surviving the period (0,1) while being in HCC state with hazard rate $\lambda_2 t$ given that the individual was in HCC state at the time of the start of the study)

$$=> P_{22}(0,1) = \exp\left(\int_0^t \lambda_2 t dt\right) \quad (7)$$

$Q_{23}(0,1)$ = P (of dying in the period (0,1) from liver disease with hazard rate $\lambda_4 t$ given that the individual was suffering from HCC at the time of the start of the study with hazard rate $\lambda_2 t$)

$$=> Q_{23}(0,1) = \int_0^t \exp\left(\int_0^t \lambda_2 \tau d\tau\right) \lambda_3 t dt \quad (8)$$

$Q_{24}(0,1)$ = P (of dying in the period (0,1) from a competing risk with hazard rate $\lambda_4 t$ given that the individual was suffering from HCC at the time of the start of the study with hazard rate $\lambda_2 t$)

$$=> Q_{24}(0,1) = \int_0^t \exp\left(\int_0^t \lambda_2 \tau d\tau\right) \lambda_4 t dt \quad (9)$$

It can be easily verified from the above mentioned (7) to (9) equations that $P_{22}(0,1) + Q_{23}(0,1) + Q_{24}(0,1) = 1$

2.2. Estimation of survival and death probabilities

Maximum likelihood estimation is the recommended procedure for estimating the parameters of the foregoing model. As the data has been collected at specific time points so information on an individual's passage through the states will not usually be complete, i.e. we will only know an individual's status at several, possibly pre-chosen, points in time. Thus, if Li patients are initially enrolled in the study, then after observing them for one year of their life, we found them in different stages.

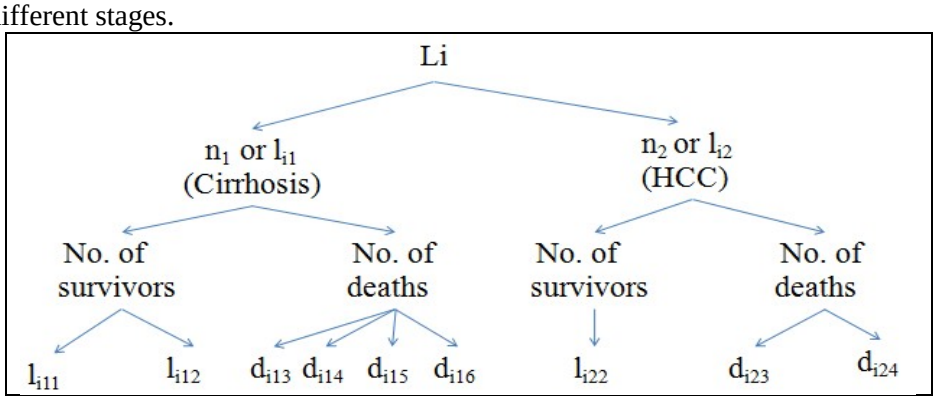


Figure2. Graphical representation of the used notations

For n_1 group of individuals:

l_{i11}, number of survived persons in cirrhotic state during the age interval (x_i, x_{i+1});

l_{i12}, number of survived persons who transfer from cirrhotic state to HCC state during the age interval (x_i, x_{i+1});

d_{i13}, number of persons died due to liver disease after suffering from HCC during age interval (x_i, x_{i+1});

d_{i14} = number of persons died due to competing risks after suffering from HCC during age interval (x_i, x_{i+1});

d_{i15} = number of persons died due to liver disease after suffering from cirrhosis during age interval (x_i, x_{i+1});

d_{i16} = number of persons died due to competing risks after suffering from cirrhosis during age interval (x_i, x_{i+1});

d_{i1}, total number of dead persons who were in cirrhotic state at the beginning of the age interval (x_i, x_{i+1});

$d_{i1} = d_{i13} + d_{i14} + d_{i15} + d_{i16}$;

P_{i1} = P (of surviving the age interval (x_i, x_{i+1}) initially in cirrhotic state);

Q_{i1} = P (of dying in the age interval (x_i, x_{i+1}) initially in cirrhotic state).

Table 1. The number of survivors and the number of deaths due to various risks in the ith age interval with corresponding probabilities

		Frequency	Probability
No. of deaths	Due to liver disease after suffering from HCC	d_{i13}	Q_{i3}
	Due to CR after suffering from HCC	d_{i14}	Q_{i4}
	Due to liver disease after suffering from Cirrhosis	d_{i15}	Q_{i5}
	Due to CR after suffering from cirrhosis	d_{i16}	Q_{i6}
No. of survivors	Individuals in cirrhotic state	l_{i11}	P_{i1}
	Individuals in HCC state	l_{i12}	P_{i2}
Total		$l_{i1+1} + d_{i1}$	1

Given l_{i1} individuals alive at the beginning of the age interval (x_i, x_{i+1}), the joint distribution of l_{i11}, l_{i12}, d_{i13}, d_{i14}, d_{i15} and d_{i16} is multinomial with probability function

$$f_{i1} = \frac{l_{i1}!}{\prod_{j=1}^{2} l_{i1j}!} \frac{1}{\prod_{k=3}^{6} d_{i1k}!} \prod_{j=1}^{2} P_{ij}^{li1j} \prod_{k=3}^{6} Q_{ik}^{di1k}$$

For a sample of u age intervals, the likelihood function is given by

$$L_1 = \prod_{i=0}^{u} \frac{l_{i1}!}{\prod_{j=1}^{2} l_{i1j}!} \frac{1}{\prod_{k=3}^{6} d_{i1k}!} \prod_{j=1}^{2} P_{ij}^{li1j} \prod_{k=3}^{6} Q_{ik}^{di1k} \quad \text{----------(a)}$$

although it is both biologically and mathematically pleasing that the proposed model has arisen from consideration of a progression of events in a stochastic process. It is more expedient for parameter estimation and subsequent interpretation to deal in terms of probability distributions of the random variables arising from the process. Using the equation (a), maximum likelihood estimates of P_{ij} and Q_{ik} can be computed.

For a group of individuals initially in HCC state,

$l_{i2,}$ number of survived individuals in HCC state at the beginning of the age interval (x_i, x_{i+1}),

$l_{i22,}$ number of survived individuals in HCC state surviving the age interval (x_i, x_{i+1}),

$d_{i23,}$ number of died individuals due to liver disease in the age interval (x_i, x_{i+1}),

d_{i24}, number of died individuals due to competing risks in the age interval (x_i, x_{i+1}),

d_{i2}, total number of individuals dying in the age interval (x_i, x_{i+1}) and were in HCC state in the beginning of the interval,

$=> d_{i2} = d_{i23} + d_{i24}$,

$l_{i2+1} = l_{i22}$,

P_{i2} = P (of surviving the age interval (x_i, x_{i+1}) when an individual is in HCC state at time x_i),

Q_{i2}= P (of dying in the age interval (x_i, x_{i+1}) when an individual is in HCC state at time x_i).

Table 2. The number of deaths due to various risks and the number of survivors in the i^{th} age interval with corresponding probabilities

		Frequency	Probability
No. of deaths	Due to HCC	d_{i23}	Q_{i3}
	Due to competing risk	d_{i24}	Q_{i4}
No. of survivors	In HCC state	l_{i22}	P_{i2}
		$d_{i2} + l_{i2+1}$	1

Given l_{i2} individuals in HCC state alive at the beginning of the age interval (x_i, x_{i+1}), the joint distribution of $l_{i22,}$ l_{i27} and l_{i28} is multinomial with probability function

$$f_{i2} = \frac{l_{i2}!}{l_{i22}!d_{i27}!d_{i28}!} P_{i3}^{\,l_{i22}} Q_{i7}^{\,d_{i27}} Q_{i8}^{\,d_{i28}}$$

For a sample of u age intervals, the likelihood function is given by

$$L_2 = \prod_{i=0}^{u} f_{i2}$$

$$\Rightarrow L_2 = \prod_{i=0}^{u} \frac{l_{i2}!}{l_{i22}!d_{i27}!d_{i28}!} P_{i3}^{\,l_{i22}} Q_{i7}^{\,d_{i27}} Q_{i8}^{\,d_{i28}}$$

Similar method can be applied as mentioned above to obtain estimates of P_{ij} and Q_{ik} for the patients initially in HCC state.

3. Application

A total of 366 patients suffering from liver cirrhosis and 30 suffering from HCC were admitted in the years 2007-2008 at Pushpawati Singhania Research Institute (PSRI). Data on these patients were collected retrospectively and in total 66 censored patients were excluded from the analysis. Patients were observed for one year of life, i.e. (0,1) at any age x_i. 310 cirrhotic and 20 HCC patients moved through various stages (Figure.1) and were observed for one year of their life as shown in the Table 3.

Table 3. Number of patients observed in various stages after 1 year

State at the beginning of the study	Transition in the interval (0,1)	State after one year	Number of individuals
Cirrhosis	Cirrhosis -› Cirrhosis	Cirrhosis	204
	Cirrhosis -› HCC Cirrhosis -› HCC-› DLD Cirrhosis -› HCC -› DCR	HCC	20
		Death due to liver disease	8
		Death due to competing risk	2
	Cirrhosis -› DLD	Death due to liver disease	70
	Cirrhosis -› DCR	Death due to competing risk	6
HCC	HCC -› HCC	HCC	8
	HCC -› DLD	Death due to liver disease	10
	HCC -› DCR	Death due to competing risk	2

4. Results

Table 4. The estimated intensities and the corresponding survival and death probabilities of patients initially in cirrhosis state

Parameters	Intensities	Probabilities
λ_1	-0.88961	$P_{11}(0,1) = 0.640149$
λ_2	0.60526	$P_{12}(0,1) = 0.208471$
λ_3	0.50000	$Q_{13}(0,1) = 0.036499$
λ_4	0.10526	$Q_{14}(0,1) = 0.007680$
λ_5	0.26415	$Q_{15}(0,1) = 0.106612$
λ_6	0.02020	$Q_{16}(0,1) = 0.008152$

Table 5. The estimated intensities and the corresponding survival and death probabilities of patients initially in HCC state

Parameters	Intensities	Probabilities
λ_2	-0.77196	$P_{22}(0,1) = 0.61402$
λ_3	0.66667	$Q_{23}(0,1) = 0.33334$
λ_4	0.10526	$Q_{24}(0,1) = 0.05263$

The estimated intensities and their corresponding survival and death probabilities for Markov model (as explained in Figure1) are presented in Table 4 and Table 5. It can be inferred from Table 4. that the probability of patients who are initially in cirrhosis stage has higher chances of staying in the same stage than that of moving to HCC stage, i.e. the probability of remaining in cirrhosis stage is threefold higher than that of moving in HCC stage (0.640/0.208) during 1 year of time span. It may also be noted that probability of a cirrhotic patient moving in HCC stage is almost twice the probability of the patient dying due to liver disease (0.208/0.107), i.e. patients are more likely to move to HCC stage than that of dying from cirrhosis. Also, the probability of patients dying due to competing risk is almost equal, irrespective of their disease severity.

On comparing both the above Tables (4 and 5), it can be inferred that the probability of patients dying due to HCC is three times the probability of patient

dying due to cirrhosis. Also, we have found that in one year of time span patients initially in cirrhosis stage are less likely to die from liver disease than the patients initially in HCC stage. The reason could be that one year of time span might be less to witness all the stages of chronic liver disease. The above discussed intensities and probabilities have been depicted by the graphs shown below.

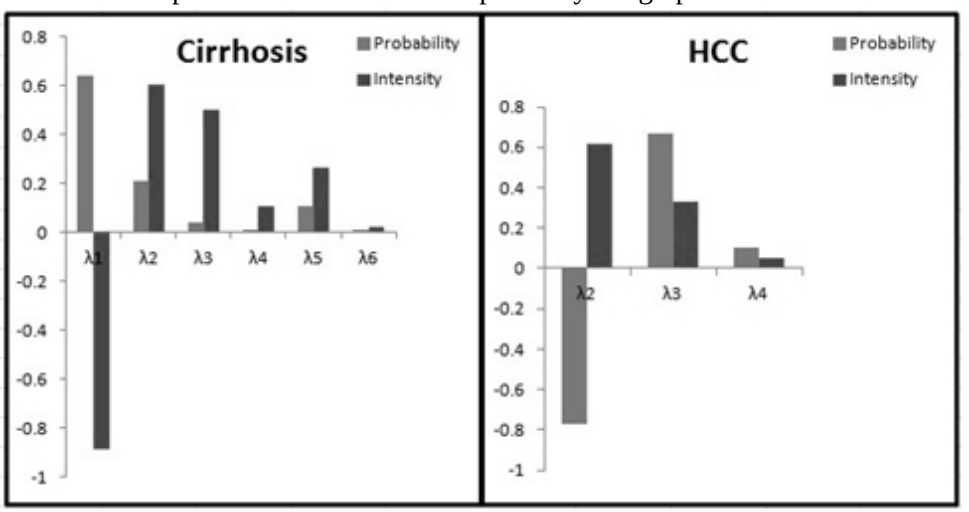


Figure 3. Survival and death probabilities and intensities

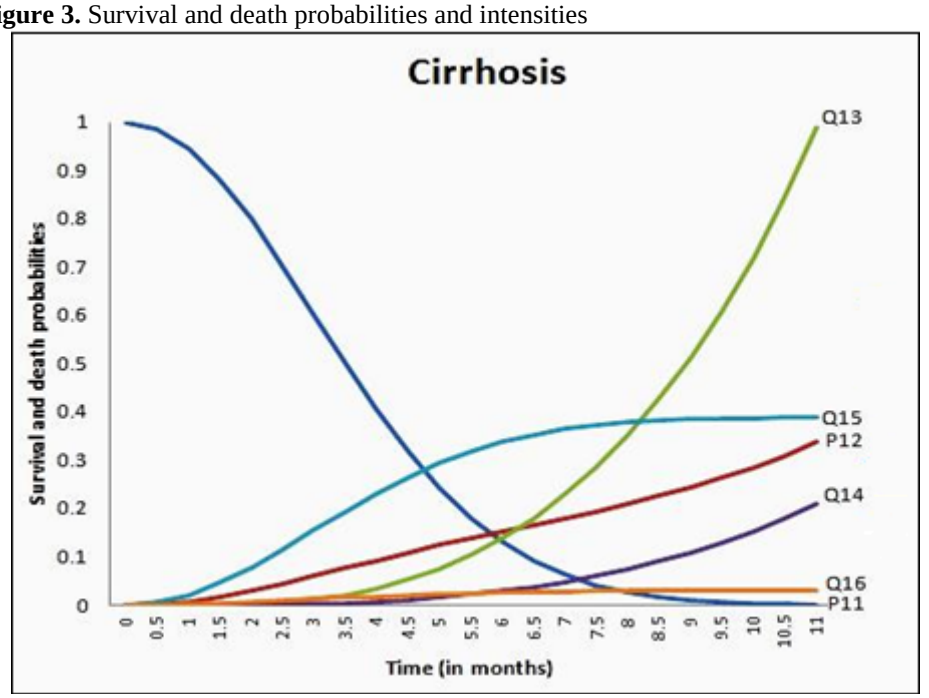


Figure 4. Illness and death transition probabilities for stages 1-4, over the period of 12 months for the patients initially in cirrhosis stage

The graph above depicts the clear picture of the movement of patients suffering from liver cirrhosis into various stages as defined in Fig1. Survival and death probabilities of the patients initially suffering from cirrhosis are illustrated in the graph at monthly intervals. A sharp decline in the curve (P_{11}) can be noticed in the first month representing that out of the patients initially in the cirrhosis stage, a few remains in the same stage after one year of their life but with very less probability. The intensity of moving to HCC (P_{12}) stage showed increment from the first month onwards, i.e. increasing with time at normal pace. The intensity of patients moving from cirrhosis to HCC and then to death is showing a parabolic trend, i.e. Q_{13} shows a steep curve depicting a sharp increase from the 6.5th month onwards. It simply shows that once the patient reaches HCC stage the probability of reaching death stage increases rapidly. The probability of patients dying due to various opportunistic causes/competing risks after suffering from HCC showed an increasing trend after the 6th month. The intensity of moving from cirrhosis to death showed a sharp increase in the 3rd month, until the 8th month it increases slowly but after that it almost remained constant. The intensity of moving to death stage due to various opportunistic causes/competing risks remains parallel to x axis after 2nd month acting like a tangent to x-axis.

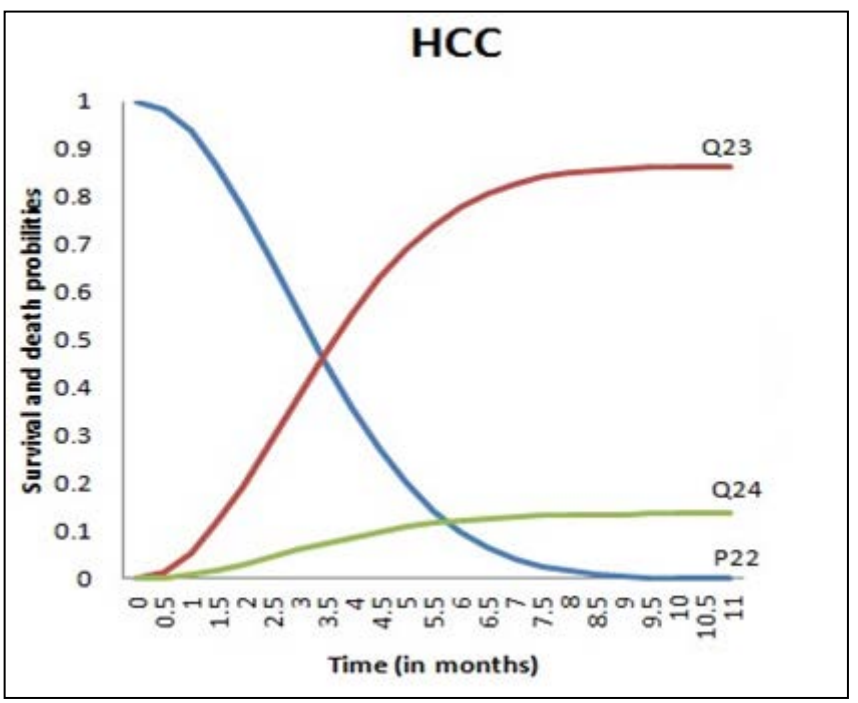


Figure 5. Illness and death transition probabilities for stages 1-4, over the period of 12 months for the patients initially in HCC stage

Figure 5 gives a depiction of the movement of the patients initially suffering from HCC to two stages, viz. death due to liver disease (Q_{23}) and death due to

competing risks (Q_{24}). A steep downfall can be observed in the intensity of patients remaining in the HCC stage after one year of their life, i.e. the probability of patients remaining in HCC state after one year is very low. Patients quickly move to death stage once they reach the HCC stage. Vice versa, intensity of moving from HCC to death stage depicts a sharp increase from 1st month onward. It increased swiftly until the 7th month and then it became almost constant with probability of 0.9. As discussed above, there are higher chances of patients moving to death stage if they are in HCC stage. The intensity of moving to death stage due to competing risks increases after 2nd month and then became constant after 7th month.

5. Discussion

In studies of many chronic medical conditions, the health status of a patient may be characterized using a finite number of disease states. When patients are observed over time, the dynamic nature of the disease process may then be examined by modeling the rates at which patients make transitions among these states. Markov models provide a convenient framework for such analyses and have been very widely adopted in fields such as infectious disease [Bailey, 1982], neurology [Hassani and Ebbutt, 1996], and rheumatology [Gladman, Farewell, and Nadeau, 1995]. Frequently, only two states are of interest to represent, e.g., the presence or absence of symptoms [Frank et al., 1990] or the presence or absence of infection [de Vlas et al., 1993]. Such disease processes operate in continuous time, but for practical and economic reasons, subjects are often only observed at discrete time points. Indeed, under the most general scenario, the observation times may be irregularly spaced and may be unique to each subject. Data obtained from such an observation scheme are termed panel data. Kalbfleisch and Lawless (1985) developed an efficient algorithm for obtaining maximum likelihood estimates of the transition rates from panel data under a time-homogeneous Markov assumption and derived the form of an asymptotic covariance matrix for the corresponding estimators based on the expected information matrix. Grover et al. have also considered a similar problem taking cardiac arrest as disease with 4 states viz. normal state, illness state and two death states. In this research paper, we have illustrated the usefulness of Markov models in the analysis of chronic liver disease. We have implemented methods introduced by Chiang [4] that allow quite general models to be fitted. Various intensities and their corresponding transition probabilities have been computed for all four stages. These results have been obtained in the absence of covariates. We have reported the risk in terms of probability and also compared the probability of death from both (cirrhosis and HCC) the illness states. The concept of competing risk has been introduced for the first time in the study related to liver cirrhosis. A further improvement of the model could be that of considering covariates. Also, survival time and stay time in every state can be computed.

REFERENCES

ARMSTRONG, G. L., WASLEY, A., SIMARD E. P., MCQUILLAN G. M., KUHNERT, W. L., ALTER, M. J., (1999). The prevalence of hepatitis C virus infection in the United States, Annals of Internal medicine, 141, 705–714.

BISWAS, S., NAIR G., SEHGAL, V. K., (1987). Successive damage shock models predicting system survival time, International Journal Systems Sciences, 18(5), 831–836.

CHEN, H. H., DUFFY, S. W., TABAR, L., (1996). A Markov chain method to estimate the tumor progression rate from preclinical phase, sensitivity and positive predictive value for mammography in breast cancer screening, The Statistician, 45(3), 307–317.

CHIANG, C. L., (1968). Introduction to Stochastic Processes in Biostatistics, John Wiley, New York.

DI BISCEGLIE, A. M., (1997). Hepatitis C and hepatocellular carcinoma. Hepatology, 26 (3 suppl 1), 34S–38S.

FATTOVICH, G., STROFFOLINI, T., ZAGNI, I., DONATO, F., (2004). Hepatocellular Carcinoma in Cirrhosis: incidence and risk factors. Gastroentrology, 127, S35–S50.

FIX, E., NEYMAN, J., (1951). A simple stochastic model of recovery, relapse, death and loss of patients, Human Biology, 23, 205–241.

GOOLEY, T. A., LEISENRING W., CROWLEY J., STORER B. E., Estimation of failure probabilities in the presence of competing risks: New Representations of old Estimators, Stat. Med.,18: 695–706.

GROVER, G., MAKHIJA, N., (2004). On the estimation of survival and death probability under cardio-vascular shocks in the presence of competing risks based on an illness-death model, Journal of institute of science and technology, 13, 130–142.

PUTTER, H., FIOCCO, M., GESKUS R. B., (2007). Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine, 26, 2389–2430.

SACKS, S. T., CHIANG, C. L., (1977). A transition probability model for the study of chronic diseases. Mathematical Biosciences, 34, 325–346.

SERIO, G., MORABITO, A., (1988). Considerations on a staging process for chronic diseases. Rivista di Statistica Applicata, 21(23), 335–348.

SERIO, G., MORABITO, A., (1986). Stochastic survival model with covariates in cancer. Modelling of Biomedical Systems, 91–96.

WILLIAMS, R., (2006). Global challenges in liver disease, Hepatology, 44, 521–526.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 443–466

THE PROPERTIES OF ATMs DEVELOPMENT STAGES- AN EMPIRICAL ANALYSIS

Henryk Gurgul[1], **Marcin Suder**[2]

ABSTRACT

This paper addresses the crucial problem of the ATM's network management which is so-called the saturation level of withdrawals. This notion refers to mean level of withdrawals after dropping particular withdrawals realized in the initial time period, (i.e. time period after activation of ATM) and the length of elapsing time period necessary to reach saturation level. One can observe that the level of withdrawals and their number stabilize as time elapses. The paper aims to define average withdrawals after achieving saturation level and mean time necessary to stabilize withdrawals (based on historical data). In addition, we established that – under condition of similarity in terms of location and date of start - ATMs exhibit similar characteristics of the development effects. This allows us for predicting the size of time necessary to achieve saturation and the average withdrawal in the state of saturation.
JEL Classification: C49, C53

Key words: ATM, withdrawals, saturation level.

1. Introduction

The data on the banks shows the changes over time in the use of technology. One can observe the shift to IT-based delivery systems like ATMs and Internet banking. In addition, the expansion of financial technologies such as financial derivatives and off-balance sheet credit commitments can be noticed. The data on banks also shows improvements in bank performance and consolidation of the industry during the deployment of new technologies. However, establishing the links between technological progress and banking industry productivity growth and the structure of this industry requires extended analyses.

There is a large body of literature on financial innovations. The contributors stress importance of the topic for understanding money demand. However, it is not easy to find in the literature the empirical analysis dealing with management problems resulting from the use of the financial innovations such as the increase in the number of bank branches and ATMs terminals.

[1] AGH University of Science and Technology in Cracow. E-mail: henryk.gurgul@gmail.com.
[2] AGH University of Science and Technology in Cracow. E-mail: m_suder@wp.pl.

The extensive analyses on this subject require large amounts of data on customers. This data is necessary in order to improve services conducted by financial institutions. The learning of habits and patterns of customer behavior on the basis of the empirical financial data is very important because it supports customer value management.

The one of the most important issues addressed in our contribution is the time necessary to achieve full functionality of the ATM (i.e. its maturity). The elapsed time determines realization of the expected profits. Besides time of full capacity (maturity) appointment very a important parameter is the size of withdrawals volume achieved after stabilization of weekly withdrawals. The knowledge of stabilization time point is essential in respect for modeling of withdrawals by time series models. The data prior to this point should be removed from the withdrawals time series used in estimation procedure. It is conjectured in the paper that the time point and the size of the stabilized withdrawals depend on consumer habits and on activity of competitive firms acting in the surroundings of the ATMs under study. To the best of our knowledge this topic has not been so far the subject of research in Poland.

The content of this paper is as follows. Next section reviews most important contributions concerning the ATMs. In the third section the dataset and methodology are presented. In the following fourth section the empirical results of estimation of saturation level and stabilization point of ATMs are showed and compared. Fifth section is concerned with the impact of ATM location on maturity of ATM. In the sixth section the examples of forecasts of the length of the development period and the saturation level are cited. Finally, in the last section we summarize major conclusions and suggest directions for future research.

2. Literature overview

The early studies on bank profitability checked many important bank-specific factors that could determine bank profitability (Rhoades and Rutz, 1982; Clarke et al., 1984; Smirlock, 1985; Evanoff and Fortier, 1988; Lloyd-Williams and Molyneux, 1994; Molyneux and Forbes, 1995; Naceur and Goaied, 2001, etc.). They found that some of them had an influence on bank profits. However, none of these studies checked the influence of investment in information technology (IT) systems on bank success.

The business strategies of banks located in Poland with respect to delivery channels have changed in recent years. In the past, the most important strategies involved how to increase the number of branches under the regulations on bank branching. However, in recent years, besides of increasing the number of branches the establishing branches that can offer full services (i.e. banking services, security services, insurance services, trust services, home loan services, etc.) at each branch location have become popular.

Banks also try to use information technology because of fast developments in this field and the popularization of IT in each country. Customers cannot only shorten their waiting time but also use bank services after closing time through Automated Teller Machines and Internet banking. Application of IT may enable banks reduction of excessive personnel costs. In addition, financial institutions can start new businesses which require the use of ATMs and other machines. An effective use of these IT tools makes possible that banks make more profits by introducing IT systems.

Holden and El-Bannany (2004) were the first who stressed the importance of investment in IT systems for bank success. In particular, they were concerned with the effects of investing in ATMs that perform routine banking transactions instead of tellers. The contributors argued that the use of ATMs can reduce banking transaction costs, the number of personnel and the number of manned branches, and can also increase bank profit. The authors used the number of ATMs of individual banks as a proxy for the measure of investment in IT systems, and analysed whether ATMs influenced profits by using data from the United Kingdom. An analysis by Holden and El-Bannany (2004) suggested that ATMs play an important role in increasing Return On Asset (ROA) of banks in the United Kingdom.

Many Japanese banks also took for granted that ATMs are important machines that can reduce banking transaction costs, excessive personal costs and branching costs.

In order to prove this conviction Kondo (2010), based on the type of analysis carried out by Holden and El-Bannany (2004) checked whether the fact that ATMs increase the ROA, as shown for banks in the United Kingdom by Holden and El-Bannany (2004), is also true for Japanese banks. He also tried to prove the effects of ATMs on the fees and commissions (income) and the interest income that could be increased by ATM actions. His results suggested that the ROA of Japanese banks is not fully explained by the variables used by Holden and El-Bannany (2004). In particular, he found that ATMs do not influence the ROA directly in Japan. However, the empirical results of the author do not contradict the thesis that ATMs contribute to increased profits, especially in transactions in which their primary functions are used. Kondo (2010) conducted also further analysis with respect to the effects of ATMs on fees and commissions, including income collected by ATMs, and on interest income, including investment profits produced by using money collected by ATMs. He found that ATMs had positive and significant effects on the fees and commissions from 2000 to 2003, but no significant effects of ATMs were found for 2004. However, positive and significant effects of ATMs on interest income were detected in the following years of the study period. This suggests that banks that actively establish ATMs and collect deposits by ATMs are investing well and succeed in increasing profits.

The findings indicate that ATMs of Japanese banks do not influence the ROA that contains the overall profits of bank transactions, which is different from the case in the United Kingdom tested by Holden and El-Bannany (2004), but they do

contribute to those particular transactions that best take advantage of their abilities. Thus, it is important for Japanese banks to invest in ATMs in order to increase profitability.

ATMs withdrawals time series are directly or indirectly reflected in daily activity of customers. They are usually recorded on a daily, monthly, quarterly basis or some other period. Cleveland and Devlin (1980) and Liu (1980) found that, e.g. the number of working days and its link with seasonal effects have impact on the size and the number of withdrawals.

Besides the number of working days, the day of the week, the week of the month, other factors such as public holidays or religious events may also affect withdrawals time series. In addition, withdrawals are often overlaid with paydays and seasonal demand, and often follow weekly, monthly and annual cycles. These issues are addressed in papers such as Simutis *et al.* (2008).

Both the correction for calendar and seasonal effects and detection of time of maturity of ATM and the level of withdrawals may help to improve the performance of short-term macroeconomic forecasts (see Carlsen and Storgaard, 2010; Esteves, 2009, Galbraith and Tkacz, 2007).

The results of statistical analysis on withdrawals data were published by Hand and Blunt (2001), Amromin and Chakravorti (2007), Boeschoten (1998) and Snellman and Viren (2009). The authors assumed that individuals replenish their cash when it falls below a certain threshold. The authors detected that ATM users typically had a lower amount of cash than non-users because the cost to them of obtaining cash was lower. Snellman and Viren (2009) modeled the relationship between the cost of obtaining cash and the number of ATMs. The contributors assumed that costs are proportional to distance from ATMs.

In opinion of Findley and Monsell (2009), Findley et al. (1998), Cleveland and Devlin (1980), Findley and Soukup (1999, 2000, 2001), McElroy and Holland (2005) distribution of withdrawals is very important with respect to modeling, forecasting and replenishment strategy. This problem is complex because of mobile holidays like Easter, Ramadan or the Chinese New Year (Gurgul and Suder, 2012).

The advanced models of ATM withdrawals were developed by Brentnall et al. (2008, 2010). The authors noticed that the empirical distribution of random effects performed well because there was large number of individual accounts.

Kufel (2010) showed by regression based on dummies and harmonic analysis that withdrawals time series for ATMs in Torun exhibited yearly, monthly, weekly and daily seasonality.

In the next section we will present dataset properties and methodology.

3. The dataset and methodology

It is not an easy task to get data on ATM's activity. This is probably one of the main reasons for low number of empirical contributions in this field. The detailed insight is in the following subsection.

3.1. The dataset description

The empirical analysis was based on weekly time series of withdrawals from 93 out of 293 EURONET ATMs located in Małopolskie and Podkarpackie provinces. Our analysis took into account the data from the ATMs, which started operating in 2008 or 2009 and operated continuously until April 20th, 2012. The ATMs which were operating also after 20th of April 2012 could not be taken into the analysis because in such cases the withdrawal data prior to 2008 was unavailable. On the other hand, the time series of withdrawals from the ATMs which started operating after 2009 were too short to conduct the empirical analysis.

The ATMs were divided according to two criteria. The first criterion, called the *population category*, is related to the size of the city/town/village in which the ATM is located. Taking into account the number of inhabitants, this criterion leads to the following clustering:

I. 0 - 20 thousand – Villages and small towns,
II. 21 - 50 thousand – Cities of average size,
III. 51 - 100 thousand – Big cities,
IV. 101 - 200 thousand – Cities of a very large size,
V. Over 201 thousand – Metropolises (Cracow).

The second criterion is related to the characteristics of the ATM's location. In this study we used six types of location of ATM's, namely: bank branch, hypermarket, petrol station, shop, shopping center and the "other" location type. Table 1 presents the data on the number of ATMs in the examined types of locations.

Table 1. The number of ATMs according to the clustering criteria

population category / location category	I	II	III	IV	V	Total
bank branch	0	7	4	7	18	36
hypermarket	0	2	4	2	6	14
other	0	1	0	3	9	13
petrol station	1	1	0	0	4	6
shop	0	5	0	1	9	15
shopping center	1	2	1	3	3	9
total	2	18	9	16	49	93

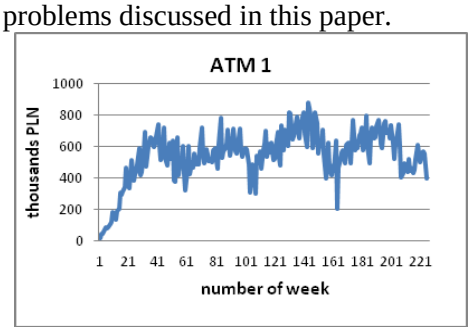
In order to discuss the problems investigated in this paper in a detailed way, we will present the outcomes of an analysis of six selected ATMs, which vary in

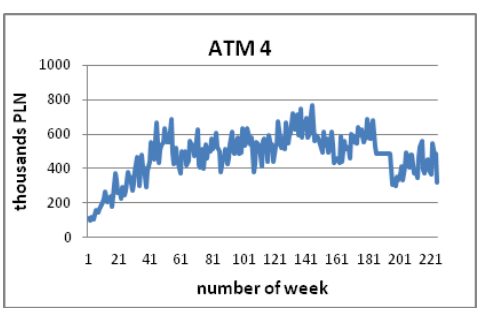
terms of the location. Table 2 contains basic information about each of the selected ATMs, while Figures 1-6 present the charts of the weekly time series of withdrawals from each individual ATM.

Table 2. Basic information on selected ATMs

ATM's number	ATM 1	ATM 2	ATM 3	ATM 4	ATM 5	ATM 6
province	Małopolskie	Małopolskie	Podkarpackie	Podkarpackie	Małopolskie	Małopolskie
town/city	Niepołomice	Andrychów	Stalowa Wola	Rzeszów	Kraków	Kraków
population category	I	II	III	IV	V	V
location type	shopping center	petrol station	bank branch	hypermarket	other	shop
start date	02-01-2008	02-01-2008	01-12-2008	02-01-2008	12-03-2008	05-11-2008
number of obs.	225	225	178	225	215	181
average of daily withdrawal	541 127	342 015	289 524	475 722	277 245	194 353
coefficient of variation	30,04%	28,56%	46,07%	27,87%	24,46%	22,67%
minimum	13 650	52 400	250	99 400	31 900	63 700
maximum	877 500	508 350	537 370	76 9900	417 850	292 650

The data presented in Table 2 indicates significant differences among six ATMs selected for the analysis, which ensures a comprehensive approach to the problems discussed in this paper.





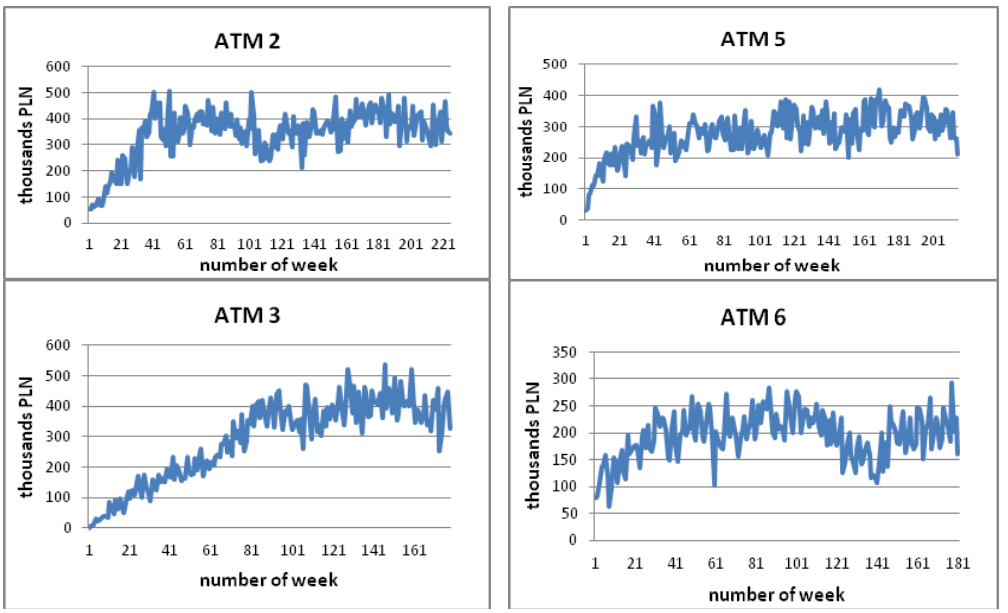


Figure 1. Weekly withdrawals from ATMs 1-6

The analysis of Figure 1 provides a basis to claim that in the initial phase the size of the withdrawals increases significantly for each ATM. However, after some time the growth rate drops markedly. Moreover, one can even see some evidence of stabilization of the withdrawals over time (taking into account the presence of seasonality). In addition, we noted that the time required for the withdrawals to stabilize differs slightly among the selected ATMs.

In the next subsection we cite simple models used in our analysis.

3.2. The methodology

In general, the analysis of the length of the development phase of ATMs was based on two methods.

The first method was based on the analysis of the local trends. In order to determine the time point, at which the withdrawals from the ATM stabilize, i.e. the ATM starts to work in its usual rhythm (taking into account the calendar effect), one should fit the following trend function to the examined dataset:

$$f(t) = \begin{cases} a_1 t + b_1 \ \ if \quad t \le t_0 \\ a_2 t + b_2 \ \ if \quad t > t_0 \end{cases}.$$

By the ordinary least squares (OLS) method the coefficients a_1, b_1, a_2, b_2 and the point t_0 were estimated. Assumption that the trend function takes the

form of the piecewise function consisting of two linear functions allows to determine the point in time at which the significant change in the dynamics of the development of the ATM takes place. In particular, we aim to prove that for the examined data the slope coefficient of the trend function $f_1(t) = a_1 t + b_1$ is much higher than the slope coefficient of the function $f_2(t) = a_2 t + b_2$. The value of the parameter t_0 may be interpreted as the length of the ATM's development phase. On the other hand, the coefficients a_1 and a_2 allow to specify the average growth rate of withdrawals during the period of development and the stable period. Finally, the value of the coefficient b_2 will determine the saturation level of the ATM.

The second method is based on modeling the size of withdrawals from ATMs with the logistic function. We assumed that the development of a specific ATM (size of withdrawals) follows the logistic curve $f(t) = \frac{a}{1+b \cdot e^{-c \cdot t}}$ where $a,b,c>0$. This type of function is often used to study the development of the market after an introduction of a new product. The parameter *a* determines the level of saturation, e.g. sales volume of the new product or (as in our case) the size of withdrawals from an ATM. The logistic function has a point of inflection at $t=\frac{\ln b}{c}$ and changes from convex to concave, which implies that in the neighborhood of the point $t=\frac{\ln b}{c}$ the increase in the value of the function changes from more than proportional to less than proportional[1].

In order to examine the impact of the calendar effect the analysis was performed on the series of weekly withdrawals and the seasonally adjusted series. Since the results obtained for both types of series were in general the same, we will present only the results obtained for the original (unadjusted) data.

In the following section the results of the conducted empirical analysis are showed.

4. Empirical results

In this section we report and evaluate the empirical results derived by local trend method and logistic regression. Next, we will compare the results by both approaches.

[1] In the literature various methods of estimation of the logistic model were introduced, e.g. the Hotelling method or the method of arbitrary fixing the parameter *a* (saturation level). One can also use some professional statistical software. However, if it comes to widely available computer software, it is recommended to perform the calculations in MS Excel Solver package. In this paper, the estimation method was based on minimizing the sum of the squared errors in MS Excel Solver framework.

4.1. Duration of development phase of ATMs by local trend approach

The analysis of the length of the development phase as well as the determination of the saturation levels was carried out for 93 time series. However, the complete results will be presented only for the six selected ATMs. In the further part of this section the descriptive statistics and the distribution of the length of the development phase for all 93 ATMs will be presented.

Figure 2 contains the result of fitting the trend function to the weekly withdrawals from ATM 5. On the other hand, Table 3 contains the results of the analysis performed for the six selected ATMs. We present the estimated values of the coefficients, the significance levels and the values of the coefficients of determination (R^2) for each examined model.

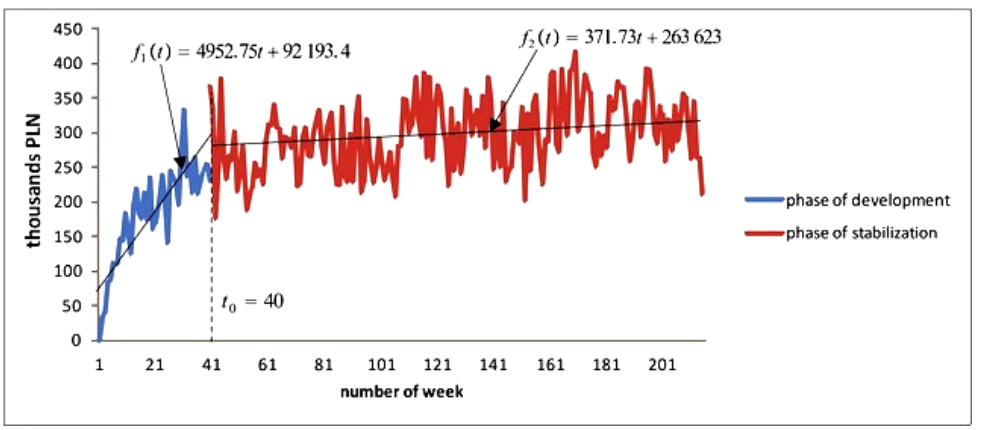


Figure 2. Local linear trend analysis for ATM 5

Table 3. Estimation results (local trend approach) for six selected ATMs

ATM's number	t_0	a_1	b_1	a_2	b_2
ATM 1	41	16 473.1***	10 230.5	351.91**	54 8735***
		R^2=90.57%		R^2=2.56%	
ATM 2	45	9355.51***	9452.55	173.75**	355 638***
		R^2=86.92%		R^2=2.45%	
ATM 3	71	6373.64***	150 213	764.77***	506 132***
		R^2=73.72%		R^2=8.99%	
ATM 4	43	8559.28***	101 182***	1428.5*	479 942***
		R^2=81.08%		R^2=7.27%	
ATM 5	40	4952.75***	92 193.4***	371.73***	263 623***
		R^2=68.53%		R^2=13.68%	
ATM 6	30	6327.84***	76 321.5***	467.22***	252 810***
		R^2=69.99%		R^2=15.65%	

The results show that for all selected ATMs one can easily indicate the point in time at which the stabilization of the size of the withdrawals takes place. For example, for the series of withdrawals from ATM 5 the period after which the stabilization occurs is approximately 40 weeks l (see Figure 2). It turns out that the slope coefficient of the trend function for the first 40 data points is equal to 4952.75. Thus, in this initial phase the size of the weekly withdrawals increases (on average) by 4952.75 zł per week. In the second period ($t_0 > 40$) the slope coefficient of the trend function is equal to 371.73, which implies that the growth rate of the withdrawals decreases by more than 13 times in comparison to the first period. The estimated saturation level of the ATM is equal to 263 623 zł. A similar analysis performed for the daily data and the seasonally adjusted data confirmed these results. Thus, the analysis of the ATM 5 clearly proves that this ATM reached the stabilization of the withdrawals after about 40 weeks. For ATMs 1, 2 and 4 the length of the stabilization period is similar as it varies between 38 and 45 weeks. For ATM 3 the development period is much longer (about 70 weeks). On the other hand, in case of ATM 6 this period is shorter (about 30 weeks).

An important issue in this analysis is the fact that for all ATMs the slope coefficient of the trend function in the first period (development phase) is much higher than the value of the slope coefficient of the trend line for the second period (stable period). In addition, the values of the slope coefficients in the second period are relatively small and their positive values may be associated with a general tendency to increase the size of withdrawals from ATMs with time (e.g. inflation rate, etc.).

An analogous analysis performed for the remaining 87 ATMs proved that in 85 cases the slope coefficients of the trend functions in the first phase were much higher than the slope coefficients in the second period. Only for two ATMs the slope coefficients of the trend function f_1 were lower than the slope coefficients of the function f_2. It is likely that this situation was caused by some additional events which took place in the close neighborhood of the ATMs, e.g. opening of a new shopping center, a new transport hub, etc.

Table 4 presents the descriptive statistics for the lengths of the development phases of the 91 ATMs (as already mentioned, for two ATMs it was difficult to determine the point t_0).

Table 4. Descriptive statistics of the lengths of the development periods of the ATMs

mean	median	Standard deviation	coefficient of variation	minimum	maximum
49.03	49	18.21	37.13%	9	83

Figure 3 presents the distribution of the lengths of the development phases of the ATMs.

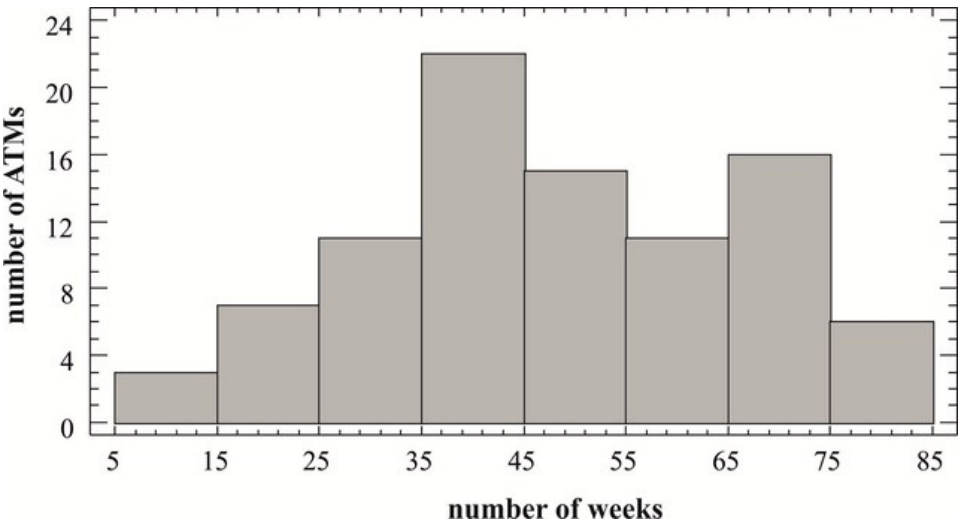


Figure 3. Distribution of the length of the development phase for weekly withdrawals from the ATMs (local trends approach)

The results of the analysis showed that the average length of the development phase of ATM is equal to 49 weeks, i.e. almost one calendar year. Variability of the length of the first phase is rather moderate and reaches the value of about 37%. The histogram suggests that the most common length of the development phase ranges between 36 and 48 weeks. The analysis showed that the development period was less than 12 weeks (3 months) only in one case. On the other hand, the group of ATMs for which the development period exceeds 72 weeks is quite large. Our analysis also showed that the results obtained for weekly data and seasonally adjusted daily data are very similar.

The results of alternative empirical approach by logistic regression are reported in the next subsection.

4.2. Saturation level estimated by logistic function

The application of the logistic function in the study on the phase of stabilization of ATMs withdrawals allows verifying and extending the previously obtained results (based on local trends approach). It also makes possible to determine the point in time in which the change of the dynamics takes place and estimate the level of saturation more accurately. On the other hand, some difficulties with determining the unique point in time in which the withdrawals stabilize may also occur. In this paper, we assumed that the point t_0 (stabilization point) is such a point in which The slope of tangent becomes not significant. On

the basis of the analysis of the figures we assumed that a dramatic drop of the slope (6-7 times) determines approximately the stabilization point.

Figure 4 contains results of fitting the logistic function to the time series describing the weekly withdrawals from ATM 5.

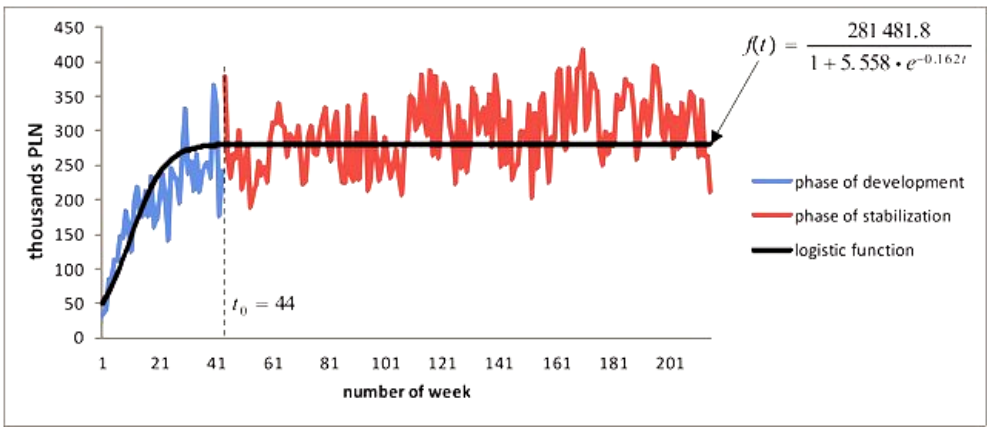


Figure 4. The results of fitting the logistic function to the weekly withdrawals from ATM 5

One can see (Figure 4) that the logistic function fits quite well to the dataset of weekly withdrawals. It does not capture the calendar effect or the seasonality. This, however, is not one of the main goals of our analysis.

Table 5 presents the results of the analysis performed for six ATMs (the estimated coefficients, the points of inflection and the stabilization points).

Table 5. The results of fitting the logistic function to the time series of weekly withdrawals from six selected ATMs

ATM's number	ATM 1	ATM 2	ATM 3	ATM 4	ATM 5	ATM 6
a	583 233.80	392 913.83	460 408.02	522 098.50	281 481.80	234 002.57
b	18.523	11.744	7.930	4.481	5.557	1.473
c	0.178	0.116	0.038	0.083	0.162	0.102
point of inflection	16.403	21.256	54.882	18.036	10.608	3.800
t_0	39	51	82	64	34	42

Similarly to the previous case, the analysis of the fit of the logistic function was performed for 91 out of 93 ATMs.

The distribution of the length of the development of ATMs and some basic descriptive statistics obtained after fitting the logistic function are presented in Table 6 and on Figure 5.

Table 6. Descriptive statistics for the length of the development phase

mean	median	Standard deviation	coefficient of variation	minimum	maximum
52.48	50	14.64	27.88%	11	84

Figure 5 presents the distribution of the length of the development phase.

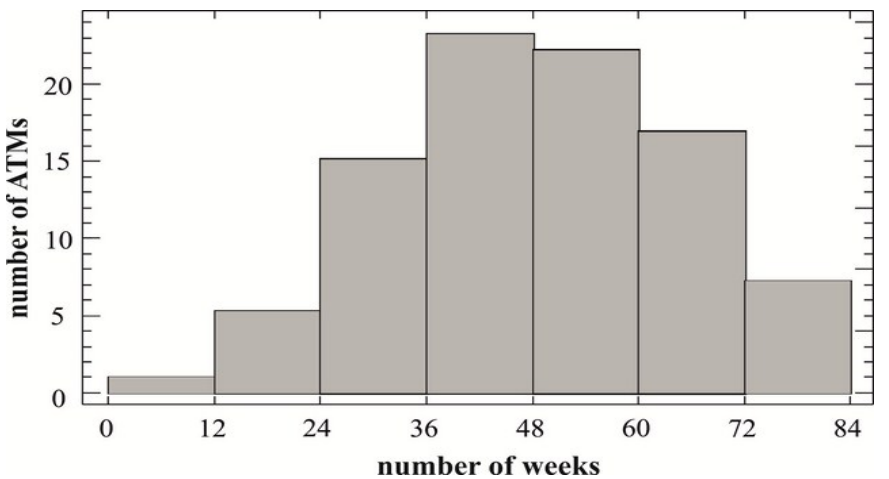


Figure 5. Distribution of the length of the development period for weekly withdrawals (logistic function approach)

The average length of the development phase of ATMs obtained after the application of the logistic function is approximately one year (52.48 weeks). Thus, it is slightly higher than the length obtained by the local trend approach. However, the variation of the length of the development period turned out to be smaller, as the coefficient of variation reached the level of 27.88%. We also found slight differences between the distributions of the length of the development period.

The analysis of the development of ATMs was carried out by two methods: the local trend approach and the application of the logistic function.

Therefore, in the next subsection we compare them.

4.3. Comparison of the empirical results

The results obtained after the application of both methods turned out to be very similar. Beside the length of the development phase of ATM, the issue of

withdrawals saturation is also very important, especially from the perspective of managing the ATMs network.

In tables 7-8 the results of the comparison of the lengths of the development periods (expressed in weeks) and saturation levels are presented. In case of the analysis performed via local trend approach the constant term in function $f_2(t) = a_2 t + b_2$ was interpreted as the saturation level.

Table 7. Comparison of the results of analysis of the lengths of the development periods

Method	ATM 1	ATM 2	ATM 3	ATM 4	ATM 5	ATM 6
Linear local trend approach	41	45	71	43	40	30
Logistic function	39	51	80	50	44	36

Table 8. Comparison of the results of analysis of the saturation levels

Method	ATM 1	ATM 2	ATM 3	ATM 4	ATM 5	ATM 6
Linear local trend approach	548 735	355 638	506 132	564 846	263 623	252 810
Logistic function	583 233	392 913	460 408	522 098	281 481	234 002

The comparison of the results of the length of the development period provides basis to claim that for 5 out of 6 selected ATMs the logistic function indicated longer length of the development period. However, the lengths of the development periods of the selected ATMs did not differ by more than about 10 weeks. Comparison of the results obtained for all ATMs leads to similar conclusion as in the case of six selected ATMs. Namely, the results of the analysis of the logistic function fitted to the dataset of 88 ATMs indicated a bit longer length of the development phase. However, in all cases the difference did not exceed 12 weeks, i.e. about 3 months. Thus, one may claim that the two methods lead to similar results in terms of estimating the length of the development phase, however, some differences are also noticeable. These differences may be a consequence of the choice of methodology of establishing the point of stabilization by application of the logistic function.

Similar results were obtained in case of determining the level of saturation. Although in this case it is difficult to indicate a method which gives higher values of the saturation level, the stabilization levels are similar. For the selected ATMs a relative difference between the saturation levels determined after the application of logistic function and the piecewise function built for local linear trend function did not exceed 11%. For all examined ATMs the average error was around 8%. The maximal value of the error did not exceed 15%.

The results of a comparison of the two methods of determining the length of the development phase and the levels of saturation of the sizes of withdrawals

provide a basis to claim that the values obtained during these analyzes are correct and may turn out to be helpful in managing the ATMs networks.

One of the main factors in development of ATMs can be location. This issue is addressed in the next section.

5. The impact of location on the development period of the ATMs

The results presented in previous section suggest that the average length of the development period of the ATMs is equal to approximately one year. However, one can see (comp. Fig. 3, Fig. 5) that there is relatively large group of ATMs for which the length of the development period differs significantly from the average. In addition, the detailed analysis conducted for six selected ATMs suggests that the type of location may have an impact on the length of the development phase. Therefore, we examined whether such characteristics as: province, size of population or type of location have an impact on the differences between the lengths of the development periods. In tables 9-11 the descriptive statistics for the lengths of the development periods for different location types are presented. In addition, the *p*-values obtained during an analysis of variance were calculated to check whether the differences among the lengths of the development periods are statistically significant.

Table 9. Descriptive statistics calculated for the lengths of the development periods in two provinces

province	Number of ATMs	mean	median	coefficient of variation	minimum	maximum	*p*-value
Małopolskie	69	48.45	51	38.59%	8	84	0.6049
Podkarpackie	24	50.70	48	33.49%	22	83	

The results of the analysis (comp. Table 9) proved that the average length of the development period of the ATMs located in Podkarpackie province is higher than the corresponding length in Małopolskie province. The difference, however, is not statistically significant.

Table 10. Descriptive statistics calculated for the lengths of the development periods among different population categories

Population category	Number of ATMs	mean	median	coefficient of variation	minimum	maximum	*p*-value
I	2	56.71	56	40.61%	40	73	0.2068
II	17	48.21	49	38.30%	15	74	
III	9	41.98	43	55.12%	9	79	
IV	16	57.87	57	27.27%	34	83	
V	49	47.41	47	36.70%	13	84	

The analysis of outcomes presented in Table 10 provides a basis to claim that the longest development period characterizes the cities with population at the level of 101-200 thousand. On the other hand, the development period is shortest in case of the cities with the population at the level of 51-100 thousand. Once again the differences in the average lengths of the development periods are not statistically significant (p-value=0.2068).

Table 11. Descriptive statistics calculated for the lengths of the development periods in respect to different location types

Location type	Number of ATMs	mean	median	coefficient of variation	minimum	maximum	*p*-value
bank branch	36	56.62	56	31.82%	22	84	0.0015
hypermarket	14	40.18	42	34.91%	9	67	
other	13	35.40	36	40.88%	13	57	
petrol station	6	53.47	46	29.92%	40	76	
shop	15	46.09	42	40.00%	15	70	
shopping center	9	54.04	57	28.97%	30	71	

The results of the analysis show that the type of location has a significant impact on the length of the development phase of the ATMs (cf. Tab. 11). The longest development period occurs in case of ATMs located at bank branches. This period was about 57 weeks long (on average). The shortest average period of development characterizes the ATMs located in “other” category. ATMs located in supermarkets also become "mature" relatively fast.

The analysis of the impact of location on the length of the development period was carried out for the results of the analysis of the local trends. However, a similar analysis performed for the results obtained from fitting a logistic function gave very similar results.

The results based on historical data allowed estimating the saturation level and the stabilization point. In the next section we will attempt to forecast these main characteristics of ATMs.

6. Forecasting of the length of the development period and the saturation level

The previous analysis was based on an examination of some properties of time series of historical data. It has been shown that for every ATM it is possible

to determine the point in time at which the size of withdrawals stabilizes. In addition, the level of the withdrawals during the stable period has been estimated. However, from the perspective of managing the ATM network, an important issue is to determine the parameters associated with the development of the ATM in case of those ATMs which are relatively "young". The ability to forecast the level of saturation and the length of the development period allows assessing whether the functioning of the ATM will be profitable in the future.

In case of the ATMs which have just started functioning, one of the methods of determining the parameters of the development can be based on a comparison of the similarity of time series. Namely, if one manages to show that for a group of ATMs with similar location type, the lengths of the development periods and the saturation levels are relatively close, then whenever starting a new ATM the predicted values of the development parameters should be equal to the average values of the parameters for the ATMs in the same location. A preliminary analysis showed that the ATMs which are characterized by the same type of location (province, size of population, surroundings) and which started at a similar time of the calendar year, are characterized by a very similar length of the development period. However, such an approach significantly reduces the possibility of forecasting because when a new ATM is started it is not certain whether it will be possible to find an already functioning ATM with similar properties, location, and a similar start date. Therefore, we were forced to look for other methods that would allow for estimation of the parameters associated with the development of ATMs.

In chapter 6 it was demonstrated that the logistic function can be successfully applied to describe the change in the size of the ATM's withdrawals in the time dimension. However, the estimation of the coefficients was based on the historical data which clearly excludes the possibility of formulating the forecasts.

In this part of the paper we investigated whether the estimated parameters of the logistic function based on the historical data covering only some initial period differ significantly from the estimates obtained from the full dataset. Such an approach would allow predicting the form of the logistic function of the ATM, in particular, forecast of such parameters as the length of the development period and the level of the saturation.

For this purpose, for all ATMs the coefficients of the logistic function were estimated based on the weekly data covering first 12, 16, 20, 24, 28.…. and 52 weeks. Then these results were compared with the results obtained from the estimation based on the full dataset. Figures 6 and 7 present the plots of the logistic functions obtained in the above-mentioned procedure based on the data for ATM 5.

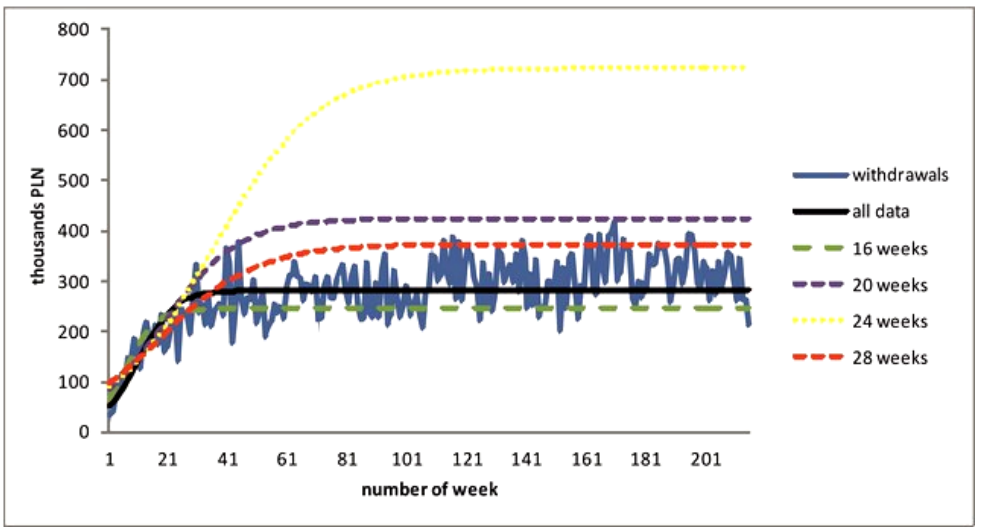


Figure 6. The logistic function based on data covering first: 16, 20, 24, 28 weeks and all data

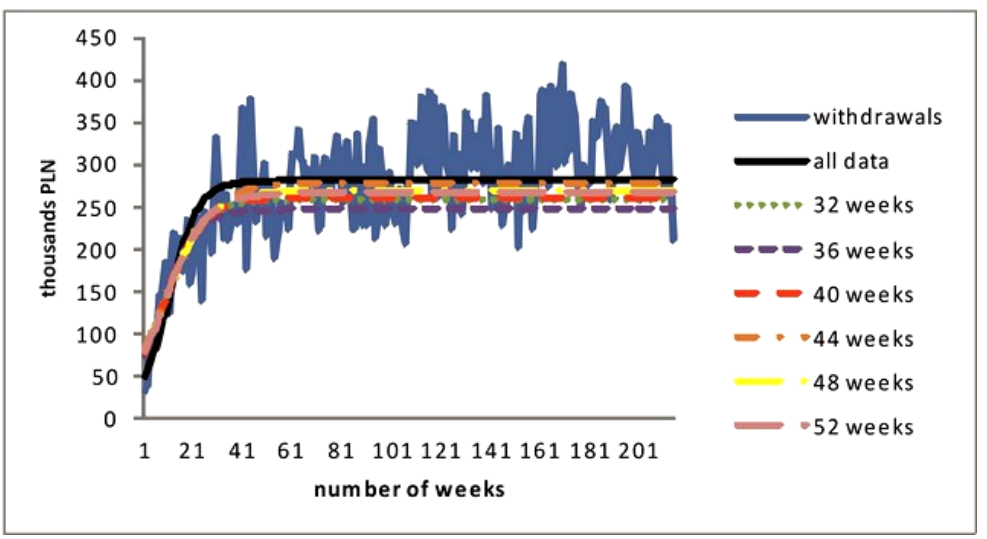


Figure 7. The logistic function based on data covering first: 32, 36, 40, 44, 48, 52 weeks and all data

The plots presented in Figure 6 provide a basis to claim that the shape of the logistic function estimated on the basis of the data from the first 20, 24 and 28 weeks differs significantly from the shape of the function estimated on the basis of the full dataset. However, in case of the function estimated on the basis of first 16 weeks the results are quite similar with the full-sample-based one. In addition,

the logistic functions estimated on the basis of the data covering first 32, 36, ..., 48, 52 weeks (Figure 7) are also very similar to the full-sample-based one. Therefore, one may note that in case of the ATM 5 the form of the logistic function could be successfully forecasted after 32 weeks. Moreover, the saturation level and the length of the development phase may also be forecasted after this period.

A similar analysis was performed for the time series describing the size of the weekly withdrawals from remaining ATMs. Table 12 contains the estimated values of the saturation level and the length of the development period based on the full dataset and the forecasts based on the data covering first: 12, 16, 20, ...,48, 52 weeks. The results are based on the analysis of the six selected ATMs.

Table 12. The results of forecasting the length of the development period and the saturation level

	ATM 1		ATM 2		ATM 3		ATM 4		ATM 5		ATM 6	
	t_0	Saturation level	t_0	Saturation level	t_0	Saturation level	t_0	Saturation level	t_0	Saturation level	t_0	Saturation level
all data	39	583233	51	392913	80	460408	50	522098	44	281481	36	234002
12 weeks	18	182040	65	6.36E+11	35	255461	75	1.10E+08	120	1.10E+08	100	1.10E+08
16 weeks	58	2576349	49	264524	75	2531592	70	238595	26	245446	52	128625
20 weeks	42	787107	50	427439	75	488065	47	426366	67	423595	40	424669
24 weeks	38	617636	52	423450	82	485498	51	521171	102	722244	38	219514
28 weeks	36	544126	54	380442	83	507194	45	567528	80	373126	34	222053
32 weeks	39	578750	58	399751	85	520531	49	578054	40	258825	37	233251
36 weeks	40	622082	60	426887	81	484344	54	585813	42	256859	38	217520
40 weeks	40	653230	58	470009	78	464126	60	606937	41	260624	37	231456
44 weeks	41	635873	57	510969	79	442209	65	752042	47	278595	35	217414
48 weeks	41	613171	56	481703	76	422527	67	822777	43	269402	33	211585
52 weeks	39	594272	60	463468	74	452394	63	803468	42	266658	37	219332

*Grey color – forecast error not greater than 10%. Slate gray color - forecast error not greater than 5%.

The analysis of the results presented in Table 12 leads to conclusion that there are two types of possible outcomes. In case of ATMs 1, 3, 5 and 6 satisfactory estimates of the unknown parameters were obtained on the basis of the data covering first 20-32 weeks. After increasing the sample size we obtained better forecasts. For ATMs 2 and 4 the prediction errors were smaller than 10% if the estimation was based on the data from the first 20-36 weeks. However, in these cases increasing the number of data points leads to prediction errors which were greater than 10%. This phenomenon may seem a bit strange, however, it can be explained quite easily. The reason of such a situation is the fact that some ATMs are more sensitive to the presence of the calendar effects. An unexpected rise or drop in the size of the withdrawals, which is due to some calendar effects (e.g. the beginning of Christmas or holiday), may cause an overestimation or underestimation of the level of saturation, and have a negative impact on the accuracy of the forecast of the length of the development phase.

However, after combining the findings from the analysis of the ATMs of the first and second group we can see that the optimal length of the period after which the forecast errors of the parameters do not exceed 10% is about 36 weeks.

The analysis performed for the remaining 85 ATMs showed that in all cases the prediction based on the data covering first 28-36 weeks leads to optimal predictability of the length of the development phase and the saturation level.

It should be underlined that the possibility of prediction of the discussed values after a period slightly longer than half a year gives the managers an ability to take appropriate decisions regarding the future of the ATMs.

On the other hand, one should not forget that the presented methodology is quite sensitive to the calendar effects. Therefore, each application of this procedure must be accompanied by an opinion of an appropriate expert specialized in the issue of cost-effectiveness of the ATMs network.

In the following final section we will summarize results and conclude.

7. Conclusions

One of the key elements of the ATM's network management is the ability to determine the time at which the ATM begins to operate in its regular rhythm and estimate the level of saturation of the withdrawals (average level of the withdrawals in the stable period). In this paper we have shown that for every ATM it is possible to determine the values of both these parameters on the basis of historical data.

In addition, we found that the ATMs characterized with similar type of location and similar start dates are also similar in terms of the development effects. This allows predicting the size of these values.

We have shown, however, that in order to forecasts the values of selected development indicators one can use a forecasting method based on the application of the logistic function estimated on the basis of the data covering first 28-36 weeks. As a consequence, after about half a year, we are able to estimate the time at which the withdrawals from ATM will stabilize reaching the level of saturation. Such information can turn out to be helpful in making the decisions on the future of the ATMs. After about half a year the managers can decide whether the ATM will be profitable in the future.

However, it should be emphasized that the presented methodology is sensitive to the calendar effects and thus it cannot be applied without second thought as it requires appropriate expert's supervision.

Acknowledgements

We would like to thank two anonymous referees for the valuable comments on an earlier version of the paper.

REFERENCES

AMROMIN, E., CHAKRAVORTI, S., (2007). Debit card and cash usage: a cross-country analysis. Technical report. Federal Reserve Bank of Chicago.

BOESCHOTEN, W. C., (1998). Cash management, payment patterns and the demand for money. De Economist, 146, 117–42.

BRENTNALL, A. R., CROWDER, M. J., HAND, D. J., (2008). A statistical model for the temporal pattern of individual automated teller machine withdrawals, Appl. Statist., 57, Part 1, 43–59.

BRENTNALL, A. R., CROWDER, M. J., HAND, D. J., (2010). Predicting the amount individuals withdraw at cash machines using a random effects multinomial model. Statistical Modelling, 10(2), 197–214.

CARLSEN, M., STORGAARD, P. E., (2010). Dankort payments as a timely indicator of retail sales in Denmark. Danmarks Nationalbank Working Papers, No. 66.

CLARKE, R., DAVIS, S., WATERSON, M., (1984). The profitability-concentration relation: market power or efficiency. The Journal of Industrial Economics, 32, 435–50.

CLEVELAND, W. S., DEVLIN, S. J., (1980). Calendar Effects in Monthly Time Series: Detection by Spectrum Analysis and Graphical Methods. Journal of the American Statistical Association, 75, 487–496.

ESTEVES, P. S., (2009). Are ATM/POS Data Relevant When Now casting Private Consumption? Banco de Portugal Working Paper, 25.

EVANOFF, D., FORTIER, L., (1988). Re-evaluation of the structure-conduct-performance paradigm in banking. Journal of Financial Service Research, 1, 277–94.

FINDLEY, D. F., MONSELL, B. C., (2009). Modeling Stock Trading Day Effects Under Flow Day-of-Week Effect Constraints. Journal of Official Statistics, 25(3), 415–430.

FINDLEY, D. F., MONSELL, B. C., BELL, W. R., OTTO, M. C., CHEN, B. C., (1998). New capabilities and Methods of the X-12-ARIMA seasonal adjustment program. Journal of Business and Economic Statistics, 16(2), 127–77.

FINDLEY, D. F., SOUKUP, R. J., (1999). On the Spectrum Diagnostics Used by X-12-ARIMA to Indicate the Presence of Trading Day Effects after Modeling or Adjustment. Proceedings of the American Statistical Association. Business and Statistics Section, 144–49.

FINDLEY, D. F., SOUKUP, R. J., (2000). Modeling and Model Selection for Moving Holidays. Proceedings of the American Statistical Association. Business and Economics Statistics Section, 102–07.

FINDLEY, D. F., SOUKUP, R. J., (2001). Detection and Modeling of Trading Day Effects. in ICES II: Proceedings of the Second International Conference on Economic Surveys, 743–53.

GALBRAITH, J. W., TKACZ, G., (2007). Electronic Transactions as High-Frequency Indicators of Economic Activity. Bank of Canada Working Paper, 58.

GURGUL, G., SUDER, M., (2012). Efekt kalendarza wypłat z bankomatów sieci Euronet, WSEI Research Journal, 8, 25–42.

HAND, D. J., BLUNT G., (2001). Prospecting for gems in credit card data. IMA Journal of Management Mathematics, 12, 173–200.

HOLDEN, K., EL-BANNANY, M., (2004). Investment in information technology systems and other determinants of bank profitability in the UK. Applied Financial Economics, 14, 361–5.

KONDO, K., (2003). How useful are ATMs for Japanese banks in the latter 1990s?, Review of Monetary and Financial Studies, 19, 43-54 (in Japanese).

KONDO, K., (2010). Do ATMs influence bank profitability in Japan? Applied Economics Letters, 17, 297–303.

KUFEL, T., (2010). Ekonometryczna analiza cykliczności procesów gospodarczych o wysokiej częstotliwości obserwowania, The Nicolaus Copernicus University Scientific Publishing House, Toruń.

LIU, L. M., (1980). Analysis of Time Series with Calendar Effects. Management Science, 26, 106–112.

LLOYD-WILLIAMS, D. M., MOLYNEUX, P., (1994). Market structure and performance in Spanish banking. Journal of Banking and Finance, 18, 433–43.

MCELROY, T. S., HOLLAND, S., (2005). A Nonparametric Test for Assessing Spectral Peaks. Research Report 2005-10. Statistical Research Division. U.S. Bureau of the Census, Washington D. C.

MOLYNEUX, P., FORBES, W., (1995). Market structure and performance in European banking. Applied Economics, 27, 155–9.

NACEUR, S. B., GOAIED, M., (2001). The determinants of the Tunisian deposit banks' performance. Applied Financial Economics, 11, 317–9.

RHOADES, S. A., RUTZ, R. D., (1982). Market power and firm risk: a test of the 'Quiet Life' hypothesis. Journal of Monetary Economics, 9, 73–85.

SIMUTIS, R., DILIJONAS, D., BASTINA, L., (2008). Cash demand forecasting for ATM using Neural Networks and support vector regression algorithms. 20th International Conference. EURO Mini Conference. "Continuous Optimization and Knowledge-Based Technologies" (EurOPT- 2008). Selected Papers. Vilnius May 20–23, 416–421.

SMIRLOCK, M., (1985). Evidence on the (non) relationship between concentration and profitability in banking. Journal of Money, Credit and Banking, 17, 69.

SNELLMAN, H., VIREN, M., (2009). ATM networks and cash usage. Applied Financial Economics, 19 (10), 841–851.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 467–486

SOME CONSIDERATIONS ON MEASURING THE PROGRESSIVE PRINCIPLE VIOLATIONS AND THE POTENTIAL EQUITY IN INCOME TAX SYSTEMS

Edyta Mazurek[1], **Achille Vernizzi**[2]

ABSTRACT

Kakwani and Lambert (1998) state three axioms which should be respected by an equitable tax system; then they propose a measurement system to evaluate at the same time the negative influences that axiom violations exert on the redistributive effect of taxes, and the potential equity of the tax system, which would be attained in absence of departures from equity. The authors calculate both the potential equity and the losses due to axiom violations, starting from the Kakwani (1977) progressivity index and the Kakwani (1984) decomposition of the redistributive effect. In this paper, we focus on the measure suggested by Kakwani and Lambert for the loss in potential equity, which is due to violations of the progressive principle: the authors' measure is based on the tax rate re-ranking index, calculated with respect to the ranking of pre-tax income distribution. The aim of the paper is to achieve a better understanding of what Kakwani and Lambert's measure actually represents, when it corrects the actual Kakwani progressivity index. The authors' measure is first of all considered under its analytical aspects and then observed in different simulated tax systems. In order to better highlight its behaviour, simulations compare Kakwani and Lambert's measure with the potential equity of a counterfactual tax distribution, which respects the progressive principle and preserves the overall tax revenue. The analysis presented in this article is performed by making use of the approach recently introduced by Pellegrino and Vernizzi (2013).

JEL Codes: C81, H23, H24

Key words: equity**,** personal income tax, progressive principle, redistribution, re-ranking.

[1] Department of Statistics, Wrocław University of Economics, Poland. E-mail: edyta.mazurek@ue.wroc.pl.
[2] Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Italy. E-mail: achille.vernizzi@unimi.it.

1. Introduction

Since their very beginning taxes have mainly been the way to gather by the state the resources necessary to ensure its proper functioning. Apart from performing the fiscal function, the state, by means of taxes, influences the 'fair' distribution of income, thus fulfilling the redistribution function. As Kakwani and Lambert (1998), thereafter KL, observe, the redistribution function through the tax system has to be performed respecting social equity principles; two basic commands of social equity are "*the equal treatment of equals and the appropriately unequal treatment of unequals*" (KL, p. 369). As Aronson, Johnson and Lambert (1994, p.262) stress, equity violations should be considered for given "*specifications of the utility/income relationship*".

The redistributive effect of the income tax system can be measured by the difference between the Gini coefficients for the pre-tax income distribution and the post-tax income one, respectively. The difference between these two indexes measures how the income tax system reduces inequality in income distribution. The potential equity in the tax system is a value of the redistributive effect which might be achieved if all inequities could be abolished, by a rearrangement of tax burdens which substantially maintains either the tax revenue or the tax schedule. Rearrangements are generally performed by means of tax credits, exemptions, allowances, income splitting or quotient. The assessment of the potential equity requires a definition of an equitable tax system.

KL propose an approach for measuring inequity in taxation. According to KL an equitable tax system should respect three axioms: (Axiom 1) tax should increase monotonically with respect to people's ability to pay; (Axiom 2) richer people should pay taxes at higher rates; (Axiom 3) no re-ranking should occur in people's living standards. In this paper we maintain the KL definition of equity in income taxation by means of the three axioms. Violations by an income tax system of each one of the three axioms provide the means to characterise the type of inequity present in an income tax system. A tax system is equitable if all axioms are satisfied.

Let X be the pre-tax income or living standard[1] T - the tax, and A - the tax-rate distribution, and Y - the disposable income. The three axioms ask that the ranking of T, A, and Y coincide with the ranking of X. It follows that, as KL suggest, the extent of each axiom violations can be measured by the Atkinson-Kakwani-Plotnick re-ranking index of each attribute T, A and Y, with respect to the X ordering.

By these re-ranking indexes, on the basis of the Kakwani (1977) progressivity index and of the Kakwani (1984) decomposition of the redistributive effect, KL evaluate the implicit or potential equity in the tax system, in the absence of inequities. In particular, by adding the tax re-ranking index to the Kakwani (1977)

[1] KL, page 372, use the term living standard for X, assuming that nominal incomes have been transformed by a proper equivalence scale.

progressivity index they evaluate the potential equity which the tax system would reach in the absence of Axiom 1 violations. Analogously, by adding the tax-rate re-ranking index to the Kakwani progressivity index, they estimate the potential equity which the tax system would reach in the absence of Axiom 2 violations, that is to say, in the absence of the progressive principle violations.

However, if the addition of the tax re-ranking index to the Kakwani progressivity index restores the progressivity which would be yielded without tax re-ranking, it is less simple to understand what happens when adding the tax-rate re-ranking index to the Kakwani progressivity index, as the next section illustrates.

The aim of this paper is first of all to contribute to a better understanding of what the KL measure of the potential equity implies. In so doing, we try to contribute to the definition of alternative measures for the potential equity, which the tax system would yield if violations in the progressive principle were eliminated. Our analysis is performed by making use of the approach recently introduced by Pellegrino and Vernizzi (2013).

In Section 2 the measure of the potential equity and the losses generated by axioms violations are presented, as suggested by KL. Next we present the potential equity for three different cases:

(*1*) both Axiom 1 and Axiom 2 are respected,
(*2*) Axiom 2 is violated, whilst Axiom 1 is respected,
(*3*) Axiom 1 is violated, which implies that Axiom 2 is violated too.

In fact, as KL observe,[1] "*a violation of minimal progression (Axiom 1) automatically entails a violation of the progressive principle*" (Axiom 2). Section 3 discusses KL's potential equity measure by analysing it at the level of income unit pairs' relations. This section also considers an alternative naïve measure for the potential equity. This measure is calculated by the Gini coefficient of the counterfactual tax distribution, which can be obtained by matching tax rates and pre-tax incomes, both ranked in non-decreasing order. Section 4 illustrates and completes the analytical considerations of previous sections by simulations performed on the income distribution of taxpayers from Wrocław (Poland). Section 5 concludes.

2. The loss due to axiom violations and the potential equity in the tax system

Let $x_1, x_2, ...x_K$ the pre-tax income levels of K income units, who are paying t_1, t_2, $...t_K$ in tax. Both incomes and taxes can be expressed either in nominal values or in equivalent values. Moreover, let $y_i = x_i - t_i$ and $a_i = t_i/x_i$ represent the disposable income and the tax rate, respectively, which result in unit i (i=1, 2,...,K), after having paid tax t_i.

[1] Kakwani and Lambert (1998), page 371.

Let G_X, G_T, G_A and G_Y be the Gini coefficients for attributes *X*, *T*, *A* and *Y*, respectively. Let $C_{Z|X}$ be the concentration coefficient for an attribute *Z* of income units (*Z*= *T*, *A*, *Y*) when the attribute *Z* is ranked by *X*, which refers to the pre-tax incomes lined up in ascending order. KL detect axiom violations via the three following Atkinson-Plotnick-Kakwani re-ranking indexes: $R_{T|X} = \left(G_T - C_{T|X}\right)$, $R_{A|X} = \left(G_A - C_{A|X}\right)$ and $R_{Y|X} = \left(G_Y - C_{Y|X}\right)$.[1] According to KL, if $R_{T|X} > 0$, Axiom 1 is violated; analogously, if $R_{A|X} - R_{T|X} > 0$ Axiom 2 is violated, and if $R_{Y|X} > 0$ it is Axiom 3 which is violated.[2]

KL use the three re-ranking indexes to evaluate the loss in the potential redistributive effect of the tax system. They represent the redistributive effect $RE = G_X - G_Y$, which, on the basis of the Kakwani decomposition, can be written as [3]

$$RE = \tau P - R_{Y|X}, \tag{1}$$

where τ is the ratio between the tax average - μ_T, and the disposable income average - μ_Y. *P* is the Kakwani progressivity index, [4] which is defined as follows:

$$P = C_{T|X} - G_X. \tag{2}$$

If no tax-re-ranking occurred, then $C_{T|X} = G_T$, and $P = G_T - G_X$. Therefore, KL can define $\tau R_{T|X}$ as the loss of redistributive effect due to tax re-ranking.[5] The potential equity after the correction for income and tax re-ranking can then be defined as:

$$PE_T = \tau P + \tau R_{T|X} = \tau\left(C_{T|X} - G_X + G_T - C_{T|X}\right) = \tau\left(G_T - G_X\right). \tag{3}$$

KL evaluate the loss due tax-rate re-ranking by the quantity $\tau\left(R_{A|X} - R_{T|X}\right)$. In analogy to (3), after having corrected also for the tax-rate re-ranking, KL define the potential equity as:

$$PE_A = \tau P + \tau R_{T|X} + \tau\left(R_{A|X} - R_{T|X}\right) = \tau\left(C_{T|X} + R_{A|X} - G_X\right). \tag{4}$$

In order to understand how the corrections yield the potential equity as per formulae (3) and (4), we now gather income unit pairs into three different groups, according to Axiom 1 and 2 violations.

---

[1] Being $G_Z \geq C_{Z|X}$, $R_{Z|X} \geq 0$.

[2] As KL observe (page 372), Axiom 3 can be violated only if Axiom 2 (and consequently Axiom 1) holds.

[3] See, e.g., Lambert (2001), pp. 238–242.

[4] Lambert (2001), ibidem. We observe that the progressivity index P does not contain any information on the incidence of taxation; τ is an indicator of the taxation incidence, which, conversely, does not contain any information on the progressivity: intuitively the redistributive effect is a function both of progressivity and of incidence.

[5] More details about equations (1) and (2) can be found in Appendix 3.

Group (*1*) includes income unit pairs presenting neither tax re-ranking nor tax rate re-ranking: for these pairs both Axiom 1 and Axiom 2 hold.

Group (*2*) includes income unit pairs presenting tax rate re-ranking but no tax re-ranking: for these pairs Axiom 1 holds, whilst Axiom 2 is violated.

Group (*3*) includes income unit pairs presenting both tax rate re-ranking and tax re-ranking: here, Axiom 1, and consequently Axiom 2, are both violated.

According to this classification, G_T, $C_{T|X}$, $R_{A|X}$ and G_X can split into three different components, $G_T^{(g)}$, $C_{T|X}^{(g)}$, $R_{A|X}^{(g)}$ and $G_X^{(g)}$ ($g = 1,2,3$), respectively, each related to one of the three different groups. This can be done by making use of the notation adopted by Pellegrino and Vernizzi (2013, page 3), and by a group selector function. Pellegrino and Vernizzi express the Gini coefficient in the following form:

$$G_Z = \frac{1}{2\mu_Z N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(z_i - z_j\right)p_i p_j I_{i-j}^{Z}, \quad \text{where} \quad I_{i-j}^{Z} = \begin{cases} 1: & z_i \geq z_j \\ -1: & z_i < z_j \end{cases}. \tag{5}$$

In (5) p_i and p_j are weights associated to z_i and z_j, respectively; $\sum_{i=1}^{K} p_i = N$; μ_Z is the average of *Z*, and I_{i-j}^{Z} is an indicator function.

When income units are lined up by ascending order of *X*, Pellegrino and Vernizzi write the concentration coefficient of attribute *Z* as:

$$C_{Z|X} = \frac{1}{2\mu_Z N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(z_i - z_j\right)p_i p_j I_{i-j}^{Z|X}, \qquad I_{i-j}^{Z|X} = \begin{cases} 1: & x_i > x_j \\ -1: & x_i < x_j \\ I_{i-j}^{Z}: & x_i = x_j \end{cases}. \tag{6}$$

Consequently, they formulate the re-ranking index of *Z* with respect to *X* as:

$$R_{Z|X} = \frac{1}{2\mu_Z N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(z_i - z_j\right)p_i p_j \left(I_{i-j}^{Z} - I_{i-j}^{Z|X}\right). \tag{7}$$

Let us introduce the indicator function $I_{i-j}^{(g)}$, which is 1 when the income unit pair $\{i, j\}$ is classified into group *g*, and it is zero otherwise: $I_{i-j}^{(g)}$ is then a group selector, which classifies each income unit pair into one of the three groups.[1] For each of the three different groups we can write:

$$G_T^{(g)} = \frac{1}{2\mu_T N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(t_i - t_j\right)p_i p_j I_{i-j}^{T} I_{i-j}^{(g)}, \tag{8}$$

[1] More details on the indicator function $I_{i-j}^{(g)}$ are in Appendix 3.

$$C_{T|X}^{(g)} = \frac{1}{2\mu_T N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(t_i - t_j\right) p_i p_j I_{i-j}^{T|X} I_{i-j}^{(g)}, \tag{9}$$

$$R_{A|X}^{(g)} = \frac{1}{2\mu_A N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(a_i - a_j\right) p_i p_j \left(I_{i-j}^{A} - I_{i-j}^{A|X}\right) I_{i-j}^{(g)}, \tag{10}$$

$$G_X^{(g)} = \frac{1}{2\mu_X N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(x_i - x_j\right) p_i p_j I_{i-j}^{X} I_{i-j}^{(g)}. \tag{11}$$

We observe that:[1]

$C_{T|X}^{(1)} = G_T^{(1)}$, $C_{T|X}^{(2)} = G_T^{(2)}$, $C_{T|X}^{(3)} = -G_T^{(3)}$;

$$C_{A|X}^{(1)} = G_A^{(1)},\ C_{A|X}^{(2)} = -G_A^{(2)},\ C_{A|X}^{(3)} = -G_A^{(3)}; \tag{12}$$

from which

$R_{T|X}^{(1)} = 0$, $R_{T|X}^{(2)} = 0$, $R_{T|X}^{(3)} = 2G_T^{(3)}$;

$$R_{A|X}^{(1)} = 0,\ R_{A|X}^{(2)} = 2G_A^{(2)},\ R_{A|X}^{(3)} = 2G_A^{(3)} \tag{13}$$

By making use of the expressions (8)-(11), we can split P, PE_T and PE_A defined by formulae (2), (3) and (4), into three components, $P^{(g)}$, $PE_T^{(g)}$ and $PE_A^{(g)}$ ($g = 1,2,3$), respectively, each related to one of the three groups. In addition, using the observations (12) we have the following relations:

- group (*1*)

$$P^{(1)} = G_T^{(1)} - G_X^{(1)},\ PE_T^{(1)} = \tau\, P^{(1)}, \tag{14}$$

$$PE_A^{(1)} = PE_T^{(1)}; \tag{15}$$

- group (*2*)

$$P^{(2)} = G_T^{(2)} - G_X^{(2)},\ PE_T^{(2)} = \tau\, P^{(2)} \tag{16}$$

$$PE_A^{(2)} = \tau \left(P^{(2)} + R_{A|X}^{(2)}\right) \tag{17}$$

- group (*3*)

$$P^{(3)} = C_{T|X}^{(3)} - G_X^{(3)},\ PE_T^{(3)} = \tau \left(P^{(3)} + R_{T|X}^{(3)}\right) = \tau \left(G_T^{(3)} - G_X^{(3)}\right) \tag{18}$$

$$PE_A^{(3)} = PE_T^{(3)} + \tau \left(R_{A|X}^{(3)} - R_{T|X}^{(3)}\right) = \tau \left(C_{T|X}^{(3)} + R_{A|X}^{(3)} - G_X^{(3)}\right) \tag{19}$$

[1] For more details on equations (12), see Appendix 3.

We can see that, according to the KL method, in case (*2*), where only tax-rate re-ranking is present, $R_{A|X}^{(2)}$ corrects just for the loss due to tax-rate re-ranking (expression 17). In case (*3*), $R_{A|X}^{(3)}$ corrects simultaneously both for tax and for tax-rate re-ranking (expression 19). Having in mind (15), KL's potential equity can then be written as:

$$PE_A = PE_A^{(1)} + PE_A^{(2)} + PE_A^{(3)} = PE_T^{(1)} + PE_A^{(2)} + PE_A^{(3)}. \tag{20}$$

However, expression (19) and, as a consequence, expression (20) are measures which are not strictly faithful to what KL state (page. 372) "*Axiom 2 is violated if the rankings by X and by A of income units pairs {i, j} for which Axiom 1 holds differ [.....]*". According to this statement, as all income units pairs, classified in case (*3*), present tax re-ranking and violate Axiom 2, they cannot be considered as violating Axiom 3, even if, as a consequence, they present tax-rate re-ranking too. In fact, the KL command specifies that tax-rate re-ranking should be considered only for income unit pairs classified in group (*2*).
Then, if we want to observe literally the KL command, the potential equity should be written as:[1]

$$PE_{A,T} = PE_T^{(1)} + PE_A^{(2)} + PE_T^{(3)}. \tag{21}$$

Even if Pellegrino and Vernizzi (page 242) specify that their new measure should be adopted "*If we want to observe literally the KL command*", they do not exclude adopting the original KL measure; they just stress that one should be aware that the KL measure "*does not involve only income units pairs for which Axiom 1 holds*". Actually, as it will appear clearer in the pursue, the tax rate re-ranking index often results to be greater that the tax re-ranking index, and this is coherent with the matter of fact that even after having eliminated tax re-ranking, tax-rate re-ranking can still persist.[2] So in this article we shall consider both the original KL measure and the one introduced by Pellegrino and Vernizzi.
In the next section we discuss how the potential equity measures act at the level of income pairs, in particular we will focus on (17) and (19).

3. The potential equity at the microscope

If we consider expression (3) for PE_T, after having added $R_{T|X}$ to $C_{T|X}$, one yields the Gini coefficient of the tax liability distribution. Conversely, for what concerns PE_A, as per expression (4), we need more considerations in order to

[1] Expression (21) can also be obtained by adding τP and S_2^*, defined in Pellegrino and Vernizzi (2013), at formula (9).

[2] Consider the following simple example: gross incomes {*X*: 1000, 2000, 3000, 5000}, taxes {*T*: 450, 400, 300, 200}. If we align taxes in ascending order and match them with incomes, the ratios {*A*: 200/1000, 300/2000, 400/3000, 450/5000} still remain in decreasing order, with respect to incomes.

interpret what one yields by adding $R_{A|X}$ to $C_{T|X}$. We shall now consider the effects of this addition at the level of the addends which constitute the sums in (9) and (10), which enter $PE_A^{(2)}$ and $PE_A^{(3)}$, given at (17) and (19), respectively.

The sum of $C_{T|X}$ and $R_{A|X}$, can be expressed as:

$$C_{T|X} + R_{A|X} = \frac{1}{2\mu_T N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} d_{i-j}\, p_i\, p_j\,, \tag{22}$$

having defined

$$d_{i-j} = \left(t_i - t_j\right) I_{i-j}^{T|X} + \left(a_i - a_j\right) \lambda \mu_X \left(I_{i-j}^{A} - I_{i-j}^{A|X}\right), \tag{23}$$

with $\lambda = \left(\mu_T / \mu_X\right) / \mu_A$.

In (23) the tax difference $\left(t_i - t_j\right) I_{i-j}^{T|X}$ is "corrected" by adding a component which is either 0 or $2\left|a_i - a_j\right| \lambda \mu_X$:[1] that is to say, when a tax rate pair does not respect the *X* ordering, the absolute difference of the two tax rates is transformed into a monetary value by the constant factor $\lambda \mu_X$.

In $PE_A^{(3)}$, the term $\left(t_i - t_j\right) I_{i-j}^{T|X}$ is negative; then d_{i-j} has to express the redistributive effect between income units *i* and *j*, which would be potentially yielded after compensation both for tax and for tax rate re-ranking. As a consequence, d_{i-j} should be not lower, at least, than the term $\left(t_i - t_j\right) I_{i-j}^{T} = \left|t_i - t_j\right|$, which expresses the potential equity, when the correction is limited to tax re-ranking. This implies that:

$$\left|a_i - a_j\right| \lambda \mu_X \geq \left|t_i - t_j\right|, \quad \text{i.e.} \quad \frac{\left|a_i - a_j\right|}{\mu_A} \geq \frac{\left|t_i - t_j\right|}{\mu_T}. \tag{24}$$

From the analysis reported in Appendix 2, there are more chances that inequality (24) is verified, than it is not. According to our simulations, reported in section 4, despite the fact that there is a not insignificant number of cases which do not verify (24), $PE_A^{(3)}$ happens to be always significantly greater than $PE_T^{(3)}$. KL "*found* $\left[R_{A|X} - R_{T|X}\right]$ *always to be non-negative in extensive simulations*".[2] The results of our simulations reinforce KL's findings. In fact, according to our experiments, not only the overall potential equity PE_A is greater than the corresponding PE_T: even when considering only pairs in case (*3*), where expression (24) is not necessarily satisfied, $PE_A^{(3)}$ is generally greater than $PE_T^{(3)}$.

---

[1] In case (*1*) $\left(a_i - a_j\right) \lambda \mu_X \left(I_{i-j}^{A} - I_{i-j}^{A|X}\right)$ is equal to 0; in cases (*2*) and (*3*) it is $2\left|a_i - a_j\right| \lambda \mu_X$, as $\left(I_{i-j}^{A} - I_{i-j}^{A|X}\right)$ is 2, when $\left(a_i - a_j\right)$ is positive, and it is -2 when the difference is negative.

[2] Kakwani and Lambert (1998, page 373).

We could try to measure the loss due to Axiom 2 violations by introducing a counterfactual tax distribution, which respects Axiom 2. This counterfactual tax distribution could be obtained by matching tax rates and pre-tax incomes, both aligned in ascending order. Tax rates are then rescaled in order to maintain the same tax revenue. However, we have to be aware that, in so doing, we would not strictly follow KL's command which asks that Axiom 2 violations should be *"confined to those income unit pairs for which Axiom 1 {i, j} holds"*. KL themselves do not fully respect their command as in measuring the extent of Axiom 2 violations by $\tau\left(R_{A|X} - R_{T|X}\right)$, as a matter of fact, they consider all unit pairs for which Axiom 2 does not hold.

In our opinion there are some reasons which could lead to considering all income unit pairs in evaluating the extent of Axiom 2 violations; in fact, after having matched both taxes and pre-tax incomes in ascending order, the tax rates derived from this matching do not necessarily become aligned in ascending order, as the example in footnote 11 illustrates. Consequently, the loss due to Axiom 2 violations can go further than the loss due to Axiom 1 violations.

If we denote this counterfactual tax distribution by T^{CF}, and the Gini index for the counterfactual tax distribution by G_T^{CF}, the loss due to Axiom 2 violations could then be measured by $\tau\left(G_T^{CF} - G_T\right)$, and the expression for the potential equity would become:

$$PE_A^{CF} = \tau\left(G_T^{CF} - G_X\right). \tag{25}$$

In the case of $G_T^{CF} < G_T$, further progressive counterfactual tax distributions could be generated by exploiting the progressivity reserve implicit in the tax system. In fact there are several counterfactual tax distributions which respect Axiom 2. For example, a more progressive counterfactual tax distribution could be generated by matching the pre-tax and income distribution and a modified tax-rate distribution which *a*) maintains the same tax rates in the upper queue of *A* distribution, and *b*) lowers tax rates in the lower queue of *A* distribution. Obviously, both distributions being in ascending order. Operations *a*) and *b*) could be calibrated in such a way that the tax revenue remains the same as that of *T*.

In the next section, we will evaluate the potential equity measures PE_T, PE_A, $PE_{A,T}$, and PE_A^{CF}, together with the incidence of cases which verify (24).

4. Simulation results

The measures of the potential equity, described in the previous section, were calculated for a Polish data set, by applying sixteen different hypothetical tax systems. The data come from the Lower-Silesian tax offices, 2001. The data set contains information on gross income for individual residents in the Municipality

of Wrocław (Poland). After deleting observations with non-positive gross income, the whole population consists of 37,080 individuals. For the analysis we used a random sample with size 10 000. The summary statistics for the sample of gross income distribution are: mean income = 18,980 PLN; standard deviation = 23,353 PLN; skewness = 13.29; kurtosis = 424.05. The Gini coefficient for the pre-tax income distribution is 0.45611.

The sixteen tax systems were constructed on the basis of four tax structures actually applied or widely discussed in Poland, from the simplest flat tax system to a more progressive tax system with four income brackets. In order to implement the "iniquity", within each tax structure net incomes were "disturbed" by introducing four different types of random errors (more details are reported in Appendix 1).

The resulting sixteen tax systems present *RE* indexes ranging from 0.141% to 3.584 %.

Table 1 reports basic indexes for the 16 tax systems, whereas Table 2 reports the potential equity measures for each tax system.

From Table 1, columns (3), (4) and (5), we can see that in each tax system the most unit pairs belong to group (*1*), where neither tax re-ranking nor tax rate re-ranking occurs. In four tax systems which derive from the Basic System 4, that is to say *T-S_4*, *T-S_8*, *T-S_12* and *T-S_16*, the percentage of income unit pairs belonging to group (*3*) is slightly greater than that of pairs belonging to group (*2*). Group (*2*) is much more crowded than group (*3*) in the remaining tax systems.

The percentage of pairs in group (*3*), which verify (24) (see Table 1, column 6), is never lower than 37.95% and greater than 83.6%. This percentage is changeable. However, if we consider column (14) in Table 1, we can observe that the ratios $\left(R_{A|X}^{(3)} / G_T^{(3)}\right)$ are stable at around 4 and a half. These empirical findings reinforce KL's simulation results: whenever $PE_A^{(3)}$ is greater than $PE_T^{(3)}$, a fortiori it is verified that $PE_A > PE_T$, as necessarily $PE_A^{(2)} \geq PE_T^{(2)}$.

Beyond any discussion about how measuring the loss in potential equity due to Axiom 2 violations, the role of this Axiom appears evident from Table 2. Considering the differences between PE_A (column 6) and PE_T (column 5), the marginal contribution to potential equity yielded in the absence of Axiom 2 violations is greater, and, in some cases, incomparably greater than that yielded by Axiom 1, which is given by the difference between PE_T and τP (column 4). This finding is not invalidated if we consider $PE_{A,T}$ (column 7), instead of PE_A, that is to say KL's measure, corrected as per formula (20). The counterfactual PE_A^{CF} potential equity gives different measures from PE_A, and in general the values estimated by the former are lower than those yielded by the latter; however, if one considers columns (6) and (8), distances present a much lower extent than those existing between column (6) and (5).

These results confirm the relevance of KL's contribution on distinguishing the different sources of inequity, and so, under this aspect, in spite of their conceptual and empirical differences, PE_A, PE_A^{CF} and $PE_{A,T}$ seem to give a coherent signal.

In any case, in our opinion, further research and discussion is still needed to conceive a measure which can solve some remaining ambiguities.

5. Concluding remarks

In this paper we have reconsidered the problem of measuring loss due to the progressive principle violations in a personal tax system. Our results confirm first of all the relevance of the contribution of Kakwani and Lambert's article (1998). The violations of minimal progression (KL's Axiom 1) and of the progressive principle (KL's Axiom 2) produce different effects and these effects have to be kept distinct. According to our simulations, on the whole, violations of the progressive principle appear to be even more relevant than those regarding minimal progression.

This note argues whether KL's measure of the progressive principle needs some refinements. The authors' measure implicitly transforms tax rate differences into tax differences by a factor which is constant, irrespective of income levels: this factor depends only on overall effects of a tax system, i.e. on the average tax rate and the average liability.

We air the idea of measuring the potential equity by introducing counterfactual tax distributions. As an example, we have simulated the behaviour of a naïve tax distribution, obtained by matching tax rates and incomes, both aligned in ascending order. Tax rates have been rescaled in order to maintain the same tax revenue. On the one hand this naïve measure produces different values from those obtained through KL's approach, on the other hand, it confirms the relevance of potential equity losses, due to Axiom 2 violations.

However, the counterfactual measures here outlined can be applied if one does not exclude income unit pairs which do not respect minimal progression. We observed that, even after having restored minimal progression by matching both the tax and the pre-tax income distribution in ascending order, the resulting counterfactual tax rates are not necessarily in ascending order too.

In conclusion, the discussion presented in this paper, which is based on the cornerstone represented by Kakwani and Lambert's (1998) article, intends to be an initial contribution towards a satisfying measure for the potential equity when the progressivity principle is violated.

Acknowledgement

We are grateful to an anonymous referee for the detailed comments and suggestions which helped us improve the final version of the article. Usual disclaimers apply.

REFERENCES

ARONSON, R. J., JOHNSON, P. J., LAMBERT, P. J., (1994). Redistributive Effect and Unequal Income Tax Treatment, *The Economic Journal*, 104, pp. 262–270.

KAKWANI, N. C., (1977). Measurement of tax progressivity: an international comparison, *Economic Journal*, 87, pp. 71–80.

KAKWANI, N. C., (1984). On the measurement of tax progressivity and redistributive effect of taxes with applications to horizontal and vertical equity, *Advances in Econometrics*, 3, pp. 149–168.

KAKWANI, N. C., LAMBERT, P. J., (1998). On measuring inequality in taxation: a new approach, *European Journal of Political Economics*, 14 (2), pp. 369–80. doi:http://dx.doi.org/10.1016/S0176-2680(98)00012-3.

LAMBERT, P. J., (2001). *The Distribution and Redistribution of Income*, Manchester University Press, Manchester and New York.

PELLEGRINO, S., VERNIZZI, A., (2013). On measuring violations of the progressive principle in income tax systems. *Empirical Economics*, 45 (1), pp. 239–245, doi: 10.1007/s00181-012-0613-1.

Table 1. Basic indexes: (1) neither tax re-ranking nor tax rate re-ranking, (2) tax rate re-ranking, without tax re-ranking, (3) tax and tax rate re-ranking. $G_X \times 100$=45.611; N=10,000

Tax system	λ	Percentage of pairs in groups			Percentage of pairs in group (3) for which $d_{i-j} \geq \lvert t_i - t_j \rvert$	$\frac{G_X^{(1)}}{G_X} \cdot 100$	$\frac{G_X^{(2)}}{G_X} \cdot 100$	$\frac{G_X^{(3)}}{G_X} \cdot 100$	$\frac{C_{T\|X}^{(1)}}{G_T^{(1)}} \cdot 100$	$\frac{C_{T\|X}^{(2)}}{G_T^{(2)}} \cdot 100$	$\frac{R_{A\|X}^{(2)}}{G_T^{(2)}}$	$\frac{C_{T\|X}^{(3)} \cdot 100}{-G_T^{(3)} \cdot 100}$	$\frac{R_{A\|X}^{(3)}}{G_T^{(3)}}$
	$(\mu_T / \mu_X) / \mu_A$	Group (1)	Group (2)	Group (3)									
(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**	**(14)**
T-S_1	1.114	59.933	35.185	4.882	81.257	60.363	39.156	0.481	31.389	15.385	0.321	-0.205	4.634
T-S_2	1.368	85.259	10.759	3.982	80.476	94.105	5.558	0.337	55.536	1.807	0.606	-0.148	4.705
T-S_3	1.310	84.297	11.692	4.012	79.782	91.795	7.853	0.351	52.102	2.818	0.449	-0.161	4.458
T-S_4	1.753	93.242	3.094	3.665	64.617	98.945	0.774	0.281	69.337	0.206	1.128	-0.093	4.712
T-S_5	1.118	59.920	34.066	6.013	83.596	61.818	37.435	0.747	32.910	14.160	0.419	-0.326	4.664
T-S_6	1.359	83.084	12.089	4.828	83.199	92.368	7.127	0.505	54.816	2.277	0.674	-0.234	4.672
T-S_7	1.304	82.120	13.041	4.840	82.374	89.989	9.490	0.521	51.466	3.297	0.525	-0.250	4.448
T-S_8	1.758	92.363	3.450	4.187	68.812	98.643	1.000	0.358	69.517	0.262	1.182	-0.140	4.674
T-S_9	1.116	59.655	38.360	1.985	64.081	60.586	39.333	0.081	29.521	17.159	0.105	-0.021	4.650
T-S_10	1.367	93.507	4.760	1.733	59.907	98.716	1.213	0.070	56.725	0.440	0.362	-0.015	4.720
T-S_11	1.306	93.159	5.113	1.728	59.224	98.198	1.731	0.071	53.902	0.701	0.250	-0.016	4.496
T-S_12	1.749	96.678	1.269	2.053	37.949	99.745	0.119	0.136	69.381	0.032	0.973	-0.010	4.730
T-S_13	1.118	60.025	34.802	5.174	82.180	61.906	37.558	0.536	32.330	14.677	0.353	-0.230	4.713
T-S_14	1.364	84.746	11.056	4.198	81.301	93.755	5.870	0.375	55.349	1.897	0.634	-0.168	4.698
T-S_15	1.304	83.707	12.095	4.198	80.416	91.412	8.206	0.382	51.744	2.939	0.466	-0.179	4.464
T-S_16	1.743	93.063	3.164	3.774	65.526	98.877	0.827	0.296	69.186	0.223	1.111	-0.104	4.630

Table 2. Potential equity measures. $G_X \times 100 = 45.611$

Tax system	$RE \cdot 100$	τ	$(\tau P/\mathrm{RE}) \cdot 100$	$(PE_T/\mathrm{RE}) \cdot 100$	$(PE_A/\mathrm{RE})\,100$	$(PE_{A,T}/\mathrm{RE})\,100$	$(PE_A^{CF}/\mathrm{RE})\,100$	$(R_{Y\|X}/\mathrm{RE})\,100$
(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
T-S_1	0.154	0.174	108.375	154.811	774.165	713.011	642.279	8.351
T-S_2	2.504	0.218	100.641	103.207	116.198	112.727	112.387	0.640
T-S_3	1.720	0.190	100.789	104.344	122.649	118.281	122.095	0.788
T-S_4	3.584	0.151	100.196	100.977	103.011	101.952	102.540	0.196
T-S_5	0.141	0.135	108.596	171.068	822.886	739.662	645.733	8.578
T-S_6	1.844	0.165	100.794	104.982	124.321	118.727	119.574	0.792
T-S_7	1.278	0.145	100.963	106.638	133.230	126.285	131.053	0.962
T-S_8	2.817	0.118	100.220	101.386	104.240	102.680	103.502	0.220
T-S_9	0.169	0.163	100.772	104.802	283.570	278.231	250.409	0.765
T-S_10	2.322	0.201	100.062	100.330	102.073	101.708	101.746	0.062
T-S_11	1.573	0.175	100.076	100.441	102.851	102.395	103.391	0.076
T-S_12	3.323	0.140	100.023	100.106	100.352	100.238	100.290	0.023
T-S_13	0.154	0.140	106.070	147.973	675.609	618.762	537.408	6.055
T-S_14	1.964	0.172	100.580	103.525	118.039	114.067	114.364	0.579
T-S_15	1.324	0.150	100.713	104.756	125.240	120.258	123.472	0.711
T-S_16	2.848	0.120	100.173	101.053	103.259	102.102	102.886	0.173

APPENDIXES

Appendix 1. The simulated tax systems

Four basic tax structures are hypothesised as follows:

BASIC SYSTEM 1. One 15 per cent tax rate is applied to all incomes. All taxpayers benefit from 556.02 PLN tax credit.

BASIC SYSTEM 2. A system with three income brackets: *i*) 19 per cent from 0 to 44,490 PLN, *ii*) 30 per cent from 44,490 to 85,528 PLN, *iii*) 40 per cent over 85,528 PLN. All taxpayers benefit from 586.85 PLN tax credit.

BASIC SYSTEM 3. A system with two income brackets: *i*) 18 per cent from 0 to 85,528 PLN, *ii*) 32 per cent over 85,528 PLN. All taxpayers benefit from 556.02 PLN tax credit.

BASIC SYSTEM 4. A system with four income brackets*: i)* 10 per cent from 0 to 20,000 PLN, *ii*) 20 per cent from 20,000 to 40,000 PLN, *iii*) 30 per cent from 40,000 to 90,000 PLN, iv) 40 per cent over 90,000 PLN. All taxpayers benefit from 500.00 PLN tax credit.

For each taxpayer, the tax $T(x_i)$ that results after the application of a basic tax system is then modified by a random factor, so that net income becomes $y_i = x_i - T(x_i) + z_i \cdot T(x_i)$; the factor z_i is drawn:

(a) from the uniform distributions:
($a1$) $Z \sim U(-0.2 \div 0.2)$, ($a2$) $Z \sim U(0 \div 0.4)$;
(b) from the normal distributions:
($b1$) $Z \sim N(0\,;0.0133)$, ($b2$) $Z \sim N(0\,;0.12)$;

Then, each basic system generates four sub-systems. When the normal distribution is applied, the random factor z_i is considered in absolute value; the programme did not allow incomes to become either negative or greater that $2x_i$.

In this way we receive the following sixteen hypothetical tax systems:

T-S_1: BASIC SYSTEM 1 modified by a random factor (*a*1)

T-S_2: BASIC SYSTEM 2 modified by a random factor (*a*1)

T-S_3: BASIC SYSTEM 3 modified by a random factor (*a*1)

T-S_4: BASIC SYSTEM 4 modified by a random factor (*a*1)

T-S_5: BASIC SYSTEM 1 modified by a random factor (*a*2)

T-S_6: BASIC SYSTEM 2 modified by a random factor (*a*2)

T-S_7: BASIC SYSTEM 3 modified by a random factor (*a*2)

T-S_8: BASIC SYSTEM 4 modified by a random factor (*a*2)

T-S_9: BASIC SYSTEM 1 modified by a random factor (*a*1)

T-S_10: BASIC SYSTEM 2 modified by a random factor (*b*1)

T-S_11: BASIC SYSTEM 3 modified by a random factor (*b*1)

T-S_12: BASIC SYSTEM 4 modified by a random factor (*b*1)

T-S_13: BASIC SYSTEM 1 modified by a random factor (*a*2)

T-S_14: BASIC SYSTEM 2 modified by a random factor (*b*2)

T-S_15: BASIC SYSTEM 3 modified by a random factor (*b*2)

T-S_16: BASIC SYSTEM 4 modified by a random factor (*b*2)

Appendix 2. On the sign of ($PE_A - PE_T$) for income unit pairs from group (*3*)

From expressions (18), (19) and (22), the pair {*i*, *j*} contributes to $PE_A^{(3)}$ at a greater extent than to $PE_T^{(3)}$ if

$$\left(t_i - t_j\right) I_{i-j}^{T|X} + \left(a_i - a_j\right)\lambda\mu_X \left(I_{i-j}^{A} - I_{i-j}^{A|X}\right) \geq \left(t_i - t_j\right) I_{i-j}^{T} . \qquad \text{(A.1)}$$

For the sake of simplicity and without any lack of generality, let us consider only the differences corresponding to incomes $x_i < x_j$; for group (*3*), when the difference $\left(x_i - x_j\right)$ is negative, $\left(t_i - t_j\right)$, $\left(a_i - a_j\right)$, $\left(I_{i-j}^{A} - I_{i-j}^{A|X}\right)$, and I_{i-j}^{T} are positive, whilst $I_{i-j}^{T|X}$ is negative. Then, for group (*3*), inequality (A.1) is verified if

$$\left(a_i - a_j\right)\lambda\mu_X \geq \left(t_i - t_j\right), \text{ for } x_i < x_j, a_i > a_j, \text{ and } t_i > t_j . \qquad \text{(A.2)}$$

As $t_i = a_i x_i$ and, analogously, $t_j = a_j x_j$, (A.2) can be rearranged as

$$a_i\left(\lambda\mu_X - x_i\right) \geq a_j\left(\lambda\mu_X - x_j\right). \qquad \text{(A.3)}$$

Being $a_i > a_j$, and $x_i < x_j$, we can conclude that, whenever $x_i < \lambda\mu_x$, strict inequality holds.

Income distributions are, in general, positive skew, so more than 50% of incomes are lower than μ_x and even more incomes are lower than $\lambda\mu_x$, as we expect that λ>1. To understand why λ is greater than 1, observe that if the tax rate schedule can be approximated by a strictly concave function of pre-tax incomes, (*i*) due to Jensen's inequality[1], it results in $\mu_A < \left[t\left(\mu_X\right)/\mu_X\right]$; moreover, (*ii*) if the

[1] See e.g. Lambert (2001), page 11.

tax distribution can be approximated by a strictly convex function, still due to Jensen's inequality[1], it results in $t(\mu_X) < \mu_T$. As in general both (*i*) and (*ii*) happen, a fortiori, we can expect that $\mu_A < (\mu_T / \mu_X)$, from which λ>1.

If $x_i > \lambda\mu_X$, strict inequality holds in (A.3), if $a_i(x_i - \lambda\mu_x) < a_j(x_j - \lambda\mu_x)$, or, which is the same, if

$$\frac{x_i - \lambda\mu_x}{x_j - \lambda\mu_x} < \frac{a_j}{a_i} \tag{A.4}$$

which can be either verified or not.

Our simulations (Table 1, column 6) confirm that in group (*3*) there are more income unit pairs which verify inequality (24) than those which do not: the share of pairs which verify the inequality is never less than 57.9 %.

Less immediate is interpreting the effect of λ. If one considers that keeping constant all the remaining components in (A.4), the left hand side of the inequality is a decreasing function of λ, it can be surprising to observe that the percentage of pairs in group (*3*), which verify (A.3), appears to be inversely related to λ. We can try to explain this by observing that the progressivity of a tax system does not act only on λ[2]: by its interactions with different sources of unfairness, it acts also on the actual tax rates and, consequently on the ratio (a_j / a_i). It is then difficult foreseeing the final outcomes concerning the inequality at (A.4). Moreover there is no reason to believe that the distribution of incomes lower than $\lambda\mu_X$ should remain equally distributed through the three groups.

As λ increases the percentage of violations of both Axiom 1 and Axiom 2 decreases as we can see from Table 1, columns (2), (4) and (5), progressivity should augment the theoretical distance of net incomes. Another not surprising result is that, for what concerns Axiom 2 violations (column 12), the relative correction, expressed by the ratio $\frac{R_{A|X}^{(2)}}{G_T^{(2)}}$, appears to be a direct function of λ.

We conclude observing that in column (14) the ratio $\frac{R_{A|X}^{(3)}}{G_T^{(3)}}$ assumes nearly constant values in all the simulated tax systems, ranging from 4.448 to 4.720.

[1] See e.g. Lambert (2001), page 223.

[2] From Table 1, column (2), we can see that λ is highest for tax systems *T-S_4*, *T-S_8*, *T-S_12* and *T-S_16*, which derive from the most progressive basic system, the fourth one. Conversely it is lowest for *T-S_1*, *T-S_5*, *T-S_9* and *T-S_13*, which derive from basic system 1, which has only one and quite low tax rate (15%).

Appendix 2. The decomposition of the redistributive effect

The Kakwani progressivity index, $P = C_{T|X} - G_X$, is based on the Jakobsson-Fellman and the Jakobsson-Kakwani theorems.[1]

First of all, from the Jakobsson-Fellman theorem it follows that if the derivative of the tax rate is non-negative, i.e. $a'(x) \geq 0$, then $P \geq 0$.

Let us now consider two different tax systems, $t_1(x)$ and $t_2(x)$, applied to a same income distribution. If the tax elasticities the two tax systems, $LP_1(x) = \left[t_1'(x) / a_1(x) \right]$ and $LP_2(x) = \left[t_2'(x) / a_2(x) \right]$, are such that $LP_1(x) \leq LP_2(x)$, then, from the Jakobsson-Kakwani theorem, it follows that $P_1 \leq P_2$.

The redistributive effect of a tax system, which is defined as the difference between $RE = G_X - G_Y$ [2], can be represented in terms of the Kakwani progressivity index and of the Atkinson-Plotnick-Kakwani re-ranking index, as per formula (1):

$$RE = \tau P - \left(G_Y - C_{Y|X} \right) = \tau P - R_{Y|X} \, .$$

By this expression one can immediately evaluate how much of the redistributive effect depends on the tax progressivity and how much is lost due to the re-ranking of incomes; it follows that τP represents the redistributive effect which would be achieved if no income re-ranking were introduced by taxes.[3]

If the tax ordering coincides with the pre-tax income ordering, we have that $C_{T|X} = G_T$, and, consequently, $P = G_T - G_X$. If the two orderings do not coincide, being $C_{T|X} < G_T$, it results that $P < G_T - G_X$, and, consequently, $\tau P < \tau (G_T - G_X)$: $\tau (G_T - G_X)$ is the potential redistributive effect which would be achieved if neither post-tax income re-ranking nor tax re-ranking occurred, with respect to the pre-tax income ordering.

We introduce the indicator function $I_{i-j}^{(g)}$ (*g=1, 2, 3*):

- $I_{i-j}^{(1)}$ is 1 when both the sign of $\left(t_i - t_j \right)$ and the sign of $\left(a_i - a_j \right)$ are not opposite to that of $\left(x_i - x_j \right)$, $I_{i-j}^{(1)}$ is 0 otherwise;

[1] Lambert (2001), pp.190-191, 199-200.
[2] Lambert (2001), pp.37-41.
[3] As already stressed in footnote 3, $R_{Y|X}$ is non-negative; it is zero when $G_Y = C_{Y|X}$.

- $I_{i-j}^{(2)}$ is 1 when the sign of $\left(t_i - t_j\right)$ is not opposite to that of $\left(x_i - x_j\right)$, and, conversely, the sign of $\left(a_i - a_j\right)$ is opposite to that of $\left(x_i - x_j\right)$; $I_{i-j}^{(2)}$ is 0 otherwise;
- $I_{i-j}^{(3)}$ is 1 when both the sign of $\left(t_i - t_j\right)$ and the sign of $\left(a_i - a_j\right)$ are opposite to that of $\left(a_i - a_j\right)$, $I_{i-j}^{(3)}$ is 0 otherwise.

Therefore, $I_{i-j}^{(g)}$ selects income unit pairs in relation to their behaviour in fulfilling Axiom 1 and Axiom 2.
In more general terms we can write as follows:

$$I_{i-j}^{(g)} = \begin{cases} 1 & \text{for income unit pairs belonging to group } (g) \\ 0 & \text{otherwise} \end{cases}, (g=1, 2, 3).$$

By applying the indicator function $I_{i-j}^{(g)}$, we can decompose $C_{T|X}$ as:

$$C_{T|X} = C_{T|X}^{(1)} + C_{T|X}^{(2)} + C_{T|X}^{(3)}$$

Having in mind expression (6), $C_{T|X}^{(1)}$ can be written as

$$C_{T|X}^{(1)} = \frac{1}{2\mu_T N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(t_i - t_j\right) p_i p_j I_{i-j}^{T|X} I_{i-j}^{(1)} \tag{A.5}$$

When $I_{i-j}^{(1)} = 1$, that is when income unit pairs are selected from group (*1*), the equality $I_{i-j}^{T|X} = I_{i-j}^{T}$ holds, so, as (A.5) yields

$$C_{T|X}^{(1)} = \frac{1}{2\mu_T N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(t_i - t_j\right) p_i p_j I_{i-j}^{T} I_{i-j}^{(1)} = G_T^{(1)}. \tag{A.6}$$

Analogously, also when $I_{i-j}^{(2)} = 1$, that is when income unit pairs are selected from group (*2*), it results that $C_{T|X}^{(2)} = G_T^{(2)}$, because, for all income unit pairs from group (*2*), we have $I_{i-j}^{T|X} = I_{i-j}^{T}$.

Differently for income unit pairs from group (*3*), which are selected by $I_{i-j}^{(3)} = 1$, we have $C_{T|X}^{(3)} = -G_T^{(3)}$. This is due to the fact that in the expression

$$C_{T|X}^{(3)} = \frac{1}{2\mu_T N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(t_i - t_j\right) p_i p_j I_{i-j}^{T|X} I_{i-j}^{(3)}, \tag{A.7}$$

the indicator function $I_{i-j}^{T|X}$ has sign opposite to I_{i-j}^{T}; using $I_{i-j}^{T|X} = (-1) \cdot I_{i-j}^{T}$, (A.7) yields

$$C_{T|X}^{(3)} = (-1) \frac{1}{2\mu_T N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \left(t_i - t_j\right) p_i p_j I_{i-j}^{T} I_{i-j}^{(3)} = -G_T^{(3)}. \tag{A.8}$$

From $C_{T|X}^{(1)} = G_T^{(1)}$, $C_{T|X}^{(2)} = G_T^{(2)}$, and $C_{T|X}^{(3)} = -G_T^{(3)}$, it follows that

$$R_{T|X}^{(1)} = G_T^{(1)} - C_{T|X}^{(1)} = 0\,,\ R_{T|X}^{(2)} = G_T^{(2)} - C_{T|X}^{(2)} = 0\,,\text{ and }\ R_{T|X}^{(3)} = G_T^{(3)} - C_{T|X}^{(3)} = 2G_T^{(3)}\,. \tag{A.9}$$

Also the re-ranking index $R_{A|X}$ can be decomposed into three components:

$$R_{A|X} = R_{A|X}^{(1)} + R_{A|X}^{(2)} + R_{A|X}^{(3)}\,, \tag{A.10}$$

with

$$R_{A|X}^{(1)} = 0,\ R_{A|X}^{(2)} = 2G_{A|X}^{(2)}\,,\text{ and }\ R_{A|X}^{(3)} = 2G_A^{(3)}\,. \tag{A.11}$$

The first element of the sum in equation (A.10), $R_{A|X}^{(1)} = G_A^{(1)} - C_{A|X}^{(1)}$, equals zero, because $C_{A|X}^{(1)} = G_A^{(1)}$. Using formulae (5) and (6) we can write:

$$G_A^{(1)} = \frac{1}{2\mu_A N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(a_i - a_j\right)p_i p_j I_{i-j}^{A} I_{i-j}^{(1)}\,, \tag{A.12}$$

$$C_{A|X}^{(1)} = \frac{1}{2\mu_A N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(a_i - a_j\right)p_i p_j I_{i-j}^{A|X} I_{i-j}^{(1)}\,. \tag{A.13}$$

When income unit pairs are selected from group (*1*), that is when $I_{i-j}^{(1)} = 1$, the differences $\left(a_i - a_j\right)$ and $\left(x_i - x_j\right)$ have the same sign and, consequently, in (A.12) and in (A.13) the equality $I_{i-j}^{A|X} = I_{i-j}^{A}$ holds, from which it follows that $C_{A|X}^{(1)} = G_A^{(1)}$.

Conversely, both $R_{A|X}^{(2)} = G_A^{(2)} - C_{A|X}^{(2)}$ and $R_{A|X}^{(3)} = G_A^{(3)} - C_{A|X}^{(3)}$ are greater than 0.

For income units pairs from group (*2*), that is when $I_{i-j}^{(2)} = 1$, having the differences $\left(a_i - a_j\right)$ and $\left(x_i - x_j\right)$ opposite sign, it follows that $I_{i-j}^{A|X} = (-1)\cdot I_{i-j}^{A}$. Therefore, we have

$$\begin{aligned} C_{A|X}^{(2)} &= \frac{1}{2\mu_A N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(a_i - a_j\right)p_i p_j I_{i-j}^{A|X} I_{i-j}^{(2)} \\ &= (-1)\frac{1}{2\mu_A N^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(a_i - a_j\right)p_i p_j I_{i-j}^{A} I_{i-j}^{(2)} = -G_A^{(2)}\,. \end{aligned} \tag{A.14}$$

As $C_{A|X}^{(2)}$ has opposite sign with respect to $G_A^{(2)}$, it is verified that $R_{A|X}^{(2)} = 2G_A^{(2)}$.

Analogously, for what concerns $R_{A|X}^{(3)}$, as $C_{A|X}^{(3)} = -G_A^{(3)}$, we can verify that $R_{A|X}^{(3)} = 2G_A^{(3)}$.

Equations (A.6)−(A.14) illustrate equivalences and relations reported in (12) and (13).

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 487–494

EMPIRICAL EVALUATION OF OCLUS AND GENRANDOMCLUST ALGORITHMS OF GENERATING CLUSTER STRUCTURES

Jerzy Korzeniewski[1]

ABSTRACT

The OCLUS algorithm and genRandomClust algorithm are newest proposals of generating multivariate cluster structures. Both methods have the capacity of controlling cluster overlap, but both do it quite differently. It seems that OCLUS method has much easier, intuitive interpretation. In order to verify this opinion a comparative assessment of both algorithms was carried out. For both methods multiple cluster structures were generated and each of them was grouped into the proper number of clusters using *k*-means. The groupings were assessed by means of divisions similarity index (modified Rand index) referring to the classification resulting from the generation. The comparison criterion is the behaviour of the overlap parameters of structures. The monotonicity of the overlap parameters with respect to the similarity index is assessed as well as the variability of the similarity index for the fixed value of overlap parameters. Moreover, particular attention is given to checking the existence of an overlap parameter limit for the classical grouping procedures as well as uniform nature of overlap control with respect to all clusters.

Key words: cluster analysis, cluster structure generation, OCLUS algorithm, genRandomClust algorithm.

1. Introduction

In cluster analysis it is not common that one can achieve far reaching theoretical results, therefore numerical simulations are a popular way of research. Therefore, generating cluster structures resembling real world data sets is a vital task. Across the decades researchers were trying to find better and better generating algorithms. Many proposals were made, however, none seems to be effective enough. In the early days, the inability to control the final degree of the overlap between clusters was the main obstacle. Kuiper and Fisher (1975) generated clusters from multivariate normal distributions, changing their means

[1] University of Lodz, Poland. E-mail: jurkor@wp.pl.

and covariance matrices but the final overlap was unknown. Gold and Hoffman (1976), also generated clusters from multivariate normal distributions adding observations drawn from distributions with different means. The resulting overlap between clusters was unknown. Blashfield (1976) generated clusters from possibly correlated distributions adding outlying observations from the uniform distribution. Also, the final overlap is unknown because some parameters are drawn at random. McIntyre and Bashfield (1980) changed the dispersion of distributions from which clusters were generated but the precise overlap remained undefined. Milligan (1985) generated very well separated distributions from truncated normal distributions adding outlying observations to blur the structures. The final overlap was unknown and some coordinates were distinguished with more stringent conditions imposed on them. Price (1993) controlled the overlap by means of trial and error, shifting the mean vectors until the desired overlap was achieved. However, this method is very time-consuming and not all values of possible overlaps can be assessed in this way. Atlas and Overall (1994) tried to keep overlap control, however, it remained unknown for the whole data set structure. Waller et al. (1999) fund a method which allows one to control overlap but only for low dimensional spaces and the control is visual.

2. OCLUS algorithm

Probably the first algorithm for generating cluster structures in which, in some cases, one has full control over the overlap between particular clusters as well as for the whole data set structure was the OCLUS algorithm (Steinley and Henson, 2005). Every cluster is generated from a predefined distribution and the overlap of each pair of two adjacent clusters also has to be predetermined for every dimension. If we simplify things slightly and assume that every pair of adjacent clusters has the same overlap $p_{j-1,j}$ then the overlap for the whole *i*-th dimension (*marginal overlap*) will be equal to

$$overlap_i = \frac{1}{k-1}\sum_{j=1}^{k-1} p_{j-1,j} \tag{1}$$

where *k* is the number of clusters. If we further assume that the dimensions are independent we can write the overall extent of overlap (*joint overlap*) as

$$overlap = \prod_{i=1}^{T} overlap_i \tag{2}$$

where *T* is the number of dimensions. Figure 1 presents an example of the scheme of generating three clusters with similar number of objects. The overlap between each two adjacent clusters is equal as well as the number of objects in each class.

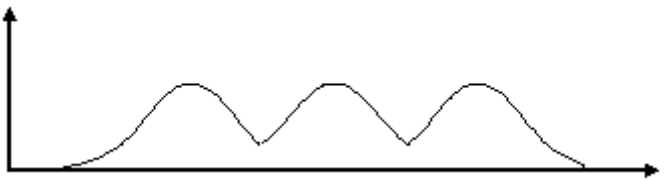


Figure 1. The marginal distribution for generating three clusters with roughly equal number of objects in each cluster.

Table 1. Distances between two adjacent clusters in dependence on overlap and the number T of dimensions.

overlap	T=2	T=4	T=6
0	6	6	6
.1	2	1.16	.82
.2	1.52	.86	.60
.3	1.2	.66	.46
.4	.96	.51	.35

Source: own calculations.

From formula (2) it follows that for the fixed marginal overlap the joint overlap tends to 0 with the growing number of dimensions. From formula (2) it also follows that for the fixed joint overlap the marginal overlap tends to 1 with the growing number of dimensions. We generated cluster structures according to the OCLUS algorithm assuming identical normal distributions with unit variance for every cluster and every dimension. The means of the normal distributions are determined by the overlap and the number of dimensions. The distances between the means of adjacent clusters are given in Table 1. The natural frontier of the overlap parameter for the classical grouping methods assigning every object to unique class is the number 0.5. If this value is exceeded then the clusters are indistinguishable from each other.

3. GenRandomClust algorithm

Quite different way of generating cluster structures can be found in the GenRandomClust algorithm (*Qiu and Joe*, 2006). The user is not obliged to determine the parameters of clusters distributions. Instead, one defines only (so the authors claim) one overlap parameter and the remaining parameters for every cluster (concerning shape and direction) are drawn at random from a defined interval.

When two clusters are generated from the normal distributions with means and covariance matrices, respectively, $\theta_i, \Sigma_i \quad i=1,2$, then their separation measures the authors use is:

$$J_{12}^{*} = \frac{a^T(\theta_2 - \theta_1) - q_{\alpha/2}\left(\sqrt{a^T\Sigma_1 a} + \sqrt{a^T\Sigma_2 a}\right)}{a^T(\theta_2 - \theta_1) + q_{\alpha/2}\left(\sqrt{a^T\Sigma_1 a} + \sqrt{a^T\Sigma_2 a}\right)} \tag{3}$$

where $\alpha \in (0; 0.5)$ is a parameter defining the percentage of outlying observations, $q_{\alpha/2}$

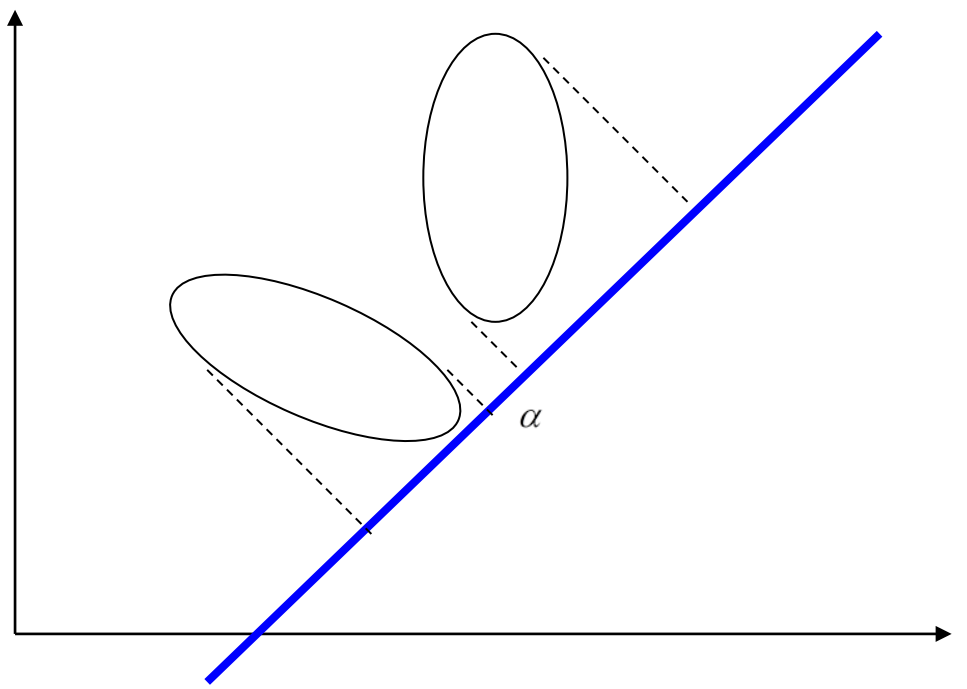


Figure 2. Graphical interpretation of the optimal projection and coefficient *α*.

is a quantile of order $\alpha/2$ of the marginal distribution (after projection see Fig. 2), vector ***a*** is the projection direction maximizing the value of measure J_{12}^{*}.

The measure thus defined takes into account the possibly different dispersions of objects in both clusters. However, the user has to specify the percentage *α.* of outlying observations. Typically, the parameter is assumed to be equal to 0.05. The algorithm works in the following way:

- cluster centers are placed in the vertices of a simplex and covariance matrices are randomly generated (by means of generating eigenvalues from a defined interval) so that the adjacent clusters would have the prespecified value of measure J_{12}^{*} - working as overlap measure.

- cluster centers and covariance matrices are randomly rotated by means of multiplication by an orthogonal matrix.
- objects are generated from the received distributions.

The algorithm thus defined does not allow one to generate clusters from other distributions type than the elliptical one. Neither there is a natural value of overlap parameter which would constitute a frontier between cluster structures meant only for traditional grouping methods and those which allow for objects belonging to more than one cluster. The authors provide function *genRandomClust* available in *R* language, which allows one to generate a vast palette of different cluster structure. However, it is probably impossible to use this function to generate big number of cluster structures with predefined means and covariance matrices. Theoretically, it is possible – the authors provide an algorithm to find the optimal projection direction, but we do not know if this algorithm has found that direction or got stuck in a local solution. Therefore, the value of the overlap parameter may be uncertain for the predefined normal clusters. The authors acknowledge also that only the separation between two closest clusters is controlled but other clusters are probably much better separated from the closest two.

4. Simulation experiment

In order to assess the quality of the cluster structures generated by both algorithms the following experiment was carried out. Data sets with similar cluster structures with respect to the number of clusters, number of dimensions, numbers of objects in each cluster and crossing the whole range of overlap parameter values characteristic for each algorithm were generated. Then, in order to get some numerical ground for the assessment, the objects were grouped into the proper number of clusters and into the number of clusters distorted by 1. When contaminating the number of clusters the number 3 was used instead of 2, number 3 instead of 4 and number 5 instead of 6. The grouping method was the traditional *k*-means with the random choice of starting points, repeated 50 times with the best grouping remembered (see *Steinley and Brusco*, 2007). The quality of the groupings was assessed by means of the corrected Rand index (*Hubert and Arabie*, 1985) of the similarity of the division resulting from the grouping with the one resulting from clusters generation.

For the Joe and Qiu algorithm, the function *genRandomClust* available in the R language was used in two variants: first, for equal cluster sizes (adding up to roughly 200 objects), second, for the size of each cluster drawn randomly from the set {20, …,120}. The value of the overlap control parameter, in the function named *sepval*, had eight values -0.35, -0.25, -0.15, -0.05, 0.05, 0.15, 0.25, 0.35. These values did not cover the whole range of possible variability of *sepval*, i.e. the interval (-1, 1), but the results were clearly predictable for the values above 0.35 and less interesting for the values below -0.35.

For the OCLUS algorithm, the clusters were generated (also in the two variants mentioned above) from the normal distribution with unit variance and the means on each dimension far away from the adjacent means by the values given in Table 1. The clusters were subsequently randomly and independently permutated on every dimension. The value of the overlap control parameter, in this case *joint overlap,* had five values 0.4, 0.3, 0.2, 0.10, 0. The two cases of equal and unequal cluster sizes were considered.

For both algorithms the possible numbers of clusters and dimensions were identical, number of clusters had three variants {2,4,6} and the number of dimensions also had three variants {2,4,6}. Every parameter set up was repeated 5 times for both algorithms.

5. Results and conclusions

The values of the Rand index presented in Tables 2 and 3 are arithmetic means from all 5 repetitions of the same set-ups. The results of the first case of equal cluster sizes are only presented as the second case of unequal cluster sizes gave very similar results for both algorithms. Apart from the mean values of the Rand index the dispersion of these values was also investigated. The examination showed that the standard deviation of the 5 values of the Rand index averaged through the whole table was similar for both algorithms (equal to about 0.1).

There are considerable differences between the two algorithms with respect to the number of clusters and the number of dimensions. The OCLUS algorithm generates structures which are more obscure for smaller number of clusters. There is also some dependence on the number of dimensions. The separability of clusters growing with the number of clusters may be the effect of the grouping method used – it is easier for the *k*-means method to hit (at least partially) the cluster centers when there are more of them than when there are just two clusters. The dependence on the number of dimensions is not considerable. There are no such correlations for the genRandomClust algorithm. All structures generated by this algorithm keep roughly the same quality for all numbers of clusters and dimensions considered.

The greatest difference between the two algorithms, however, was revealed for distorted number of clusters. It is very common that while examining the real world data sets we do not know the true number of clusters, therefore it is very vital to check the efficiency of cluster analysis methods in this case. The structures generated by the OCLUS algorithm did not suffer very much – there was a 5% or smaller drop (or rise) for the distorted number of clusters. The structures generated by the genRandomClust algorithm turned out to be very non-robust to these distortions – there always was a 20%-30% drop of the values of the Rand index. This feature seems to have some serious consequences. It corroborates the authors surmise of the lack of overlap control over the rest of all clusters apart from the two closest. The consequences depend on what kind of

efficiency investigation one will use the cluster structures for. If one wants to test a number of clusters index, then it is highly unadvisable to use the genRandomClust to generate cluster structures because the index may tend to be satisfied with the smaller number of well separated clusters. The interpretation is not clear when one wants to test a grouping method, however, since the case of the distorted number of clusters should be considered, it seems that the results may be seriously affected. A good example is that of the model based approaches to cluster analysis in which both the number of clusters assessment as well as the classification of objects are done simultaneously.

Table 2. Values of the Rand index for the *genRandomClust* algorithm

sepval		-.35	-.25	-.15	-.05	.05	.15	.25	.35
Number of clusters	Number of dimen.								
2	2	.29	.49	.68	.86	.91	.97	1.00	1.00
	4	.32	.53	.60	.84	.92	.96	1.00	1.00
	6	.30	.51	.62	.82	.89	.94	.99	1.00
4	2	.27	.47	.62	.79	.91	.98	.99	0.99
	4	.30	.41	.62	.80	.88	.94	.99	1.00
	6	.31	.46	.61	.83	.87	.95	.99	1.00
6	2	.30	.48	.69	.83	.91	.93	.99	1.00
	4	.28	.49	.65	.81	.88	.96	.98	1.00
	6	.29	.50	0.70	.81	.87	.92	1.00	1.00

Source: Own calculations.

Table 3. Values of the Rand index for the *OCLUS* algorithm

Overlap		.4	.3	.2	.1	0
Number of clusters	Number of dimen.					
2	2	.35	.43	.56	.81	1.00
	4	.21	.36	.46	.68	1.00
	6	.20	.35	.45	.62	1.00
4	2	.47	.62	.66	.83	1.00
	4	.68	.71	.85	.91	1.00
	6	.70	.74	.86	.92	1.00
6	2	.56	.61	.71	.76	.91
	4	.72	.77	.80	.85	.89
	6	.86	.89	.89	.95	1.00

Source: Own calculations.

REFERENCES

ATLAS, R., OVERALL J., (1994). Comparative Evaluation of Two Superior Stopping Rules for Hierarchical Cluster Analysis, Psychometrika, 59, 581–591.

BLASHFIELD, R. K., (1976). „Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods", Psychological Bulletin, 83, 377–388.

GOLD, E., HOFFMAN P., (1976). Flange Detection Cluster Analysis, Multivariate Behavioral Research, 11, 217–235.

HUBERT, L., ARABIE, P., (1985). Comparing Partitions, Journal of Classification 2.

KUIPER, F. K., FISHER, L., (1975). A Monte Carlo Comparison for Six Clustering Procedures, Biometrics, 31, 777–784.

MCINTYRE, R., BLASHFIELD, R., (1980). A Nearest-Centroid Technique for Evaluating the Minimum Variance Clustering Procedure, Multivariate Behavioral Research, 15, 225–238.

MILLIGAN, G., (1985). An Algorithm for Generating Artificial Test Clusters, Psychometrika, 50, 123–127.

PRICE, L., (1993). Identifying Cluster Overlap with NORMIX Population Membership Probabilities, Multivariate Behavioral Research, 28, 235–262.

QIU, W., JOE H., (2006). Generation of random clusters with specified degree of separation, Journal of Classification 23, 315–334.

STEINLEY, D., BRUSCO, M., (2007). Initializing k-means batch clustering: A critical evaluation of several techniques, Journal of Classification 24, 99–121.

STEINLEY, D., HENSON, R., (2005). OCLUS: An Analytic Method for Generating Clusters with Known Overlap, Journal of Classification 22, 221–250.

WALLER, N., UNDERHILL, J, KAISER, H., (1999). A Method for Generating Simulated Plasmodes and Artificial Test Clusters with User-Defined Shape, Size and Orientation. Multivariate Behavioral Research, 34, 123–142.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 495–506

A MODIFICATION OF THE PROBABILITY WEIGHTED METHOD OF MOMENTS AND ITS APPLICATION TO ESTIMATE THE FINANCIAL RETURN DISTRIBUTION TAIL

Marta Małecka[1], **Dorota Pekasiewicz**[2]

ABSTRACT

The issue of fitting the tail of the random variable with an unknown distribution plays a pivotal role in finance statistics since it paves the ground for estimation of high quantiles and subsequently offers risk measures. The parametric estimation of fat tails is based on the convergence to the generalized Pareto distribution (GPD). The paper explored the probability weighted method of moments (PWMM) applied to estimation of the GPD parameters. The focus of the study was on the tail index, commonly used to characterize the degree of tail fatness. The PWMM algorithm requires specification of the cdf estimate of the so-called excess variable and depends on the choice of the order of the probability weighted moments. We suggested modification of the PWMM method through the application of the level crossing empirical distribution function. Through the simulation study, the paper investigated statistical properties of the GPD shape parameter estimates with reference to the PWMM algorithm specification. The simulation experiment was designed with the use of fat-tailed distributions with parameters assessed on the basis of the empirical daily data for DJIA index. The results showed that, in comparison to the commonly used cdf formula, the choice of the level crossing empirical distribution function improved the statistical properties of the PWMM estimates. As a complementary analysis, the PWMM tail estimate of DJIA log returns distribution was presented.

Key words: PWMM, generalized Pareto distribution, tail estimate, distribution function estimate.

1. Introduction

In finance the estimates of the loss distribution tails are used for risk valuation. Using the extreme value theory, risk is assessed on the basis the extreme quantile of the loss distribution or the expected overshoot of a specified

[1] University of Lodz, Department of Statistical Methods. E-mail: marta.malecka@uni.lodz.pl.
[2] University of Lodz, Department of Statistical Methods. E-mail: pekasiewicz@uni.lodz.pl.

threshold. The estimation of the fat tails of the loss distribution is based on the empirical distribution of the auxiliary variable which is defined as the excess over the threshold and converges asymptotically to the generalized Pareto distribution (GPD). Since the shape parameter of the GPD distribution is used to characterize the degree of tail fatness, its estimation receives particular attention.

The paper explored the PWMM applied to estimation of the generalized Pareto distribution parameters (McNeil, Frey and Embrachts, 2005, Rasmussen, 2001). The PWMM requires specification of the cdf estimate of the excess variable and depends on the choice of the order of the probability weighted moments used for parameters estimation. In the paper we suggested the modification of the PWMM through the application of the level crossing empirical distribution function. In the simulation study we compared the GPD shape parameter estimate properties obtained with the use of the level crossing distribution function and the other empirical cdf proposed in the literature. The study involved the bias and variance of the estimates. The simulation experiment was designed with the use of fat-tailed distributions: t-Student, Pareto, log gamma and Burr with parameters assessed on the empirical daily data for DJIA index. As a complementary analysis, the PWMM tail estimate of DJIA log returns distribution was presented.

The paper is organized as follows. The next section sets the notation and introduces the probability weighted method of moments. Section 3 outlines the estimation procedure of the generalized Pareto distribution parameters and the tail of the underlying loss variable with the use of the PWMM. That section is followed by a comparative study of PWMM estimates of GPD parameters according to the chosen empirical distribution function and order of moments. At the end of section 4 we turn to empirical study for the daily DJIA data. The final section summarizes and concludes the article.

2. Probability weighted method of moments

The probability weighted method of moments is the procedure of estimation of the distribution parameters $\theta_1,...,\theta_s$ of the random variable X with cdf F that utilizes the idea of probability weighted moments of the distribution and their estimates. The probability weighted moment of the X distribution is defined as (if the relevant expectation exists):

$$M_{1,j,0} = E\left(XF^{j}(x)\right), \tag{1}$$

$$M_{1,0,k} = E\left(X\left(1-F(x)\right)^{k}\right), \tag{2}$$

for $j,k = 0, 1, ...$ (Rasmussen, 2001).

Suppose that we have a sequence of i.i.d. observations $X_1, X_2, ..., X_n$ from an unknown distribution *F*. The moments $M_{1,j,0}$ and $M_{1,0,k}$ estimates are sample probability weighted moments given by:

$$m_{1,j,0} = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}^{(n)}\left[F_n\left(x_{(i)}^{(n)}\right)\right]^j, \tag{3}$$

$$m_{1,0,k} = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}^{(n)}\left[1-F_n\left(x_{(i)}^{(n)}\right)\right]^k, \tag{4}$$

where $X_{(i)}^{(n)}$ are position statistics, F_n is the empirical cdf *F* and $j,k = 0, 1, ...$

Solving the system of *s* equations of the form:

$$\begin{cases} m_{1,j,0} = M_{1,j,0}, \\ m_{1,0,k} = M_{1,0,k}, \end{cases} \tag{5}$$

where $j = 0, 1, ..., s_1$, $k = 1, ..., s_2$ and $s_1 + s_2 = s$, gives the estimates of $\theta_1,...,\theta_s$.

Statistical properties of the estimates depend on the form of empirical distribution function. Various cdf estimates are presented in Hosting and Wallis (1987), Landwehr, Matalas and Wallis (1979), Seckin et al., (2010). We considered the commonly used cdf estimate:

$$F_n(x) = \frac{1}{n-1}\sum_{i=1}^{n}\left(I_{(-\infty,x\rangle}\left(x_{(i)}^{(n)}\right)-1\right), \tag{6}$$

and proposed the level crossing empirical distribution function (Huang and Brill, 1999):

$$F_n(x) = \sum_{i=1}^{n} w_{n,i} I_{(-\infty,x\rangle}\left(x_{(i)}^{(n)}\right), \tag{7}$$

where

$$w_{n,i} = \begin{cases} \dfrac{1}{2}\left[1-\dfrac{n-2}{\sqrt{n(n-1)}}\right] & \text{dla } i = 1, n, \\ \dfrac{1}{\sqrt{n(n-1)}} & \text{dla } i = 2, 3, ..., n-1. \end{cases} \tag{8}$$

and *I* is the indicator taking the value of one for $x_i \le x$ and 0 in the opposite case.

In the initial study also the popular formula $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x\rangle}\left(x_{(i)}^{(n)}\right)$ for the cdf was used. The comparative analysis including this cdf formula can be found in (Małecka and Pekasiewicz, 2013).

3. Estimation of the generalized Pareto distribution parameters and the distribution tail

Fitting the tail of the variable X distribution with the cdf F gives grounds for estimation of high quantiles and is therefore used for risk valuation. Of particular importance are distributions with fat tails which demonstrate financial data features. In such cases the estimate of the distribution function F for $x > u$ is given by (McNeil, Frey and Embrechts, 2005):

$$\hat{F}(x) = 1 - \frac{n_u}{n}\left(1 + \hat{\xi}\frac{x-u}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}}, \tag{9}$$

where u is the fixed threshold, n_u is the number of observations larger than u and $\hat{\xi}, \hat{\beta}$ are the estimates of the ξ, β parameters of the generalized Pareto distribution with the distribution function:

$$F_{\beta,\xi}(y) = 1 - \left(1 + \xi\frac{y}{\beta}\right)^{-\frac{1}{\xi}} \quad \text{for } \xi > 0 \text{ and } y \geq 0, \tag{10}$$

where β is the scale parameter and ξ is the shape parameter, whose reciprocal is commonly referred to as a tail index.

The GPD parameters are estimated from the sample $Y_1, Y_2, ..., Y_n$, where $Y_i = X_i - u$ and $Y_i > 0$ for $i = 1, 2, ..., n_u$.

If $\xi < 0.5$ parameters ξ and β of the GPD may be estimated by the method of moments. Then $\hat{\xi} = \frac{1}{2} - \frac{\overline{Y}^2}{2S^2}$ and $\hat{\beta} = \frac{\overline{Y}\left(\overline{Y}^2/S^2 + 1\right)}{2}$, where $\overline{Y}$, S^2 are subsequently the mean and the variance of n_u - element sample $Y_1, Y_2, ..., Y_n$.

Since for $0.5 \leq \xi < 1$ the variance of the GPD does not exist, the parameter estimates may only be obtained with the use of the expectation and the probability weighted moments. In such case, solving the system of equations (5) for the expected value ($M_{1,0,0}$) and the moment $M_{1,0,1}$ yields the estimates of the form:

$$\hat{\xi}^{mMWP} = 2 - \frac{\overline{Y}}{\overline{Y} - 2\alpha}, \tag{11}$$

$$\hat{\beta}^{mMWP} = \frac{2\alpha\overline{Y}}{\overline{Y} - 2\alpha}, \tag{12}$$

where α is the estimate of the moment $M_{1,0,1}$. Depending on the chosen form of the empirical distribution function we have subsequently:

- for the empirical distribution (6):

$$\alpha = \frac{1}{n_u}\sum_{i=1}^{n_u}\frac{n_u - i}{n_u - 1}Y_{(i)}^{(n_u)} \tag{13}$$

- for the proposed level crossing empirical distribution function (7):

$$\alpha = \frac{1}{n_u}\left(\frac{1}{2} + \frac{n_u - 2}{2\sqrt{n_u(n_u - 1)}}\right)Y_{(1)}^{(n_u)} + \frac{1}{n_u}\sum_{i=2}^{n_u-1}\left(\frac{1}{2} + \frac{n_u - 2i}{2\sqrt{n_u(n_u - 1)}}\right)Y_{(i)}^{(n_u)}, \tag{14}$$

where $Y_{(i)}^{(n)}$ is the *i*-th position statistic

For $\xi \geq 1$ the expectation and the variance of the GPD do not exist. The estimates of its parameters are based on moments $M_{1,0,k}$ and $M_{1,0,k+1}$ (if they exist) and, for the empirical distribution function (6), it holds that (Seckin et al., 2010)

$$\hat{\xi} = \frac{(k+1)^2\alpha_1 - (k+2)^2\alpha_2}{(k+1)\alpha_1 - (k+2)\alpha_2}, \tag{15}$$

$$\hat{\beta} = \frac{(k+2)(k+1)\alpha_1\alpha_2}{(k+1)\alpha_1 - (k+2)\alpha_2}, \tag{16}$$

where α_1 and α_2 take the following forms:

$$\alpha_1 = \frac{1}{n_u}\sum_{i=1}^{n_u}\frac{(n_u - i)(n_u - i - 1)\ldots(n_u - i - k + 1)}{(n_u - 1)(n_u - 2)\ldots(n_u - k)}Y_{(i)}^{(n_u)}, \tag{17}$$

$$\alpha_2 = \frac{1}{n_u}\sum_{i=1}^{n_u}\frac{(n_u - i)(n_u - i - 1)\ldots(n_u - i - k)}{(n_u - 1)(n_u - 2)\ldots(n_u - k - 1)}Y_{(i)}^{(n_u)}. \tag{18}$$

The application of the level crossing empirical distribution function leads to the sample moments:

$$m_{1,j,0} = \left(\frac{1}{2} - \frac{n-2}{2\sqrt{n(n-1)}}\right)^j \frac{X_{(1)}^{(n)}}{n} + \sum_{i=2}^{n-1}\frac{X_{(i)}^{(n)}}{n}\left(\frac{1}{2} - \frac{n-2i}{2\sqrt{n(n-1)}}\right)^j + \frac{1}{n}X_{(i)}^{(n)}, \tag{19}$$

$$m_{1,0,k} = \left(\frac{1}{2} + \frac{n-2}{2\sqrt{n(n-1)}}\right)^k \frac{X_{(1)}^{(n)}}{n} + \sum_{i=2}^{n-1}\frac{X_{(i)}^{(n)}}{n}\left(\frac{1}{2} + \frac{n-2i}{2\sqrt{n(n-1)}}\right)^k. \tag{20}$$

and

$$\hat{\xi} = \frac{(k+1)^2 m_{1,0,k} - (k+2)^2 m_{1,0,k+1}}{(k+1)m_{1,0,k} - (k+2)m_{1,0,k+1}}, \tag{21}$$

$$\hat{\beta} = \frac{(k+2)(k+1)m_{1,0,k}m_{1,0,k+1}}{(k+1)m_{1,0,k} - (k+2)m_{1,0,k+1}}. \tag{22}$$

4. Statistical properties of GPD parameter estimates for chosen return distributions

For its practical importance, in the simulation study we focused on parameter ξ estimation. Statistical properties of the estimates were examined in a simulation study. We considered four fat-tailed distributions: t-Student, Pareto, log gamma and Burr. The distribution parameters assumed for the purpose of the simulation experiment were assessed on the basis of the historical distribution of the daily returns from the stock market index. In order to reflect typical features of real financial series the DJIA index was chosen, which has a long tradition and is frequently used as a benchmark index in empirical analysis.

As the initial sample estimates showed that the ξ values usually fell into the interval (0,1) therefore we considered ξ from that range and presented the results for $\xi = 0.2, 0.4, 0.6, 0.8$. We set the distribution parameters on relevant levels to obtain such values. The threshold u was fixed at the level of the quantile of order 0.95 and the sample size was set to 1000, which guaranteed the number of observations over *u* large enough for the use of the asymptotic theorems.

In cases where values of ξ were below 0.5 we considered the use of the method of moments and PWMM estimates for weighted moments of orders 0 to 3. In cases with ξ between 0.5 and 1 it was only possible to use the expected value and further probability weighted moments of orders higher than 0. We used moments of orders 0 to 3. In the first step we investigated how the moment choice affects the estimate properties. The initial study was conducted with the use of the usual cdf formula (6)[1]. The results, obtained over 20000 replications[2], are presented in Tables 1 - 4. The simulation results showed that, while the estimate bias did not always exhibit a regular behavior, the variance influence resulted in the mean square error (MSE) increase for higher weighted moments. On the other hand, the ordinary methods of moments in most cases gave larger MSE than PWMM. Hence the results support the choice of the PWMM with expectation and the weighted moment of order 1, for ξ estimation, according to formulas (11) and (12).

[1] The simulations were afterwards repeated for the modified method with the crossing level distribution function, which gave the same conclusions about the choice of the order of weighted moments. The results are available from the authors upon request.

[2] In most cases 10000 is a sufficient number of repetitions in simulations (Białek, 2014), however, in our study we noticed differences in the second decimal point when diminishing the number of repetitions from 20000 to 10000.

Table 1. Statistical properties of parameter $\xi = 0.2$ estimates for different weighted moments

Distribution	Moments	$BIAS\left(\hat{\xi}\right)$	$D^2\left(\hat{\xi}\right)$	$MSE\left(\hat{\xi}\right)$
t-Student (5)	EX , D^2X	-0.139	0.041	0.060
	EX , $M_{1,0,1}$	-0.099	0.041	0.051
	$M_{1,0,1}$, $M_{1,0,2}$	-0.118	0.116	0.130
	$M_{1,0,2}$, $M_{1,0,3}$	-0.134	0.292	0.310
Pareto (1, 5) $\xi = 0.2$	EX , D^2X	-0.088	0.027	0.035
	EX , $M_{1,0,1}$	-0.027	0.029	0.030
	$M_{1,0,1}$, $M_{1,0,2}$	-0.020	0.093	0.094
	$M_{1,0,2}$, $M_{1,0,3}$	-0.023	0.257	0.254
Log gamma (2, 0.2, 0)	EX , D^2X	-0.084	0.026	0.033
	EX , $M_{1,0,1}$	-0.023	0.030	0.031
	$M_{1,0,1}$, $M_{1,0,2}$	-0.016	0.093	0.093
	$M_{1,0,2}$, $M_{1,0,3}$	-0.021	0.254	0.254
Burr (0.5, 10)	EX , D^2X ,	0.015	0.015	0.015
	EX , $M_{1,0,1}$	0.130	0.043	0.060
	$M_{1,0,1}$, $M_{1,0,2}$	0.188	0.113	0.149
	$M_{1,0,2}$, $M_{1,0,3}$	0.208	0.253	0.297

Table 2. Statistical properties of parameter $\xi = 0.4$ estimates for different weighted moments

Distribution	Moments	$BIAS\left(\hat{\xi}\right)$	$D^2\left(\hat{\xi}\right)$	$MSE\left(\hat{\xi}\right)$
t-Student (2.5)	EX , D^2X	-0.185	0.0522	0.086
	EX , $M_{1,0,1}$	-0.081	0.0398	0.046
	$M_{1,0,1}$, $M_{1,0,2}$	-0.076	0.092	0.099
	$M_{1,0,2}$, $M_{1,0,3}$	-0.086	0.236	0.243
Pareto (1, 2.5)	EX , D^2X	-0.168	0.044	0.072
	EX , $M_{1,0,1}$	-0.050	0.033	0.036
	$M_{1,0,1}$, $M_{1,0,2}$	-0.022	0.082	0.082
	$M_{1,0,2}$, $M_{1,0,3}$	-0.024	0.221	0.222
Log gamma (2, 0.4, 0)	EX , D^2X	-0.151	0.038	0.061
	EX , $M_{1,0,1}$	-0.022	0.031	0.031
	$M_{1,0,1}$, $M_{1,0,2}$	-0.011	0.081	0.081
	$M_{1,0,2}$, $M_{1,0,3}$	-0.009	0.2218	0.222

Table 2. Statistical properties of parameter $\xi = 0.4$ estimates for different weighted moments (cont.)

Distribution	Moments	$BIAS\left(\hat{\xi}\right)$	$D^2\left(\hat{\xi}\right)$	$MSE\left(\hat{\xi}\right)$
Burr (0.5, 5)	EX , D^2X	-0.111	0.025	0.037
	EX , $M_{1,0,1}$	0.055	0.031	0.034
	$M_{1,0,1}$, $M_{1,0,2}$	0.140	0.094	0.114
	$M_{1,0,2}$, $M_{1,0,3}$	0.166	0.227	0.255

Table 3. Statistical properties of parameter $\xi = 0.6$ estimates for different weighted moments

Distribution	Moments	$BIAS\left(\hat{\xi}\right)$	$D^2\left(\hat{\xi}\right)$	$MSE\left(\hat{\xi}\right)$
t-Student (5/3)	$M_{1,0,0}$, $M_{1,0,1}$	-0.111	0.046	0.058
	$M_{1,0,1}$, $M_{1,0,2}$	-0.057	0.080	0.083
	$M_{1,0,2}$, $M_{1,0,3}$	-0.059	0.205	0.208
Pareto (1, 5/3) $\xi = 0.2$	$M_{1,0,0}$, $M_{1,0,1}$	-0.099	0.042	0.052
	$M_{1,0,1}$, $M_{1,0,2}$	-0.034	0.077	0.078
	$M_{1,0,2}$, $M_{1,0,3}$	-0.033	0.197	0.198
Log gamma (2, 0.6, 0)	$M_{1,0,0}$, $M_{1,0,1}$	-0.056	0.034	0.037
	$M_{1,0,1}$, $M_{1,0,2}$	-0.032	0.074	0.075
	$M_{1,0,2}$, $M_{1,0,3}$	-0.039	0.189	0.191
Burr (0.5, 3.33)	$M_{1,0,0}$, $M_{1,0,1}$	-0.032	0.029	0.030
	$M_{1,0,1}$, $M_{1,0,2}$	0.086	0.075	0.082
	$M_{1,0,2}$, $M_{1,0,3}$	0.114	0.190	0.203

Table 4. Statistical properties of parameter $\xi = 0.8$ estimates for different weighted moments

Distribution	Moments	$BIAS\left(\hat{\xi}\right)$	$D^2\left(\hat{\xi}\right)$	$MSE\left(\hat{\xi}\right)$
t-Student (1.25)	$M_{1,0,0}$, $M_{1,0,1}$	-0.180	0.062	0.095
	$M_{1,0,1}$, $M_{1,0,2}$	-0.059	0.076	0.080
	$M_{1,0,2}$, $M_{1,0,3}$	-0.054	0.180	0.183
Pareto (1; 1.25)	$M_{1,0,0}$, $M_{1,0,1}$	-0.176	0.061	0.093
	$M_{1,0,1}$, $M_{1,0,2}$	-0.051	0.075	0.078
	$M_{1,0,2}$, $M_{1,0,3}$	-0.041	0.181	0.182

Table 4. Statistical properties of parameter $\xi = 0.8$ estimates for different weighted moments (cont.)

Distribution	Moments	$BIAS(\hat{\xi})$	$D^2(\hat{\xi})$	$MSE(\hat{\xi})$
Log gamma (2, 0.8,0)	$M_{1,0,0}$, $M_{1,0,1}$	-0.127	0.044	0.060
	$M_{1,0,1}$, $M_{1,0,2}$	0.041	0.075	0.077
	$M_{1,0,2}$, $M_{1,0,3}$	0.061	0.175	0.179
Burr (0.5, 2.5)	$M_{1,0,0}$, $M_{1,0,1}$	-0.136	0.044	0.063
	$M_{1,0,1}$, $M_{1,0,2}$	0.043	0.070	0.072
	$M_{1,0,2}$, $M_{1,0,3}$	0.075	0.172	0.178

In the next step we evaluated, over 20000 repetitions, the bias and the variance of estimates $\hat{\xi}_1, \hat{\xi}_2$ for empirical distribution functions respectively (6) and (7). According to the initial results we restricted the analysis to the use of PWMM with expectation and the weighted moment of order 1. In cases where $0 < \xi < 0.5$ we also considered the ordinary methods of moments, whose estimates were denoted $\hat{\xi}_0$. The results are presented in Tables 5 and 6 (ξ values below 0.5) and 7 and 8 (ξ values between 0.5 and 1). For the considered range of ξ values, in comparison to the commonly used empirical cdf formula (6), the choice of formulas given by (7) improved both the bias and the variance of parameter ξ estimates. The estimate bias for the formula $\hat{\xi}_1$ was mainly negative and reached values up to -0.18, which showed that this formula produced substantially lower estimates than the real parameter value. For the examined fat-tailed distributions better statistical properties were obtained for estimate $\hat{\xi}_2$, which used the suggested level crossing empirical distribution function defined by formula (7). Nearly in all cases both the bias and variance were lower for the modified method. Moreover, the PWMM estimates for both cdf formulas offered systematically lower bias and variance then the ordinary method of moments.

Table 5. Bias and variance of parameter $\xi = 0.2$ estimates for different cdf estimates

Distribution	$BIAS(\hat{\xi}_0)$	$BIAS(\hat{\xi}_1)$	$BIAS(\hat{\xi}_2)$	$D^2(\hat{\xi}_0)$	$D^2(\hat{\xi}_1)$	$D^2(\hat{\xi}_2)$
t-Student(5)	0.139	0.099	-0.064	0.041	0.041	0.032
Pareto(1, 5)	-0.088	-0.027	0.004	0.027	0.029	0.025
Log gamma(2, 0.2,0)	-0.084	-0.023	0.008	0.030	0.030	0.026
Burr(0.5, 10)	0.015	0.130	0.153	0.015	0.043	0.047

Table 6. Bias and variance of parameter $\xi = 0.4$ estimates for different cdf estimates

Distribution	$BIAS\left(\hat{\xi}_0\right)$	$BIAS\left(\hat{\xi}_1\right)$	$BIAS\left(\hat{\xi}_2\right)$	$D^2\left(\hat{\xi}_0\right)$	$D^2\left(\hat{\xi}_1\right)$	$D^2\left(\hat{\xi}_2\right)$
t-Student(2.5)	-0.185	-0.081	-0.045	0.052	0.040	0.031
Pareto(1, 2.5)	-0.168	-0.050	-0.015	0.044	0.033	0.027
Log gamma(2, 0.4,0)	-0.151	-0.022	-0.011	0.038	0.031	0.027
Burr(0.5, 5)	-0.111	-0.055	-0.084	0.024	0.031	0.032

Table 7. Bias and variance of parameter $\xi = 0.6$ estimates for different cdf estimates

Distribution	$BIAS\left(\hat{\xi}_1\right)$	$BIAS\left(\hat{\xi}_2\right)$	$D^2\left(\hat{\xi}_1\right)$	$D^2\left(\hat{\xi}_2\right)$
t-Student(5/3)	-0.111	-0.084	0.046	0.037
Pareto(1, 5/3)	-0.099	0.073	0.042	0.034
Log gamma(2, 0.6,0)	-0.056	-0.032	0.034	0.029
Burr(0.5, 3.33)	-0.032	-0.010	0.029	0.025

Table 8. Bias and variance of parameter $\xi = 0.8$ estimates for different cdf estimates

Distribution	$BIAS\left(\hat{\xi}_1\right)$	$BIAS\left(\hat{\xi}_2\right)$	$D^2\left(\hat{\xi}_1\right)$	$D^2\left(\hat{\xi}_2\right)$
t-Student(1.25)	-0.180	-0.169	0.062	0.056
Pareto(1, 1.25)	-0.176	-0.165	0.061	0.055
Log gamma(2, 0.8,0)	-0.127	-0.117	0.044	0.039
Burr(0.5, 2.5)	-0.136	-0.126	0.044	0.040

In finance, the estimates of the generalized Pareto distribution are used for fitting tails of fat-tailed return distributions. The functional form of the tail gives basics for calculating high quantiles of the distribution and in consequence offers risk measures. As an illustration of the simulation results we compared the use of the estimate in the form $\hat{\xi}_2$, characterized by the lowest bias and variance, with the usual $\hat{\xi}_1$ formula for the empirical time series. We applied the ordinary and the modified method to calculate the tail estimate from the sample of 1000 observations of DJIA dating back to March 2009. The estimated distribution function took the form:

$$\hat{F}_2(x) = 1 - 0.05\left(1 + 0.104\frac{x - 0.025}{0.008}\right)^{-\frac{1}{0.104}} \quad \text{for} \quad \hat{\xi}_2 \tag{23}$$

compared to

$$\hat{F}_1(x) = 1 - 0.05\left(1 + 0.052\frac{x - 0.025}{0.008}\right)^{-\frac{1}{0.052}} \quad \text{for} \quad \hat{\xi}_1 \tag{24}$$

In empirical analysis the value of parameter ξ, referred to as an extreme index, is often of particular interest, since it gives information about the degree of fatness of the distribution tail. In Figure 1 we presented the estimated values of this parameter for DJIA index, with the use of the three considered cdf estimates. The results were obtained

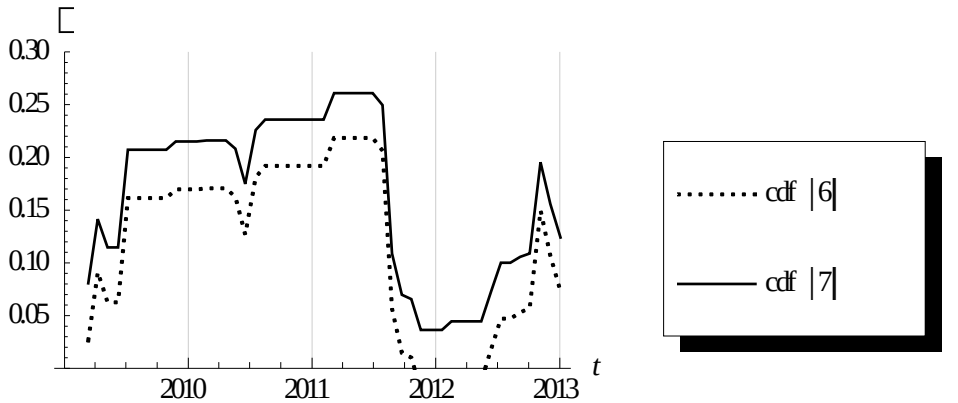


Figure 1. Parameter ξ estimates from the rolling estimation based on daily DJIA observations

from the rolling estimation with the window length of 1000 observations and the step of 20 observations. The dates included in the picture indicate the end point of the estimation window: starting sample ranged from March 2005 to March 2009 and the final sample range was from March 2009 to March 2013. The graphical presentation confirmed that parameter values had tendency to lie below the level of 1. The estimate $\hat{\xi}_1$ obtained with the use of the usual empirical distribution function in the form (6) lied at some distance from the proposed estimate $\hat{\xi}_2$, characterized by lower bias and variance, and were systematically below them.

5. Conclusion

The paper explored the probability weighted method of moments applied to estimation of the parameters of the generalized Pareto distribution. We suggested modification of the PWMM method through the application of the level crossing empirical distribution function. Statistical properties of the estimates of GPD parameter ξ, which characterizes the degree of tail fatness, were examined in a simulation study. We considered four fat-tailed distributions: t-Student, Pareto, log gamma and Burr and the typical parameter ξ values, which were assessed on the basis of the initial study for DJIA daily returns. The simulation exercise showed that suggested modification offered better statistical properties of the ξ estimate for ξ range (0,1) and the examined distributions. In case of other distributions, applicability of our results requires further research.

Since the PWMM method depends on the choice of the order of the probability weighted moments used for parameters estimation, the first step of the presented study concentrated on estimate statistical properties with relation to the chosen moment order. It was shown that there is a general increasing tendency in mean square error with increasing the moment order. This result supported the choice of the expectation and the weighted moment of order 1 for the considered ξ range and distributions.

Following the moment choice analysis we turned to examining statistical properties of parameter ξ estimates with reference to the cdf estimate required for PWMM parameter estimation. The simulation study results indicated that for the considered range of the parameter values it was possible to improve estimate statistical properties by the choice of the empirical distribution function given by formulas different that the commonly used cdf estimate.

REFERENCES

BIAŁEK J., (2014). Simulation Study of an Original Price Index Formula. *Communications in Statistics – Simulation and Computation* 43(2), 285–297.

HOSTING J. R. M., WALLIS J. R., (1987). Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics* 29, 339–349.

HUANG M. L., BRILL P. H., (1999). *A level crossing quantile estimation method*, Statistics & Probability Letters 45, 111–119.

LANDWEHR J. M., MATALAS N. C., WALLIS J. R., (1979). Probability Weighted Moments Compared with Some Traditional Techniques in Estimating Gumbel Parameters and Quantiles. *Water Resources Research* 15(5), 1055–1064.

MAŁECKA M., PEKASIEWICZ D., (2013). Application of probability weighted method of moments to estimation of financial return distribution tail in: *Proceedings of the 31st International Conference Mathematical Methods in Economics 2013* [ed. Hana Vojáčková], College of Polytechnics Jihlava, 569–574.

MCNEIL A. J., FREY R., EMBRACHTS P., (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princetown.

RASMUSSEN P. F., (2001). Generalized probability weighted moments: Application to the generalized Pareto distribution. *Water Resources Research* 37(6), 1745–1751.

SECKIN N., YURTAL R., HAKTAIR T., DOGAN A., (2010). Comparison of Probability Weighted Moments and Maximum Likelihood Methods Used in Flood Frequency Analysis for Ceyhan River Basin. *The Arabian Journal for Science and Engineering* 35, 49–69.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 507–516

THE DISTRIBUTION OF THE NUMBER OF CLAIMS IN THE THIRD PARTY'S MOTOR LIABILITY INSURANCE

Anna Szymańska[1]

ABSTRACT

In the automobile insurance tarification consists of two stages. The first step is to determine the net premiums on the basis of known risk factors, called *a priori* ratemaking. The second stage, called *a posteriori* ratemaking is to take into account the driver's claims history in the premium. Each step usually requires the actuary's selection of the theoretical distribution of the number of claims in the portfolio. The paper presents methods of consistency evaluation of the empirical and theoretical distributions used in motor insurance, illustrated with an example of data from different European markets.

Key words: distribution of the number of claims, civil liability motor insurance of vehicle owners.

1. Introduction

Calculation of the premium in property insurance is a complex process. The insurer at the time of setting the premium does not know the future costs of compensation, but they can be estimated on the base of the expected number and the amount of claims (Śliwiński, 2002, p.82). Estimation of the expected values of the random variable distributions representing the amount and number of claims requires the determination of theoretical distributions of the random variables.

The aim of the paper is to present methods of evaluation of the goodness-of-fit of theoretical distributions to empirical distributions of the number of claims in motor liability insurance. The data on the number of claims from the Polish market (from one of the insurance companies from Lodz, from the Lodz Region for the years 2000, 2001, 2002) were analyzed. Empirical distributions were chosen deliberately so as to assess functioning in the literature reselection methods of theoretical distribution of the number of claims.

[1] Department of Statistical Methods, University of Lodz. E-mail: szymanska@uni.lodz.pl.

2. Distributions of the number of claims used in car liability insurance

Let the random variable X represent the number of claims from individual policy or a policy portfolio. In motor liability insurance different theoretical distributions may be used to model the number of claims (Lemaire, 1995). In the following most commonly used distributions of random variables are presented.

Bernouli binominal distribution is described with the probability distribution function:

$$P(X=k)=\binom{n}{k}p^k q^{n-k}, \text{where } k=0,1,...,n \text{ and } \binom{n}{k}=\frac{n!}{k!(n-k)!}. \quad (1)$$

Poisson distribution is a distribution with the function of the probability defined by the formula:

$$P(X=k)=\exp(-\lambda)\frac{\lambda^k}{k!},\ k=0,1,.... \quad (2)$$

The random variable X has a negative binomial distribution (Polya) when its probability distribution function has the form:

$$P(X=k)=\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{k}. \quad (3)$$

If the random variable X has a Poisson distribution with parameter λ and the parameter λ has the inverse normal distribution, then the random variable X has a *Poisson-inverse normal* distribution (Denuit, Marechal, Pitrebois, Walhin, 2007, p.31). The probability function of Poisson-inverse normal distribution is given by:

$$P(X=k)=\sqrt{\frac{2\alpha}{\pi}}\exp\left(\alpha\sqrt{1-\theta}\right)\frac{(\alpha\theta/2)^k}{k!}K_{k-1/2}(\alpha)\quad k=0,1,...\ , \quad (4)$$

where $K_{k-1/2}(\alpha)$ is a modified third kind Bessel function (for positive and real arguments) in the form of:

$$K_{k-1/2}(\alpha)=\sqrt{\frac{\pi}{2\alpha}}\exp(-\alpha)\left(\sum_{i=0}^{k-1}\frac{(k-1+i)!}{(k-1-i)!i!}(2\alpha)^{-i}\right), k=1,2.... \quad (5)$$

Probability distribution function of the *Poisson-Poisson* distribution (*Neyman type A*) is given by:

$$P(X=k)=\exp(-\lambda_1)\frac{\lambda_2}{k!}^k\sum_{n=0}^{\infty}\frac{n^k}{n!}(\lambda_1\exp(-\lambda_2))^n,\quad k=0,1,2,... \quad (6)$$

Generalized Poisson-Pascal distribution is described by the moment generating function, which has the form:

$$M_X(t)=\exp\left\{\lambda\left[\frac{[1-\beta(t-1)]^{-\alpha}-(1+\beta)^{-\alpha}}{1-(1+\beta)^{-\alpha}}-1\right]\right\},\quad \alpha>-1,\lambda>0,\beta>0. \quad (7)$$

For $\alpha > 0$ the above distribution is called the *Poisson-Pascal* distribution. For $\alpha = 1$ the distribution is called *Polya-Aeppli* distribution. For $\alpha = -0{,}5$ it is called Poisson-inverse normal distribution.

Heilmann suggests the choice of the distribution of the number of claims in civil motor liability insurance depending on the relationship between the expected value and variance of the sample (Heilmann, 1988, p.46). Three distributions are considered: binomial, Poisson and negative binomial, which belong to the class (*a, b, 0*) (Otto, 2002, p.95). Wherein the family of distributions is called family distributions class *(a, b, 0),* where *a* and *b* are constants such that:

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k = 1,2,3,..., \tag{8}$$

where p_k is a function of the probability distribution of the discrete random variable. Treating the pair of parameters *(a, b*) as a point on the coordinates space of the areas corresponding to certain types of distributions can be designated. For $\{(a,b) : a = 0 \wedge b > 0 \wedge b \in R\}$ the distribution (*a, b, 0*) is the Poisson distribution, for $\{(a,b) : a \le 0 \wedge b > -a \wedge a \in R \wedge b = 1,2,3,...\}$ it is the binomial distribution, for $\{(a,b) : a \in (0,1) \wedge b > -a \wedge a,b \in R\}$ – the negative binomial distribution (Otto, 2004).

According to the paper (Panjer, Willmot, 1992, p.292) pre-selection of the theoretical distribution of the number of claims can be based on the calculated moments of the sample and the frequency coefficients.

Let $X_1, X_2, ..., X_n$ be an i.i.d. random sample. In case of aggregated data, where we know only the number of policies for the number of claims, simple sample moments usually are:

$$M_r = \frac{1}{n}\sum_{l=1}^{\infty} k^r N_k, \quad r = 1,2,..., \tag{9}$$

where N_k is the number X_i for which $X_i = k$, (k = 0,1,2, ...), $n = \sum_{k=0}^{\infty} N_k$ and $\overline{X} = \frac{1}{n}\sum_{k=0}^{\infty} kN_k$. The first three central moments of the sample are: $\overline{X} = M_1$; $S^2 = M_2 - M_1^2$; $K = M_3 - 3M_2M_1 + 2M_1^3$. Frequency coefficients are described by the following equation:

$$T(k) = (k+1)\frac{N_{k+1}}{N_k}, \quad k = 0,1,2,.... \tag{10}$$

Let:

$$T(k) = (a+b) + ak, \quad k = 0,1,2,... \tag{11}$$

be a function. When the function given by equation (11) is linear, whose slope coefficient:

- is zero and $\overline{X} = S^2$; then to describe the distribution of the number of claims the Poisson distribution is suggested;
- is negative and $\overline{X} > S^2$; then the binomial distribution can be assumed;
- is positive and $\overline{X} < S^2$; then the negative binomial distribution should be chosen.

When the function described by equation (11) grows faster than linearly, the skewness of the distribution should be considered. If the equation $K = 3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$ holds, the negative binomial distribution should model the number of claims well. If inequality $K < 3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$ holds, the generalized Poisson Pascal distribution, or its special case the Poisson-inverse normal distribution can be used to describe the distribution of the number of claims. If the inequality $K > 3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$ holds, the Neyman type A, Polya-Aeppli, Poisson-Pascal or negative binomial distributions are suitable for modeling the distribution of the number of claims.

3. Statistical methods of assessing the fitness of empirical and theoretical distributions

In the actuarial literature the tests which are most commonly used for evaluation of the relevance of the theoretical distribution to empirical data are: goodness-of-fit test χ^2 and test statistics based on λ – Kolmogorow (Domański, 1990, p.61). However, in case of distribution of the number of claims in car automobile insurance the number of classes is often not larger than four, which means that the number of degrees of freedom of the chi-squared test is too small. Additionally, most policies in the insurance portfolios are concentrated in the number zero class, which results in the distortion of the distribution. Portfolios are usually large, with the consequence that chi-squared test generally rejects the null hypothesis even though empirical data closely match theoretical distribution. In such cases, measures assessing the degree of fit of the theoretical distribution to empirical data may be found in statistical literature, such as the standard deviation of the differences in relative frequencies, the index of structures similarity, index of distribution similarity, ratio of the maximum difference of relative frequencies, ratio of the maximum difference of cumulative distribution functions (Kordos, 1973, p.115 -118).

Deviation of the differences in relative frequencies is a measure given by:

$$S_r = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(\gamma_i - \hat{\gamma}_i)^2}, \tag{12}$$

where: k - the number of classes, γ_i - empirical frequencies, $\hat{\gamma}_i$ - theoretical frequencies.

The measure is equal to zero in the case of full compliance of the empirical and theoretical distribution. The practice shows that the value $S_r \leq 0.005$ is an evidence of high compliance of schedules, if $0.005 \leq S_r < 0.01$ the compatibility of tested distributions is satisfactory and $S_r \geq 0.01$ shows significant deviations between the studied distributions.

The index of *structures similarity* is given by:

$$w_p = \sum_{i=1}^{k} \min(\gamma_i, \hat{\gamma}_i). \tag{13}$$

The index value is in the range [0,1]. The closer the value is to the unity, the more similar the structures of the studied distributions are.

Index of distribution similarity is determined by the equation:

$$W_p = 1 - \frac{1}{2}\sum_{i=1}^{k}|\gamma_i - \hat{\gamma}_i|. \tag{14}$$

Distribution similarity index is equal to 100% for fully compatible distribution. The distributions show high compatibility when $W_p \geq 0.97$. If $W_p < 0.95$ distributions show significant differences.

Ratio of the maximum difference of relative frequencies is given by the formula:

$$r_{\max} = \max_i |\gamma_i - \hat{\gamma}_i|. \tag{15}$$

This ratio is equal to zero for distributions fully compatible. If $r_{\max} < 0.02$, it is believed that the distributions are quite compatible.

Ratio of the maximum difference of cummulative distribution functions is given by the equation:

$$D_{\max} = \max_i |F_i - \hat{F}_i|. \tag{16}$$

where: $F_i = \sum_{j=1}^{i} \gamma_j$ - value of the empircial cummulative distribution function, $\hat{F}_i = \sum_{j=1}^{i} \hat{\gamma}_i$ - value of the theoretical cummulative distribution function. This ratio is equal to zero for fully consistent distributions.

In the analyses of the consistency of distributions of the number of claims with theoretical distributions, comparisons of distribution parameters such as mean, median, first and third quartile as well as measurement variation distributions can be used, in addition to the data indicators formulas (12) - (16). It is assumed that if the relative differences in ratings of all parameters between theoretical and empirical distribution do not exceed 5%, the distributions are fairly consistent.

4. Empirical example

This part of the paper presents the evaluation of the distribution of the number of claims in motor insurance for the actual sample data sets.

Table 1 presents data from the Polish market respectively in the years: 2000 (P1-option 1), 2001 (P2-option 2), 2002 (P3-option 3).

Table 1. Distribution of the number of claims

Number of claims	Number of policies			T_k		
	P1	P2	P3	P1	P2	P3
0	21 570	21 922	22 451	0.0313	0.0332	0.0284
1	676	730	638	0.0947	0.0712	0.0689
2	32	26	22	0.1875	0.2308	0.2727
3	2	2	2			
4	2	0	0			
sum	22 282	22 680	23 113			

Table 2 presents estimated parameters of the distributions presented in table 3.

Table 2. Parameters of the distribution of the number of claims

Parameters of the distribution	Distribution		
	P1	P2	P3
a	0.063334732	0.037932995	0.040548075
$\overline{X}$	0.033838973	0.034744268	0.029766798
S^2	0.037181825	0.036358973	0.031303615
K	0.04618218	0.039907237	0.034732823
$3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$	0.044527988	0.039738468	0.034535935

For each of the considered distributions the relationships $a > 0$, $\overline{X} < S^2$ and $K > 3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$ hold. Further analysis included the following theoretical distributions: negative-binomial, Poisson-inverse normal and Neyman type A.

Table 3. Measures of the degree of fit of theoretical distributions to distribution P1

Measure	Theoretical distribution		
	Negative-binominal	Poisson-inverse normal	Neyman A type
S_r	0.00029986	0.00047542	0.01311831
w_p	0.99940006	0.99941457	0.96913855
W_p	0.99999978	0.99999943	0.99956977
r_{max}	0.00000027	0.00000091	0.00085846
D_{max}	0.00032057	0.00054293	0.03061508

Table 4. Measures of the degree of fit of theoretical distributions to distribution P2

Measure	Theoretical distribution		
	Negative-binominal	Poisson-inverse normal	Neyman A type
S_r	0.00006487	0.00031944	0.01390219
w_p	0.99986817	0.99949218	0.96772665
W_p	0.99999999	0.99999974	0.99951682
r_{max}	0.00000001	0.00000031	0.00096509
D_{max}	0.00006366	0.00048811	0.03223130

Table 5. Measures of the degree of fit of theoretical distributions to distribution P3

Measure	Theoretical distribution		
	Negative-binominal	Poisson-inverse normal	Neyman A type
S_r	0.00007375	0.00026573	0.01198220
w_p	0.99985053	0.99962895	0.97220702
W_p	0.99999999	0.99999982	0.99964107
r_{max}	0.00000001	0.00000026	0.00071699
D_{max}	0.00007263	0.00034735	0.02774574

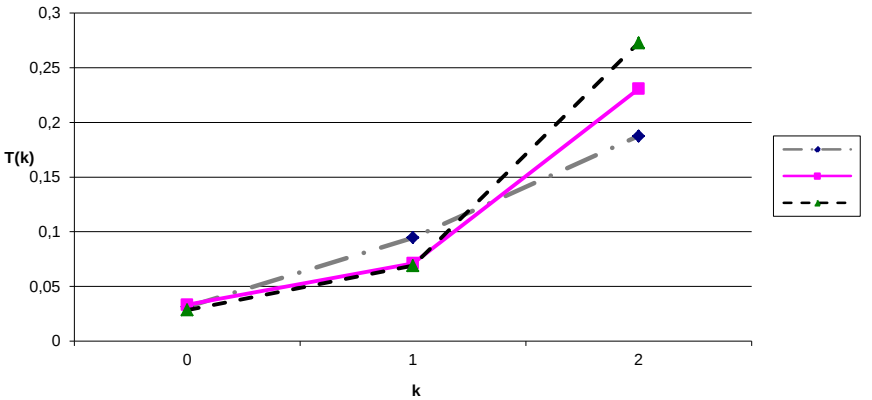


Figure 1. *T*(*k*) function of analyzed distributions

All empirical distributions received on the basis of data from the Polish market comply with the negative binomial distribution. This distribution gives the best fit to the empirical distribution of P2 (see tables 3-5), the worst to the distribution P1. The pre-fit empirical distribution of the negative binomial distribution can be assessed by the graph of *T*(*k*) function. The closer the graph of the function *T*(*k*) is to a straight line, the weaker fit it gives.

5. Conclusions

The conducted study shows that the methods of the selection of theoretical distribution of the number of claims proposed in the actuarial literature usually show the theoretical distribution that gives the best fit to the empirical data. Further studies are needed to confirm that for a linear $T(k)$ function with a positive slope coefficient and for $\overline{X} < S^2$ the negative binomial distribution should be chosen to describe the number of claims. In the case of the distributions considered, the results of the fit of the empirical distribution to the negative binomial distribution were worst the closer the function $T(k)$ was to a linear function.

In assessing the consistency of distributions, in most cases, the chi-square test cannot be used due to the nature of the data on the number of claims in motor liability insurance. Measures proposed in the paper offer a possibility to assess the goodness-of-fit of empirical and theoretical distributions.

It is not possible to unequivocally specify the type of theoretical distribution of the number of claims in motor liability insurance, although the distribution that gives the best fit is the negative binomial distribution. However, for each insurance market the distribution of the number of claims can be consistent with different theoretical distributions.

REFERENCES

DENUIT M., MARECHAL X., PITERBOIS S., WALHIN J., (2007). Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems, John Wiley & Sons, England.

DOMAŃSKI CZ., (1990). Testy statystyczne, PWE, Warszawa.

HEILMANN W. R., (1988). Fundamentals of Risk Theory, Verlag Versiecherungswirtschaft, Karlsruhe.

KORDOS J., (1973). Metody analizy i prognozowania rozkładów płac i dochodów ludności, PWE, Warszawa.

LEMAIRE J., (1995). Bonus-Malus Systems in Automobile Insurance, Kluwer, Boston.

OTTO W., (2002). Matematyka w ubezpieczeniach. Ubezpieczenia majątkowe, WNT, Warszawa.

PANJER H. H., WILLMOT G. E., (1992). Insurance risk models, Society of Actuaries, Schaumburg.

ŚLIWIŃSKI A., (2002). Ryzyko ubezpieczeniowe, taryfy - budowa i optymalizacja, poltext, Warszawa.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 517–520

BOOK REVIEW

Arthur A. Stone and Christopher Mackie (eds.) 2013, **Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience.** ***Panel on Measuring Subjective Well-Being in a Policy-Relevant Framework***; Committee on National Statistics, National Research Council. The National Academies Press. Washington, D.C.

by **Włodzimierz Okrasa**

One of the recently released by the US National Academies of Press reports brings the much awaited (by research community at large, and particularly by statisticians in public statistics offices) product of the CNSTAT-organized Panel on Measuring Subjective Well-Being in a Policy-Relevant Framework[1], entitled ***Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience*** (edited by Stone and Mackie). It accords with systematically growing interest from various sides, including policy analysts and policy makers and practitioners, in extending the measurement of objective aspects of economic achievements of countries and populations - with GDP as their most celebrated indicator - through accounting also for people's experiences, feelings and perceptions of how their lives are going. Or, in the Panel's nomenclature, subjective well-being (SWB) that "refers to how people *experience* and *evaluate* their lives and specific domains and activities in their lives" (p. 15).

It should be mentioned that the Panel's project was sponsored by the National Institute on Aging (NIA) of the National Institutes of Health, and by the UK Economic and Social Research Council (ESRC).

The recent impetus to reorient a long-term concern in social science literature about people's welfare (with primary focus on income) into interest in essentially psychological constructs, such as individual happiness and satisfaction (cf. Eid and Larsen, 2008) - sparked by the so-called *Stiglitz Report* (Stiglitz, Sen, and Fitoussi, 2009) - has already flourished in variety of studies conducted at the national and international levels (e.g. Gallup, World Values Survey, etc.), creating at the same time a demand for methodologically coherent and operationally useful framework that could underpin the measurement and evaluation works being conducted not only by research agencies but by official statistics government

[1] Members of the Panel: Arthur A. Stone (*Chair*, Stony Brook University); Norman M. Bradburn (University of Chicago); Laura L. Carstensen (Stanford University); Edward F. Diener (University of Illinois at Urbana-Champaign); Paul H. Dolan (London School of Economics and Political Science); Carol L. Graham (The Brookings Institution, Washington, DC); V. Joseph Hotz (Duke University); Daniel Kahneman (Princeton University), Arie Kapteyn (The RAND Corporation); Amanda Sacker (University of Essex, UK); Norbert Schwarz (University of Michigan); Justin Wolfers (University of Pennsylvania).

institutions as well, providing also a needed yardstick for comparison of the appropriate indices cross-culturally.

The complexity of the issue has already been recognized and addressed by the OECD in its well-received by the international research community and national statistical offices *Guidelines on the Measurement of Subjective Well-being* (2013). Envisaged as a part of the system of measures of societal progress, subjective well-being has clearly been recognized as only telling part of a person's story, which need to be complemented by information about more objective aspects of well-being to provide a full and rounded picture of how life is (p. 3). The CNSTAT Panel members admitted helpfulness of several conceptual and methodological advancements proposed in the *Guidelines*, especially as regards its recommendations for drafting appropriate questionnaires and scaling the respective categories in government surveys, as well as how to present the results correctly. Some prototype survey modules on subjective well-being developed as its part are currently under 'field' testing by several national and international agencies, including projects being organized under the auspice of the EUROSTAT with a SWB module proposed for the European Social Survey System (in EU member countries). However, indicators for 'quality of life and well-being' are proposed along with 'comprehensive environmental index' under the joint heading 'indicators complementing GDP' (De Smedt, 2014). Several national offices for statistics have already included SWB scales in their surveys (prominently, in the UK, France, Germany, Canada and others). The topic of measuring SWB was also discussed extensively during the last, 59th World Statistics Congress, held in Hong Kong (August 24-30, 2013) - mainly from data collection and measurement point of view (e.g. session on *Subjective Indicators and Their Role in Measuring Countries' Progress and Wellbeing: Definition, Construction and Analysis*), but also from analytical point of view (e.g. Okrasa, 2013).

On general, several sector-oriented international agencies have also been engaged, some for decades, in collecting data on well-being, but on limited scope only. For example, the World Health Organization's long-run system of quality of life indicators covers essentially evaluative, as opposed to experienced, aspect of SWB. On the contrary, the latter - experienced subjective well-being, ExWB - is of main focus of the Panel, concerned about people's emotions associated with a recent time period and the activities that occurred during that period. Reflecting pragmatic reason for its work - i.e., whether or not to continue collecting data on a SWB module within the American Time Use Survey (ATUS) conducted by the Bureau of Labor Statistics - the Panel's original interest was in 'hedonic' or experienced dimension of well-being ["directly related to the environment or context in which people live—the quality of their jobs, their immediate state of health, the nature of their commute to work, and the nature of their social networks—and is reflected in positive and negative affective states', Panel..., 2013, p.6]. Other aspects of SWB were deliberately neglected by the Panelists - in addition to evaluative also eudaimonic well-being - also due to their conviction that they are already more advanced for research purposes, in terms of both measurement and analytical approaches.

In spite of richness of the existing and newly created reports and manual-type publications devoted to subjective well-being, the above mentioned significance, usability and timeliness of the CNSTAT Panel's product remain unquestionable. Rather contrary, it may make the originality of its overall approach and the relevance of its recommendations even more salient and praiseworthy. Especially by experts engaged in collecting data in the US institutional context of public statistics, since agencies collecting (or planning to do it) self-reported well-being information are assumed to be the major target audience of the Panel's report, And by researchers interested in fundamental issues that arise at each stage of the process of measuring subjective well-being (from 'dimensionalization' through 'operationalization' to 'utilization'), as well as by all those who share the view that subjective well-being is "much more nuanced and difficult to measure than can be understood simply by asking people if they are happy" (p. 26).

The structure of the report reflects its conceptual and methodological orientations framed on distinguishing between three underlying dimensions or types of SWB - evaluative Well-Being, Experienced Well-Being (ExWB), and Eudaimonic Well-Being. While concentrating on the ExWB - on both positive and negative emotions - the Panelists recognize its mutual intertwining with each other. Especially with *eudaimonic* well-being, which refers to person's perceptions of meaningfulness, sense of purpose, and the value of life (p.3). Accordingly, the report addresses all the fundamental issues of conducting the appropriate research as a task of government statistical office. From recommendations on how to conceptualize it - starting from ideal type definition of all key notions through their operationalization for the data collection purposes, including how to measure it by phrasing the proper items of the questionnaires and combining them in scales - to recommendations on collecting good quality data while using some of the existing surveys.

Admitting that the American Time Use Survey's module (and method) called Day Reconstruction Method (DRM) is the most valuable source of information on ExSWB, the Panelists suggest to use multiple sources (different surveys and administrative data) to generate information on ExWB. Practically, each of the major surveys with subject-matter relevance could have included ExWB module - with primary candidates being: the Survey of Income and Program Participation (administered by the U.S. Census Bureau and providing data on income, program participation and care-giver links); Health and Retirement Study (health work transition, and health links); the American Housing Survey (Neighborhood Social Capital module - community amenities and social connectedness links); the Panel Study of Income Dynamics (care-giving arrangements, connectedness, and health links); the National Longitudinal Survey of Youth (patterns of obesity); and the National Health Interview Survey.

The high richness of data on SWB in terms of domains covered makes them more relevant for better informed policy, but only as far as it refers to targeted rather than universal policy objectives. The Panel expressed clear skepticism about possibility to construct and use an aggregate type of measure for assessing a population level quantity. Moreover, the panelists emphasize the necessity to complement data on ExWB by data on evaluative well-being due to a risk of

misinterpretation of people's behavior and subsequent misadvising on policy decisions.

Although few if any questions related to the SWB - though within a domain limited to ExWB - may be found unanswered by the Panel report, some important issues remain beyond the scope of its consideration. One relates to the relation between SWB and community well-being (CWB), Not only in the sense indicated as important in the latest version of the report by Stiglitz et al. (2012) - i.e., how strongly is SWB linked to community characteristics, connectedness and resilience (what panelists suggest as possible to be explored using the Neighborhood Social Capital Module of the American Housing Survey, p. 107). Given that simple aggregation of individual-level measures of SWB is not recommendable, even more interesting would seem to have systematized knowledge and recommendations on needed standard procedures allowing to account for residents' SWB while measuring community well-being as a quality in itself - as attempted, for example, in Canada (Stemeroff et al., 2009), Scotland (Project Report, 2003), and United Kingdom (Steuer and Marks, 2007) - to mention just few works along this line. Hopefully, this may inspire another group of first-rate experts to synthesize the conceptual and methodological advancements in a forward look type of report for directing new wave of research devoted to community well-being.

REFERENCES

DE SMEDT, M., (2014). Measuring Progress, GDP and Beyond Looking at Horizon 2020. Web-COSI Kick-off Meeting, Rome, 9 January.

OECD, (2013). Guidelines On Measuring Subjective Well-Being. OECD, Paris.

OKRASA, W., (2013). Spatial aspects of community well-being. Analyzing contextual and individual sources of variation using multilevel modeling. Presentation at the 59 World Statistics Congress, Hong Kong, August 24–30.

Panel on Measuring Subjective Well-Being in a Policy-Relevant Framework, (2013). The Subjective Well-Being Module of the American Time Use Survey: Assessment for Its Continuation. Committee on National Statistics; Division of Behavioral and Social Sciences and Education; National Research Council.

Project Report to the Scottish Executive, (2003). Community Well-Being. An Exploration of Themes and Issues. Scottish Executive. Edinburgh.

STEMEROFF, M., RICHARDSON, D., T., WLODARCZYK, (2009). Context and Application of Community Well-Being. NWMO SR-2009-04. AECOM Canada Ltd.

STEUER, N., MARKS, N., (2007). Local Well-Being: Can we measure it? http://www.communities.gov.uk/publications/localgovernment/nationalindicatorsupdate.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, p. 521

Graham Kalton, PhD - Laureate of the Jerzy-Splawa Neyman Medal

Professor Graham Kalton, a Senior Vice President and Chairman of the Board of Directors of Westat, has had a distinguished career as a researcher and teacher in the area of survey statistics and methodology. He is a co-founder of the Joint Program in Survey Methodology at the University of Maryland, where he holds the title of research professor. Prior to joining Westat in 1992, he was a research scientist in the Survey Research Center, a professor of biostatistics, and a professor of statistics at the University of Michigan, where he served a term as chairman of the Department of Biostatistics. Dr. Kalton is co-author with Claus Moser of the second edition of *Survey Methods in Social Investigation*, published in 1971, a widely-acclaimed text that covers all aspects of survey research. He has published many papers on survey research, particularly in the areas of sampling methods for rare populations, weighting and imputation, and panel surveys. He served on the Committee for National Statistics (CNSTAT) of the U.S. National Academy of Sciences for 6 years and chaired or participated in several CNSTAT panels. He has also served on the Board of Scientific Counselors of the U.S. National Center for Health Statistics, the Federal Economic Statistics Advisory Committee, and Statistics Canada's Advisory Committee on Statistical Methods, which he now chairs. He has served as president of the International Association of Survey Statisticians. Dr. Kalton is a Fellow of the American Statistical Association, a Fellow of the American Association for the Advancement of Science, an elected member of the International Statistical Institute, and a National Associate of the National Academies, National Research Council.

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 523—526

REPORT

The XXXII International Conference on Multivariate Statistical Analysis, 18–20 November 2013, Łódź, Poland

The 32nd edition of the International Conference on Multivariate Statistical Analysis was held in Łódź, Poland on November 18-20, 2013. The MSA 2013 conference was organized by the Department of Statistical Methods of the University of Łódź, the Institute of Statistics and Demography of the University of Łódź, 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, and Marta Małecka, M.Sc., and Artur Mikulec, Ph.D. – both from the Department of Statistical Methods of the University of Łódź – served as the conference's scientific secretaries.

The Mayor of the City of Łódź, Hanna Zdanowska took the honorary patronage of the Multivariate Statistical Analysis MSA 2013 conference. Its organization was financially supported by the National Bank of Poland, the Polish Academy of Sciences and the Łódź City Council. The official partners of the MSA 2013 conference were StatSoft Polska Sp. z o.o. and BazaKonferencji.pl. The conference was held as a participating event of the International Year of Statistics 2013.

The 2013 edition, as all previous Multivariate Statistical Analysis conferences, aimed to create 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 presented and discussed statistical issues were based on questions identified during previous MSA conferences. The scientific programme covered various statistical problems, including multivariate distributions, statistical tests, non-parametric inference, discrimination analysis, Monte Carlo analysis, Bayesian inference, application of statistical methods for finance, economy, insurance, capital market, risk management, design of experiments and survey sampling methodology, mainly for the social sciences purposes. The conference was attended by 86 participants from Poland and abroad (scientists, academic tutors, representatives of the National Bank of Poland, local statistical offices, business). In 11 sessions 58 papers were presented, including 2 invited lectures.

The conference was opened by the Head of the Organizing Committee, Professor Czesław Domański. The subsequent speakers on the conference opening were Professor Zofia Wysokińska, Pro-Rector in Charge of International Cooperation, who was the representative of the Rector of the University of Łódź, Professor Włodzimierz Nykiel and Professor Paweł Starosta, the Dean of the Faculty of Economics and Sociology of the University of Łódź.

The opening plenary session was chaired by Professor Tadeusz Bednarski (University of Wrocław) and it was started by Professor Mirosław Krzyśko (Adam Mickiewicz University in Poznań), who presented to all participants his invited lecture *Statistical methods of data analysis for multivariate functional data*. The second statistician to present his invited paper was Professor Eugeniusz Gatnar (National Bank of Poland) with the lecture *On sovereign debt and statistics.*

The historical plenary session, chaired by Professor Mirosław Krzyśko was dedicated to eminent Polish scientists. Professor Józef Pociecha (Cracow University of Economics) recalled the work and profile of outstanding Cracow statisticians: Kazimierz Zając and Andrzej Iwasiewicz. Professor Tadeusz Bednarski dedicated his presentation to Witold Klonecki and Professor Wojciech Zieliński (Warsaw University of Life Sciences) to Ryszard Zieliński. Professor Czesław Domański presented the profiles of Julian Perkal and Tadeusz Czacki.

Further sessions were chaired respectively by:

SESSION III	Professor Józef Pociecha (Cracow University of Economics)
SESSION IV A	Professor Daniel Kosiorowski (Cracow University of Economics)
SESSION IV B	Professor Eugeniusz Gatnar (National Bank of Poland)
SESSION V A	Professor Andrzej Sokołowski (Cracow University of Economics)
SESSION V B	Professor Grażyna Dehnel (Poznań University of Economics)
SESSION VI	Professor Grzegorz Kończak (Katowice University of Economics)
SESSION VII A	Professor Józef Dziechciarz (Wrocław University of Economics)
SESSION VII B	Professor Tomasz Michalski (Warsaw School of Economics)
SESSION VIII A	Professor Jerzy Korzeniewski (University of Łódź)
SESSION VIII B	Professor Grażyna Trzpiot (Katowice University of Economics)
SESSION IX A	Professor Alina Jędrzejczak (University of Łódź)
SESSION IX B	Professor Tadeusz Gerstenkorn (University of Łódź)
SESSION X A	Professor Bronisław Ceranka (Poznań University of Life Sciences)
SESSION X B	Professor Marek Walesiak (Wrocław University of Economics)

SESSION XI Professor Wojciech Zieliński (Warsaw University of Life Sciences)

The titles of MSA 2013 papers are listed in the Appendix (by the order of the presented papers).

The MSA 2013 conference was closed by the Chairman of the Organizing Committee Professor Czesław Domański, who summarized the Conference as very effective and he added that all discussions and doubts should become inspirations and strong motivations to the 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 2014 is planned on November 17-19, 2014 and will be held in Łódź, Poland. The Chairman of the Organizing Committee, Professor Czesław Domański informed that this will be the 33rd edition of the conference and kindly invited all interested scientists, researchers and students to take part.

Prepared by:
Marta Małecka
Artur Mikulec

STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 527–528

The international conference on **Small Area Estimation (SAE 2014)** will take place from 3rd to 5th September 2014, and will be devoted to the methodology of small area statistics, which, following arrangements made by the European Working Group on Small Area Estimation, will be organized by the Chair of Statistics at the Poznan University of Economics. The conference will be co-organized by the Central Statistical Office (CSO) in Warsaw and the Statistical Office in Poznan. Prof. Janusz Witkowski, The President of CSO, and Prof. Marian Gorynia, the Rector of the Poznan University of Economics have taken Honorary Patronage of the conference.

The idea behind the SAE 2014 conference is to provide a platform for the exchange of ideas and experiences between statisticians, scientists and experts from universities, statistical institutes, research centres as well as other government agencies, local government and private companies involved in developing and applying the methodology of regional surveys, in particular small area estimation. The Poznan conference is another one in the series of conferences (Jyväskylä 2005, Piza 2007, Elche 2009, Trier 2011) intended to combine theoretical considerations and practical applications of SAE in public statistics.

The SAE 2014 conference will focus on applications of SAE in censuses, model-based estimation and its evaluation, the use of spatio-temporal models, robust methods, non-response, issues in sample selection, poverty estimation, the teaching of SAE and its applications in public statistics. The conference will also feature a discussion panel and a specialist workshop devoted to the theory and practice of indirect estimation methodology.

Some eminent experts in this field of statistics have already confirmed their participation in the conference, including: Malay Ghosh (University of Florida), J.N.K. Rao (Carleton University), Ray Chambers (University of Wollongong), Li-Chun Zhang (Statistics Norway, University of Southampton), Partha Lahiri (University of Maryland), Danny Pfeffermann (Hebrew University of Jerusalem),

Risto Lehtonen (University of Helsinki), Ralf Münnich (University of Trier), Domingo Morales (Universidad Miguel Hernández de Elche) and Isabel Molina (Universidad Carlos III de Madrid).

Detailed information about the conference is available on the conference website at www.sae2014.ue.poznan.pl. You are welcome to take part in what is planned to be a continuation of the discussion started during one of the first conferences on the topic: *International Scientific Conference on Small Area Estimation and Survey Design*, which was held from 30th September to 3rd November 1992 in Warsaw.